

EMERY AND RIMOIN'S
PRINCIPLES AND PRACTICE OF
MEDICAL GENETICS AND GENOMICS
SEVENTH EDITION

Foundations

Edited by

Reed E. Pyeritz, Bruce R. Korf,
Wayne W. Grody



EMERY AND RIMOIN'S PRINCIPLES AND PRACTICE OF MEDICAL GENETICS AND GENOMICS

Foundations

Seventh Edition

Edited by

Reed E. Pyeritz

Perelman School of Medicine at the University of Pennsylvania,
Philadelphia, PA, United States

Bruce R. Korf

University of Alabama at Birmingham, Birmingham, AL, United States

Wayne W. Grody

UCLA School of Medicine, Los Angeles, CA, United States



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2019 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-812537-3

For information on all Academic Press publications visit our website at <https://www.elsevier.com/books-and-journals>



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Publisher: Andre Wolff

Senior Acquisition Editor: Peter B. Linsley

Editorial Project Manager: Pat Gonzalez

Production Project Manager: Punithavathy Govindaradjane

Designer: Matthew Limbert

Typeset by TNQ Technologies

LIST OF CONTRIBUTORS

Stylianos E. Antonarakis

Department of Genetic Medicine and Development,
University of Geneva Medical School, Geneva,
Switzerland

T. Mark Beasley

Department of Biostatistics, School of Public Health,
University of Alabama at Birmingham, Birmingham,
AL, United States

Darci T. Butcher

Genetics and Genome Biology, Research Institute, The
Hospital for Sick Children, Toronto, ON, Canada

Lucas Calais-Ferreira

Centre for Epidemiology and Biostatistics, Melbourne
School of Population and Global Health, University
of Melbourne, Melbourne, VIC, Australia; CAPES
Foundation, Ministry of Education, Brasilia, Brazil

Rita M. Cantor

Department of Human Genetics, David Geffen School
of Medicine at UCLA, Los Angeles, CA, United States

Stephen D. Cederbaum

Research Professor and Professor Emeritus of
Psychiatry, Pediatrics and Human Genetics, University
of California, Los Angeles, Los Angeles, CA, United
States

Sanaa Choufani

Genetics and Genome Biology, Research Institute, The
Hospital for Sick Children, Toronto, ON, Canada

Jackie Cook

Consultant in Clinical Genetics, Sheffield Clinical
Genetics Service, Sheffield Children's NHS Foundation
Trust, Sheffield, United Kingdom

David N. Cooper

Institute of Medical Genetics, Cardiff University,
Cardiff, United Kingdom

Jeffrey M. Craig

Deakin University School of Medicine, Geelong, VIC,
Australia; Murdoch Children's Research Institute, Royal
Children's Hospital, Parkville, VIC, Australia

Michael R. Crowley

Department of Human Genetics, Emory University,
Atlanta, GA, United States; The Department of
Genetics, The University of Alabama at Birmingham,
Birmingham, AL, United States

Cheryl Cytrynbaum

Genetics and Genome Biology, Research Institute,
The Hospital for Sick Children, Toronto, ON, Canada;
Clinical and Metabolic Genetics, The Hospital for Sick
Children, Toronto, ON, Canada

Allison Fialkowski

Department of Biostatistics, School of Public Health,
University of Alabama at Birmingham, Birmingham,
AL, United States

Geoffrey S. Ginsburg

Center for Applied Genomics & Precision Medicine,
Duke University School of Medicine, Durham, NC,
United States

Wayne W. Grody

UCLA School of Medicine, Los Angeles, CA, United
States

Susanne B. Haga

Center for Applied Genomics & Precision Medicine,
Duke University School of Medicine, Durham, NC,
United States

Judith G. Hall

Departments of Medical Genetics and Pediatrics,
British Columbia's Children's Hospital, Vancouver, BC,
Canada

Madhuri R. Hegde

Department of Human Genetics, Emory University,
Atlanta, GA, United States; The Department of
Genetics, The University of Alabama at Birmingham,
Birmingham, AL, United States

Fuki M. Hisama

Division of Medical Genetics, Department of Medicine,
University of Washington, Seattle, WA, United States

H. Richard Johnston

Department of Human Genetics, Emory University
School of Medicine, Atlanta, GA, United States

Bronya J.B. Keats

Department of Genetics (Emeritus), Louisiana State
University Health Sciences Center, New Orleans, LA,
United States

Bruce R. Korf

Department of Genetics, University of Alabama at
Birmingham, Birmingham, AL, United States

Marie T. Lott

Center for Mitochondrial and Epigenomic Medicine,
Children's Hospital of Philadelphia, Philadelphia, PA,
United States

George M. Martin

Department of Pathology, University of Washington,
Seattle, WA, United States

Fady M. Mikhail

Cytogenetics Laboratory, Department of Genetics,
University of Alabama at Birmingham, Birmingham,
AL, United States

Daniel W. Nebert

Department of Environmental Health and Center for
Environmental Genetics, University of Cincinnati
School of Medicine, Cincinnati, OH, United States;
Department of Pediatrics, University of Cincinnati
School of Medicine, Cincinnati, OH, United States;
Division of Human Genetics, Cincinnati Children's
Hospital Medical Center, Cincinnati, OH, United States

Junko Oshima

Department of Pathology, University of Washington,
Seattle, WA, United States

Vincent Procaccio

Biochemistry and Genetics Department, MitoVasc
Institute, UMR CNRS 6015 – INSERM U1083, CHU
Angers, Angers, France

Reed E. Pyeritz

Perelman School of Medicine at the University of
Pennsylvania, Philadelphia, PA, United States

Katrina J. Scurrah

Centre for Epidemiology and Biostatistics, Melbourne
School of Population and Global Health, University of
Melbourne, Melbourne, VIC, Australia

Stephanie L. Sherman

Department of Human Genetics, Emory University
School of Medicine, Atlanta, GA, United States

Michelle T. Siu

Genetics and Genome Biology, Research Institute, The
Hospital for Sick Children, Toronto, ON, Canada

Hemant K. Tiwari

Department of Biostatistics, School of Public Health,
University of Alabama at Birmingham, Birmingham,
AL, United States

Benjamin Tycko

Division of Genetics & Epigenetics, Hackensack
Meridian Health Center for Discovery and Innovation,
Nutley, NJ, United States

Mark P. Umstad

Department of Maternal-Fetal Medicine, The Royal
Women's Hospital, Melbourne, VIC, Australia;
University Department of Obstetrics and Gynaecology,
University of Melbourne, Melbourne, VIC, Australia

Douglas C. Wallace

Center for Mitochondrial and Epigenomic Medicine,
Children's Hospital of Philadelphia, Philadelphia, PA,
United States; Perelman School of Medicine, University
of Pennsylvania, Philadelphia, PA, United States

Rosanna Weksberg

Genetics and Genome Biology, Research Institute,
The Hospital for Sick Children, Toronto, ON, Canada;
Clinical and Metabolic Genetics, The Hospital for Sick
Children, Toronto, ON, Canada

Ge Zhang

Department of Pediatrics, University of Cincinnati
School of Medicine, Cincinnati, OH, United States;
Division of Human Genetics, Cincinnati Children's
Hospital Medical Center, Cincinnati, OH, United States

PREFACE TO THE SEVENTH EDITION OF *EMERY AND RIMOIN'S PRINCIPLES AND PRACTICE OF MEDICAL GENETICS AND GENOMICS*

The first edition of *Emery and Rimoin's Principles and Practice of Medical Genetics* appeared in 1983. This was several years prior to the start of the Human Genome Project in the early days of molecular genetic testing, a time when linkage analysis was often performed for diagnostic purposes. Medical genetics was not yet a recognized medical specialty in the United States, or anywhere else in the world. Therapy was mostly limited to a number of biochemical genetic conditions, and the underlying pathophysiology of most genetic disorders was unknown. The first edition was nevertheless published in two volumes, reflecting the fact that genetics was relevant to all areas of medical practice.

Thirty-five years later we are publishing the seventh edition of *Principles and Practice of Medical Genetics and Genomics*. Adding “genomics” to the title recognizes the pivotal role of genomic approaches in medicine, with the human genome sequence now in hand and exome/genome-level diagnostic sequencing becoming increasingly commonplace. Thousands of genetic disorders have been matched with the underlying genes, often illuminating pathophysiological mechanisms and in some cases enabling targeted therapies. Genetic testing is becoming increasingly incorporated into specialty medical care, though applications of adequate family history, genetic risk assessment, and pharmacogenetic testing are only gradually being integrated into routine medical practice. Sadly, this is the first edition of the book to be produced without the guidance of one of the founding coeditors, Dr David Rimoin, who passed away just as the previous edition went to press.

The seventh edition incorporates two major changes from previous editions. The first is publication of the text in 11 separate volumes. Over the years the book had grown from two to three massive volumes, until the electronic version was introduced in the previous

edition. The decision to split the book into multiple smaller volumes represents an attempt to divide the content into smaller, more accessible units. Most of these are organized around a unifying theme, for the most part based on specific body systems. This may make the book more useful to specialists who are interested in the application of medical genetics to their area but do not wish to invest in a larger volume that covers all areas of medicine. It also reflects our recognition that genetic concepts and determinants now underpin all medical specialties and subspecialties. The second change might seem on the surface to be a regressive one in today's high-tech world—the publication of the 11 volumes in print rather than strictly electronic form. However, feedback from our readers, as well as the experience of the editors, indicated that access to the web version via a password-protected site was cumbersome, and printing a smaller volume with two-page summaries was not useful. We have therefore returned to a full print version, although an eBook is available for those who prefer an electronic version.

One might ask whether there is a need for a comprehensive text in an era of instantaneous internet searches for virtually any information, including authoritative open sources such as *Online Mendelian Inheritance in Man* and *GeneReviews*. We recognize the value of these and other online resources, but believe that there is still a place for the long-form prose approach of a textbook. Here the authors have the opportunity to tell the story of their area of medical genetics and genomics, including in-depth background about pathophysiology, as well as giving practical advice for medical practice. The willingness of our authors to embrace this approach indicates that there is still enthusiasm for a textbook on medical genetics; we will appreciate feedback from our readers as well.

The realities of editing an 11-volume set have become obvious to the three of us as editors. We are grateful to our authors, many of whom have contributed to multiple past volumes, including some who have updated their contributions from the first or second editions. We are also indebted to staff from Elsevier, particularly Peter Linsley and Pat Gonzalez, who have worked patiently with us in the conception and production of

this large project. Finally, we thank our families, who have indulged our occasional disappearances into writing and editing. As always, we look forward to feedback from our readers, as this has played a critical role in shaping the evolution of *Principles and Practice of Medical Genetics and Genomics* in the face of the exponential changes that have occurred in the landscape of our discipline.

PREFACE TO *FOUNDATIONS*

All previous editions of *Principles and Practice of Medical Genetics* have started with a section on basic principles. Although the primary audience for the book is medical geneticists, who might be expected to have mastered basic genetic principles, in fact the breadth of the field makes it difficult for any individual to keep up in all areas. Moreover, in the current edition we have divided the text into 11 volumes in order to make the book more accessible to practitioners who are not medical geneticists but need a resource on genetic and genomic approaches in their discipline. They might especially find an overview of basic principles to be useful.

This volume begins with a summary of the history of medical genetics, initially written by the late Dr. Victor McKusick, who had a unique vantage point on the origin and development of the discipline. It is impossible to replace that perspective, and the history as he recorded it has not changed, but we are grateful to Dr. Stephen Cederbaum for being willing to add commentary on

more recent events. We then provide broad overviews on medicine in a genetic context, the newer concept of precision medicine, and the nature and frequency of genetic disease. Subsequent chapters provide overviews of critical topics in three major themes—the flow of genetic information in the cell, the transmission of genetic traits in families, and the forces that mold allele frequencies in the population. We do not focus on medical applications in this volume. Broad coverage of medical genetic and genomic approaches is provided in Volume 2 and underlies the basis of virtually all the other volumes in this text.

Medical genetics and genomics is the quintessential moving target, advancing at a pace that could never be captured in a textbook. Nevertheless, we hope that this overview volume will provide a comprehensive snapshot of genetic and genomic principles, providing a foundation for the subsequent volumes on medical applications and system-specific genetic and genomic approaches.

Medicine in a Genetic and Genomic Context*

Reed E. Pyeritz

Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, United States

1.1 INTRODUCTION: OUR HISTORY

The history of science is characterized by an exponential rate of expansion [1]. No aspect has escaped, but biology, which is relatively new, has by all accounts exploded. Naturally, these changes are reflected in new principles, new thinking, and new ways of handling new information. Among the problems created is that of making these novelties available to practitioners of science of all kinds.

Among the ways suitable to medicine are massive print volumes that contain detailed summaries of diseases, usually of one class, such as endocrine, gastroenterological, or, as in the case of *Principles and Practice of Medical Genetics and Genomics* (PPMGG), inherited. And of course, the pace of change requires revisions, always characterized by increases that reflect the rate of accumulation. Fission adds volumes whose pages, chapters, contributors, and diseases all do their best to obey the exponential imperative. Each chapter represents one topic more or less, an expansible topic capable of embracing new disorders with each new edition, so the number of chapters is no guide to the number of diseases. During the past decade, print volumes are increasingly being supplemented, or even replaced, by documents available electronically. In addition to new diseases, new paths of basic science are added, a characteristic of books that mirrors progress in reductionist investigation.

But where is reductionist biology taking us? Clearly, one direction is toward fragmentation; more and more is learned about increasingly restricted fields so that even specialties bifurcate and medicine becomes ever more splintered. But despite such assaults, whatever it is we call “medicine” has at the bottom some integrity, some consistency, and common grounds that are clearly revealed in PPMGG as well as in its sister enterprise, *The Metabolic and Molecular Basis of Inherited Disease* (MMBID) [2]. One such common ground is genetics. And as such a striking variety of disorders of cellular structures and metabolic mechanisms engaging every organ and organ system are included in both of these books, it is easy to imagine that genetic variation is the basis of all diseases. This idea is far from new, having been suggested even in the 18th and 19th centuries, when it took the form of diathesis and idiosyncrasy [3]. Then, in this century, it appears in the shape of a continuity between clear-cut segregating monogenic diseases and varying degrees of familial aggregation of cases that suggest the outcomes of the actions of more than one gene acting in environments favorable for the onset of a disease. But now, with the advent of genomics, which makes possible the study of the genetics of diseases of complex origin in families of patients who have affected relatives, as well as in those who do not, we are learning

*This chapter was authored in earlier editions by Barton Childs, MD, of the Johns Hopkins University School of Medicine. Professor Childs died in 2010.

that genetic variation underlies the latter no less than the former. So the continuity of segregating to nonsegregating familial aggregation is extended to include cases where there is neither segregation nor aggregation [4]. Perhaps we should require a disease to be shown not to be associated with any genetic variation, before saying it has no genetic basis. (refer to the chapter on The Genetics of Common Disease).

All professions undergoing rapid change and increasing specialism face the same dilemma. The generalists, who must keep up, find the density of new information daunting, even impossible, to assess and retain. So, books such as the PPMGG are intended to present this information in an orderly way and in relation to specific diseases. But the job is no sooner done than even newer information arrives to change how the various disorders are perceived and, of course, treated. Furthermore, new diseases have been described and must be included. Hence another edition must appear. And that's not all. The various sciences that contribute to our understanding are all changing, too, providing new insights that challenge conventional thinking. Editors respond to this intense pressure by including articles that present not only new information but also new insights, new ways of thinking about groups of diseases or perhaps all, and these usually appear at the beginning, hinting strongly that the reader of any later chapter would do well to read these preliminary ones.

Perhaps focusing first on the principles of chromosomal organization, genomics, and the investigations of diseases of complex origin will assist in understanding the chapters on developmental anomalies, the origins of high blood pressure, or inborn errors. And this may happen, given the effort. But each reader who makes this synthesis for himself or herself is likely to do it in the context of some specific disorder rather than to generalize the principles to all diseases. Indeed, we lack a clearly articulated set of principles of disease as opposed to diseases. That is not to say that medicine lacks principles; the idea of the body as a machine that breaks and needs fixing is one, and the medical history, diagnosis, pathogenesis, treatment, prognosis, and prevention all have a conceptual basis, as do the basic sciences related to medicine. But disease as a concept seems to be taken for granted. However, PPMGG certainly does not suggest that a student of medicine (and we are all students throughout the length of our careers) might take profit in an account of disease as opposed to diseases,

including why we have it, who is likely to be affected and how, and when in the lifetime and what forms can it take, as well as what are its constraints. That is, what are the explanatory generalizations that compose a context within which to fit all diseases?

Similarly, definitions of disease have fallen by the wayside. It is true that many such definitions have been offered; there is a sizable literature on the subject [5–8]. Perhaps today's reluctance stems from physicians' perception that we have not had the wherewithal for any but descriptions based on signs and symptoms rather than anything at its core. But today we are satisfied with a definition of a disease when pathogenesis is explained by reference to abnormality of some metabolic or homeostatic system, and we can describe the qualities of the proteins that compose the system. Now, if that is so, why may we not define disease as a consequence of incongruence of a metabolic or homeostatic system with conditions of life? And as all such systems are composed of proteins capable of reflecting the variations of their genetic origins, is it not appropriate to agree with Vogel and Motulsky, who, in the third edition of their book *Human Genetics*, proclaimed genetics as the principal “basic science for medicine” [9]?

If genetics is the basic science for medicine, it should be possible to construct a set of principles that characterize disease in a genetic context—that is, a set of generalizations shared by all diseases and framed in genetic terms. And there should be hierarchies of principles, inclusive and of increasing generality and forming a matrix embracing them all. What follows is one such matrix.

1.2 THE PRINCIPLES OF DISEASE

A foundation for developing principles of disease exists in the ideas of Ernst Mayr [10]. Mayr perceived biology as divided into two areas differing in concept and method. One, functional biology, is concerned with the operation and interaction of molecules, systems, and organisms. Causes are proximate, the viewpoint is inward, and questions are commonly preceded by how; how does the organism function? The other area, Mayr calls evolutionary. It is concerned with the history of functional biology, its causes are called ultimate, and its questions are prefaced by why; why in the sense of, what is the history of organisms, what are the conditions of the past that have made it possible to ask for answers

to the how questions? The two areas of biology meet, or overlap, at the level of the DNA, so that the functional deals with everything after transcription, whereas the evolutionary centers on the history of the DNA as well as, presumably, with the evolution of the conditions of the environment within which organisms have attained their current state.

Mayr did not include disease in his description of the two biologies, but disease is no less biological than the ideal state, so there should be no difficulty in applying his principles to biological abnormality. Thus, in relation to disease, the proximate causes are the products of the variant genes and the experiences of the environment with which they are incongruent. Ultimate or remote causes are first, the mechanisms of mutation and the causes of fluctuations through time of the elements of the gene pool, including selection, mating systems, founder effects, and drift, and second, the means whereby cultures and social organization evolve. In disease, the variant gene products and the experiences of the environment with which they are incongruent account for characteristic signs and symptoms, but in making available the particular proximate causes assembled by chance in particular patients, it is the remote causes that impart the stamp of individuality to the case.

So the model relates disease to causes, to the gene pool and ultimately to biological evolution, as well as to the evolution of cultures, and to individuality, the latter a consequence of the specificities of both causes. Here there are also elements for constructing a context of principles of disease, always remembering that the word context derives from the Latin word *contexere*, meaning “to weave.” That is, the principles must be seen to be related and interdependent so as to form a network of ideas within which to compose one’s thoughts about each specific example of each disease.

There is a further feature of Mayr’s views on biology, also crucial in its application to disease [11]. It is the state of mind in which to observe patients. In medicine, we tend to think of patients in relation to their disease, that is, as a class of people characterized by the name of the disease. This is what Mayr calls “typological thinking.” Although patients do differ somewhat from one another, they all share an essence: the disease. In contrast, Mayr proposes population thinking, in recognition that populations consist not of types but of unique individuals. So, in this context, disease has no essence; its variety is imparted by that of the unique individuals

who experience it, each in their own private version, and the name of the disease is a convenience, an acknowledgment of the necessity to group patients for logistical purposes. The fruits of the Human Genome Project (HGP) can be accommodated only with such a population perspective.

Why do we need such principles? Physicians are pragmatic; their way is determined by what they see before them, and students and especially residents are intolerant of anything they can label “philosophical.” But the principles are there, explaining the qualities and behaviors of diseases, and they await exposition.

But have we not already discovered them? Medicine is at the peak of success in diagnosis and treatment and moving rapidly to ever new heights of achievement. But all changes may not be equally evident. For example, the analysis of pathogenesis, traditionally a top-down process, is beginning to give way to a bottom-up approach in which discovery of variant genes leads to variant protein products and thence to the same molecular analysis of pathogenesis (refer to the chapter on Pathogenetics). Also, the genetic heterogeneity and individuality of disease are not easily accommodated in traditional thinking. So we are changing how we look at disease, how we define and classify it, and the language we use in describing it. For example, “genomics” and “proteomics” are words that embody ways of thinking new in the past several decades [4,12,13]. These developments are changing our relationships to biology and society. Biologists are expressing interest in the fates of the molecules they discover, and the public is becoming aware of what molecular biology and genetics mean to them, as risk factors, for example [14,15]. So, because this same molecular genetics gives us new insights into the principles that govern—and have always governed—disease, should we not articulate those principles and weave them into our thinking?

Reasons for doing so lie in the need for coherence in medicine, coherence in the face of reductionist dispersion, coherence in bringing new developments to the whole of the medical enterprise and to the public, and coherence in medical education and the thinking that goes into it. No one can possibly know all the information there is, but we all need a context that can supply both a substrate on which to apply the new and a receptacle within which to encompass our own field. The principles of disease bear a relationship to diseases that resemble the relationship of military strategy to

tactics, of historiography to the practice of history, or of grammar to precision in learning and speaking a language. Once such principles have sunk into the unconscious, they remain there as a context and a basis from which the conscious thinking about the subject takes off. They are no longer “philosophy” but the basis for daily thought.

1.3 DEFINING DISEASE

If we are to define “disease,” it must be as loss of adaptation; the open system has had difficulty in maintaining homeostasis. So our question is, how is this failure of adaptation attained? The straightforward answer is to say that a variation in a homeostatic system was incongruent with its environment, whether within the cell or outside, and the mechanisms for compensation were inadequate, momentarily or permanently, to restore congruence (refer to the chapter on Pathogenetics). As a result, other systems were affected, and then still others. But this only tells us that the machine broke down. If we would define disease, we must know what variations can lead to what levels of incongruence. We must know the weaknesses in the evolution of organisms, or if not weaknesses, the degrees of flexibility. That is, the origins of human disease lie in both human evolution and the environment with which the human species has evolved to be congruent. And because both biological and cultural evolutions are continuous, although at markedly different rates, congruence must be relative and changing.

Such questions have always been germane to the definition of disease, however infrequently posed, but they assume a new relevance now because of the frequent assertion that one needs to know only the molecular form of the incongruence to devise an appropriate treatment. That is, all our problems could be solved at the molecular level. If this were true, we do not need to define disease except molecularly, and that is the vision we pursue. But before committing wholly to such a concept, it is as well to probe further into the question, what is disease?

It may seem odd to ask such a question; surely it has been answered again and again. And so it has, and many times, but always within the descriptive limits of the period. Descriptions and definitions of disease have in history proceeded from top down and from outside in. That is, from a history of the illness only, to history plus

inspection, then on to physical examination in life and at autopsy, to increasingly intimate inspections by radiological and newer visual means, as well as biochemical examination, and now molecular analysis. Now that we can proceed from the bottom up, beginning with genes identified by genomics to their protein products and to the homeostatic systems into which they are integrated, the definition of disease should be reconsidered.

History reveals two opposing definitions. One, called essentialist, proposes that diseases exist somehow and in some way as entities that attack their victims. The other, called nominalist, is represented as a change within, an altered state, or a deviation in response to some stimulus. In the essentialist view, the patient is healthy and is brought low by the disease, whereas in the nominalist, the disease is an expression of the particularity of the individual response to a stimulus. This modern-sounding construction was popular in the nineteenth century in the form of “diathesis,” in which there was suggested some element of heredity as well as individual vulnerability [3,16]. It was swept aside by the essentialist version, which emerged, in the later part of the 19th and early 20th centuries, when microorganisms were discovered. Then the nominalist began to regain favor as, first, biochemistry, then genetics, and then molecular biology flourished. Today, although we still experience microbial scourges, the nominalist view prevails, perhaps because we can so easily see that the responses to diseases, even to microbial assaults, are a product of the individuality of the systems of homeostasis, and because we are more perceptive of the interrelationships of proximate and remote causes. But there is a lingering residuum of essentialism in molecular diagnosis, so often proclaimed to be a preliminary to some “designer” treatment, usually to be concocted by pharmaceutical companies who initially made the word “pharmacogenetics” their own. (Fortunately, today these efforts are subsumed by all health practitioners by the notion of precision medicine, a theme that pervades this edition of PPMGG.) Such a diagnosis is unobjectionable as far as it goes, but it has implications. An essentialist view, first, emphasizes the disease without differentiating the patient; second, includes only proximate causes, the gene and its product are perceived as no less essentialist than the microbe that attacks; and, third, is typological. Even though allelic heterogeneity may be acknowledged, the variation is around the expressions of the

“classic case.” There is no recognition that each patient will respond to the effects of the products of each gene individually, to say nothing of each designer drug. For many years, the monogenic diseases were regarded in an essentialist typological vein, but we have become more nominalist, more prone to population thinking, and more ready to recognize the significant effects of variability of both the genetic and the environmental settings in which the principal gene effect is measured [17]. So, in defining disease, we must not only take into account the gene(s) that seems most relevant to the phenotype—after which the phenotype may be neglected in the interest of molecular treatment—but also keep alive the relationships of genes (or better, of their products) and phenotypes, the better to grasp the individuality of each, so as to tailor the particularity of the molecular treatment to the biological individuality of a very particular patient. Then, having that principle in mind, the necessity to group patients for treatment can be managed rationally.

So, in the end, how shall we define disease? The elements of Mayr’s model must be satisfied. That is, the definition must include remote as well as proximate causes and the relationship of both to DNA. Also, it must be populational in concept rather than typological, which is to say it must be nominalist. So, one way of expressing it is as follows: disease is a consequence of incongruence between genetically variable homeostatic systems and the kinds, intensities, and durations of exposures to elements of the environments to which they are called on to adapt.

No doubt objections will be raised. Is cyanide poisoning a disease? No human variation is needed. Nor is there variation in susceptibility to scurvy, although it is unquestionably a disease. But although cyanide will extinguish all life, scurvy is a disease of species; only bats, guinea pigs, a few other species, and the order of the apes (including us) are vulnerable. Still, it is fair to say that human homeostasis is uniformly incongruent in the presence of cyanide and the absence of ascorbic acid, and so both qualify for this definition. Others see poisoning and trauma as something other than disease and call them by other names—accidents for one. But again the human constitution is incongruent with bullets and car crashes and so is vulnerable. How about infections? We have genetically determined mechanisms, both well developed and efficient, in coping with microbial invaders.

Here, the variability includes individual vulnerability in the many immunodeficiencies, as well as individual invulnerability, both relative and absolute, in individuals who are immune to infections caused by many organisms, including one malarial parasite, the tubercle bacillus, and the polio virus [18].

And if there are those who are immune to, who can doubt that there are individualities susceptible to particular organisms? In general, microorganisms attack at cellular sites we define as strengths but which they have defined for their purposes as our weaknesses; for example, cell surface molecules designed for high efficiency as elements of metabolism, but which the organisms have adapted themselves to use as means of access. The point is that it is usually variation in the microbial, rather than the human, cell that brings particularity to the encounter, although there may be both. So, because variation in either human victim or microbial attacker or both determine the nature of the encounter, infections fulfill the nominalist definition, even while as entities they are compatible with, and are even the prototype of, the essentialist definition. This suggests that both definitions are of historical interest only, suitable for the levels of description of disease that we have left behind. But they are still of value in showing how the more intimate we become with the human body, organ, and molecule, the more our concepts change and the more we need to shed old ideas and their locutions and adapt to what the new is telling us. But still we should observe that much of what is new was foreshadowed in the old. A leisurely reading of the chapter “The Inborn Factors in Infective Diseases” in Archibald Garrod’s book *Inborn Factors in Disease*, published in 1931, underlines the validity of that observation [19].

1.4 THE “HOW” QUESTIONS

In the Mayr model, it is by way of DNA that remote causes leave their imprint on the proximate and it is the protein gene products that are the effectors of both. Once we spoke of gene–environment interactions, but now we know that the actual contact between these sources of variation is by way of those proteins. Indeed, one way of perceiving the DNA is as a molecule that is helpless without proteins that carry out all its ends, including transcription and translation too [20]. So the proteins carry on the life of the cell as elements in

integrated systems, responding to influences from adjacent cells, distant organs, and the outside, to maintain the open system in its uncertain relationships in life. They are, therefore, unit steps of homeostasis, and as such are pivotal in concepts of life, development, aging, health, and disease.

Such a list of attainments is banal without explanation and illustration. In the following section, I discuss several ways in which the unit steps of homeostasis fulfill the purposes of the cell. They are called “unit steps” to convey their elemental state as units of pathways and cascades, structural elements, protein machines, transducers in signaling systems, and transporters or receptors of molecules that are going somewhere. The phrase further implies units of integration into systems intended to maintain the organism’s steady state; they are the node between nature and nurture, and the phrase has the virtue of being indifferent to whether the specific protein fulfills a useful purpose or is disruptive. And finally, the unit steps have an important historical meaning, representing the central idea of Garrod’s inborn error, Beadle and Tatum’s one gene–one peptide, and Pauling’s molecular disease [21–23].

1.4.1 Some Qualities of the Unit Step of Homeostasis

1.4.1.1 As a Unit of History

Clearly, DNA is an instrument of memory, a memory that in preserving the past, gives guidance for the future. That is, the future must always reflect the past, and the means whereby this Janus vision is attained is the protein gene product that repeats its phylogenetic history in its current composition and function and predicts its future in its reincarnation through subsequent generations as itself or in the form of variants. Some of the variants have no future and their incongruence is noted by natural selection. Others are contingent, favorable for some conditions and inappropriate for others. And then there are those proteins that have hardly changed from microbial ancestry and that represent core functions. In human society, political and religious systems have similar capacities for endurance, revealing fundamental unchanged dogma associated with adaptation in ways that promote the cause with little change in the fundamentals. So the proteins that constitute our proteomes descend to us not only from our parents and other human antecedents but also, with variable conservation, from both the ancient and recent past.

1.4.1.2 As Effectors of Gene Intention

We often speak of a gene or genes as being “for” something, by which we indicate some sort of direct relationship to a phenotype. That is, we seem to be saying that the gene’s influence is determining. And so it is, if by determining we mean the sequences of bases in mRNA and of amino acids in a protein product. In this sense, the gene is indeed “for” something. But each gene product has in addition, an emergent career of its own, not predicted at all, or only indirectly, by its gene. It assumes a position in the homeostatic device to which it belongs and can now be said to be “for” that system, as the factor VIII gene is both “for” the factor VIII protein and “for” clotting. But it is far from determining clotting; all the other elements are needed, too, or, as we all well know, life-threatening bleeding occurs. We also know that in physiology, system is integrated with system in hierarchical relationships, so that the farther away from the steps of translation and first integration, the more dilute becomes the gene’s determining power. No doubt the genes are involved wherever their products are to be found, but indirectly, and any one may have little power to shape the ultimate phenotype. In another sense, the genes appear to be hardly involved at all beyond transcription because it is the quality of the protein product that determines its role in the economy of the cell, a role that is determined by how the protein folds and takes shape, a shape that must accommodate to the shapes of the products of other genes and they with still others. No doubt the protein’s folding and shape reflect the information residing in its parent gene, but its gene has no control over the shapes of those other proteins with which it fits, to say nothing of how multiprotein machines work [24]. Here is a question not of genes but of how proteins interact. It is a matter of physiology. Indeed, it is possible that as the fruits of the HGP and the proteomists filter into medicine, we will hear a good deal less about genes and more about proteins [25]. This could be less than ideal were the proteins not perceived to be as closely identified with the concept of variation as the genes. Let us see to it that they are.

1.4.1.3 As a Unit of Development

T. H. Morgan adopted *Drosophila* as an organism suitable for the study of development [26]. But it did not work out that way, and his students led the way to the operational definition of the gene. So it is ironic that

technology initially made the fruit fly ideal for the very study that defied Morgan's efforts [27,28].

In development, the genes fulfill their intentions in the ways just described. Their products are the units of developmental change, assuming positions in systems appropriate for their conformation so as to give each organism a matrix, embodying a trajectory of change that is a product of how the embryo, fetus, and infant meet and respond to experiences of intrauterine and external environments. That is, development is a historical process; what the organism is today is built on what it was yesterday and leads to what it will be tomorrow [29]. And because the genes see to the continuity of their products throughout the changes of a lifetime, it is hardly likely that the influences of the past, however, distant, would fail to influence the present. So if we would understand the origins and expressions of disease, it must be in the context of three timescales, all at once: that of phylogeny; that of development maturation and aging; and that of the present [30]. To know what we begin with is to know potential incongruence; to know where development is taking, or has taken, us is to clothe the potential with the probable, one way or the other; and to know where we are at the moment is to know the strengths and weaknesses that we will face tomorrow. There is an increasing interest in the idea that some diseases of middle life have precursors, manifestations dating to early, even intrauterine, life. These expressions may not appear to relate to the disease they are said to characterize. Rather, they may represent subtle changes in trajectories that, if pursued, emerge finally as disease [31]. How else could birth weight be related to type 2 diabetes or heart attack?

1.4.1.4 As a Unit of Individuality

In medicine, patients are seen one at a time. Each one is biologically unique, has different experiences, and tells a different story. These expressions, together with the help of the laboratory and observations over time, are compared with those of the classic case to reach a diagnosis, and, allowing something for variation, treatment or management is devised. This thinking is typological, individuality is usually ignored, and the doctor is in thrall to nosology. The method works well enough, but heterogeneity of proximate cause may be overlooked and patients are likely to be aware when they are being perceived as representative of a class rather than as their unique selves. Now, molecular biology has given us the

wherewithal to observe molecular individuality, that is, the capacity to make comparisons between individuals of variations in base pairs in the DNA and differences in amino acid sequences in proteins. The unit of individuality is the unit step of homeostasis, and the expression of uniqueness lies in how the variant proteins affect each its own system and the integrations of the latter with others, as well as how the systems respond to nongenetic proximate causes. Genomic analysis of single nucleotide polymorphisms (SNPs) suggests that the number of polymorphic loci expressed in amino acid substitution in proteins will turn out to be somewhat greater than the 30% we are accustomed to [32,33]. This is the substrate of variability within which additional "private" variants as well as clearly bad mutants express their effects, and all this variation is manifested in how the integrated homeostatic devices are fulfilling their duties. So, if each human being is unique by virtue of the variant proteins in his or her whole physiological apparatus, why should not each such human being express an experience with disease as variously as a career of health?

Variation contributed by variant proteins is far from all. Such variability is compounded by the individuality of the developmental and maturational trajectory characteristic of each person, a path determined no less strongly by the kinds, intensities, and durations of experiences than by the protein gene products with which they interact. But the final arbiter of individuality is the remote causes, which determine the specificity of both genes and experiences. The variation in the parental gene pool is a sample of what is available to the species, but it is necessarily limited, characterized by ethnicity and made local by founder effect, migration, and mating customs. These are all influences that determine the particularity of an individual's genetic endowment. But if genetic individuality is both determined and constrained by the genetic raw material inherited at conception, so is the variety of experiences made possible and limited by the mores of the social and cultural milieu, itself often inherited, which shape our likes and dislikes, our indulgences and restraints, in short, the qualities and quantities of the experiences we encounter. So, in the end, it is the remote causes that confer the specificity of individuality, but the unit steps of homeostasis that supply the substrate. Of course, the idea of variant proteins as units of individuality is not a new one, having been proposed by Archibald Garrod as "chemical individuality" as early as 1902 [22].

1.4.1.5 The Unit Steps as Effectors of Disease

If the gene product is the implement of homeostasis, it follows that it is the effector of disease; certain of its variants are in some degree incongruent with the environment, inside the cell or out. That is, wherever the origins and mechanisms of pathogenesis have been laid bare, there are proteins at the root of it. How could it be otherwise, given that both structures and motivators of the functions of cells are proteins, and disease is a consequence of homeostatic incongruence? A critic might suggest infections as exceptions, but it is the congruence of the microorganism's structures with our unit steps of homeostasis that allows them to attach themselves to cell surfaces and then to release toxin or to gain access to the cell's interior and to reproduce there. It is they who define our strengths as weaknesses and our congruence as incongruence. And they do so by using the human gene products, the human unit steps of homeostasis.

The history of the realization of this role of the unit step in disease is of interest; it paralleled the successive descriptions and definitions of both genes and proteins [5]. We all know that Archibald Garrod was the first to call attention to alkaptonuria as a hereditary alternative form of metabolism because of failure of an enzymatically catalyzed step [22]. He called this, and other such metabolic aberrations, inborn errors to distinguish them from diseases. This was an insight of extraordinary penetration in which he recognized that the differences in protein composition that distinguished species must also differentiate individuals within species [34]. But even by 1909 when his first book, *Inborn Errors of Metabolism*, was published, he could go no further [16,34,35]. Little was known about protein structure and nothing of sequence of amino acids, and it was not even established yet, to everyone's satisfaction, that enzymes were proteins [36].

As for the gene in 1909, it was still defined statistically, and although phenotype and genotype were differentiated in that year, the gene was an unknown entity, perceived by Johannsen as "an accounting or calculating unit." But by 1915 the gene had been defined operationally, so by then Garrod could have proposed the inborn errors as products of mutants of single genes. But he never did. Even in his 1931 book, *Inborn Factors in Disease*, he did not use the word "gene" despite a general recognition that genes were involved somehow in some diseases [19,34]. For example, in 1927 Barker

reported that, "No less than 223 heritable anomalies have been described in man already" [37]. And others, not in medicine, recognized a biochemical relationship between genes and phenotypes: Wright in coat colors of guinea pigs and Wheldale in flower pigments [38,39]. Then in the late 1930s and early 1940s, the studies of Ephrussi, Beadle, and Tatum, first in *Drosophila*, then in *Neurospora*, provided a functional definition of the gene that brought gene and protein unequivocally together to clarify Garrod's observations, and capitalizing on rapid advances in biochemistry, to begin in the 1950s an era of biochemical genetics [21]. Biochemical genetics was an ecumenical enterprise. If Garrod was its icon and Harry Harris its chief expositor, there were also contributions of nongeneticists, including Pauling's concept of molecular disease and the elaboration of the enzyme deficiencies in (type I) glycogen storage disease, galactosemia and other disorders, all classic inborn errors, described by biochemists with no primary interest in genetics and who made no reference to Garrod or to Beadle-Tatum [23,40,41]. But whatever the influence, the list of inborn errors expanded rapidly, soon attaining an exponential rate of increase that has barely slackened.

It is worth noting that biochemical genetics flourished before the impact of the discovery of the double helix could be felt. But the later developments led first to Yanofsky's definition of the structural gene with its correspondence to sequences in amino acids in proteins [42] and later to the definition of the gene that includes both transcribed and nontranscribed DNA. And this led, in turn, to the development of genomics as an analytical method. Thus biochemical genetics, whose analysis proceeds from the phenotype to the protein and its gene, met genomics, whose analysis proceeds from the gene toward the phenotype via its protein product [43]. And in time, the glamor passed from biochemical genetics to genomics, perhaps principally because the former had no way to tackle the genetics of complex disorders. Actually both are needed because phenotypes are not necessarily explained on discovery of the gene or genes whose products are acting as proximate causes.

As the focal point in pathogenesis, the protein gene product provides an economical answer to the question of the origins of monogenic diseases. But the question of the moment is how to explain those called complex. The approach includes genomics, by which salient genes can be found and characterized [13]. Further steps involve discovery of their proteins and the homeostatic devices

to which they belong, after which the pathophysiology may be elucidated. Additional participation by genetically inclined thinkers lies in sorting out the heterogeneity by means of appropriate family studies, work that must be done before, or together with, efforts to tie treatments to the consequences of particular protein variants.

Today we scoff at such diagnostic “entities” as dropsy and consumption, having begun long since to resolve their heterogeneity. But the sequencing of the HGP will provide the means to show how much more we have to go to characterize distinctive versions of, say, heart attack and stroke. So numerous are the genetic contributors likely to be that a case might be made for everyone having his or her own version of heart attack, stroke, or other multigenic multifactorial disorders. So family studies are vital for deciding which genes play important roles in which versions of the disease. The results will resemble those in the study of monogenic disorders; the heterogeneity will be of both loci and alleles, and the sets thereof will vary from family to family and individual to individual [17]. This kind of genetic thinking, not yet routine in medicine, is crucial to our understanding and represents an important principle of disease.

1.4.1.6 The Protein Product as a Unit of Selection

Neodarwinism is the outcome of a debate in which geneticists agreed that the object of selection must be phenotype, not genes, whereas evolutionary biologists, to whom the phenotype had been that object all along, agreed that both phenotypes, and their variation, originated in the genes [44–46]. If so, although the phenotype remains the unit of selection, it is the variable unit step or steps that cause it to qualify for that fate. In medicine, we are not much concerned with the selection by which species attain their characteristics but with what evolutionary biologists call “purifying” selection, that which removes “undesirable” genes prior to reproduction. So here again the protein product of the gene occupies a central position between two aspects of human biology. And here is yet another example of the cleavage between biology and medicine. The irony in the word purifying is not lost on the physician to whom the protection of life is uppermost, while to biology, with no stake in the individual, the question is purely one of understanding the rise and decline of species. But, in fact, variations in unit steps of homeostasis are no less the stuff of positive selection than negative.

1.4.1.7 As a Hedge Against Genetic Determinism

Institutions change and renew themselves but they usually retain residual signs of their origins. No one would deny that all the genetics of today stems from the concepts elaborated in the fly room at Columbia, or that we continue to use both concepts and language appropriate to the drosophilists’ definition of the gene [20]. Theirs was an operational definition in which authority for both heredity and cellular function was accorded to the gene. In his book, *What is Life*, Schroedinger spoke of the gene as “law code and executive power” as well as “architect’s plan and builder’s craft in one” [47]. So the language of *Drosophila* genetics included such locutions as genes “for” gene–environment interaction, modifiers, penetrance, and pleiotropy, all of which are perceived as properties of the gene, although we know now that they refer to events mediated by the protein unit steps of homeostasis. There is no question of the latter’s specification by the genes, but in folding and assuming an appropriate position in a relevant homeostatic device, they become a part of mechanisms that regulate both themselves and the DNA (see chapter on Pathogenetics). Thus, it is not the genes that are penetrant or pleiotropic or that interact with the environment; it is the proteins that do these actions that are removed from the genes’ control.

It might be correct to speak of a “gene for,” say, an enzyme or even its pathway; for example, there is a “gene for” phenylalanine hydroxylase and “for” phenylalanine degradation. But in their further integration, proteins lose their identity in those of integrated functions, for which any single gene can no longer be perceived to have any authority.

There is another way in which the locution “gene for” is used. When we observe that a disease segregates, we say there is a “gene for” that disease, that one or more mutants act as proximate cause. That is exactly what the drosophilists did for their mutants, unconcerned with their ignorance of how a gene could shorten bristles or deform wings. We continue to use their discourse, even though we know that the protein product is the actual agent of function [20]. But “genes for” is a tricky phrase. When we use it unthinkingly, as in genes for high blood pressure, say, we obscure our own inner view of the reality, whereas when we speak of variant proteins, there springs immediately to mind pathways, cascades, receptors, transducers, and feedback loops such as those that

determine blood pressure. Incidentally, it is amusing to imagine that had Archibald Garrod come to alkaptonuria thinking like a geneticist of the time, he would probably have perceived it only as a recessive character, not an inborn error. But he came to it as a biochemist and saw it for what it was: a metabolic alternative due to the absence of an enzyme. He used the genetic evidence expressed in consanguinity to support the idea of heredity, not as evidence of a gene. So, rather than perceiving his lack of interest in genetics as a shortcoming, we should be glad of it because the idea of a “gene for” alkaptonuria could have stood in the way of his biochemical insights. But equally, had he pondered the work of the drosophilists emerging in print from 1905 to 1920, and which included their operational definition of the gene, the second edition of his *Inborn Errors of Metabolism* published in 1926 must surely have anticipated the Beadle–Tatum one gene–one enzyme principle [48].

So, if we human geneticists of today revert occasionally to the drosophiline mentality, how likely are patients, their families, and the public to escape? How are they to know that the words “gene for,” say IQ, artistic ability, or criminal behavior obscure the unfathomable complexity of the identity and actions of gene products integrated in hierarchies to compose cells, organs, and whole organisms, all in touch with one another and with the outside? The extremes to which “genes for” can go are summarized in a book called *The DNA Mystique: The Gene as a Cultural Icon* by Nelkin and Lindee [49]. But fortunately, we have our mental image of the products of the genes, the unit steps of homeostasis, with their multifarious behaviors as a bulwark against loose thinking.

1.4.1.8 As the Goal of the HGP

One road to the discovery of new principles of disease has derived from the HGP [50]. Lander has suggested that this bears the same relationship to biology that the periodic table bears to chemistry [13]. So it compels our attention. Furthermore, it is the ultimate identifier of those homeostatic units that lie at the basis of pathogenesis.

About 20,000 or so genes and their products have been identified, and sooner or later, the products’ roles in homeostasis will follow, with obvious benefits for investigation of pathogenesis, treatment, and prevention. In addition, definitive samples of gene products, useful in defining disease, will be available for characterizing

human biological properties hitherto unknown. A few examples of questions that are being asked are

1. How variable is the human genome? Is it more, or less, than the estimates of Harris and Lewontin? Studies of whole genome sequences suggest more [32,33]. Nothing could be more useful than this answer because it is the common genes that so often act as modifiers and furnish the wherewithal for complex diseases. And we now understand the much of the “junk” DNA encodes various RNA species that exert important control over the expression of genes that encode proteins.
2. Is there an inborn error for every locus? And are all classes of proteins equally involved in diseases? In a comparison of 348 mutant proteins associated with inborn errors (MMBID) and a list of 3000 “core” proteins shared by yeast and *Caenorhabditis elegans*, the distribution of protein types in the two samples was remarkably similar [51]. Although indirect, the suggestion is there that all protein types are involved in inborn errors, but we cannot yet say that there is an inborn error for every locus, however plausible the idea may be.
3. Are diseases characterized by the qualities of the proteins that are their proximate causes? For example, do enzyme deficiencies differ in some systematic way from disorders associated with receptors, transcription factors, or structural proteins?
4. Are conserved genes overrepresented or underrepresented in disease? One might expect them to be overrepresented on the assumption that they fulfill critical functions or underrepresented because their mutants might be so often lethal.
5. What is the role, if any, of developmental constraints in fostering or suppressing disease? These are limitations on the evolution of phenotypic variation expressed in developmental blind alleys. Kirschner and Gerhart have examined ways by which such constraints are loosened to allow new mutation and evolutionary progress. But would some of the latter be disease [52]? And Rutherford showed how, in *Drosophila*, such constraint was exerted by a heat shock protein [53]. When altered by mutation, the constraining force was lifted and the effects of mutants suppressed by the wild-type protein were observed. Some of these effects were developmental anomalies.
6. Are diseases characterized by the evolutionary age of the proteins that lie at their root? That is, we might

suppose that inborn errors of housekeeping genes shared by remotely related species were the oldest. Do they differ in any particular from diseases of the most recent mammalian or human genes?

7. What are the implications for aging? Are some proteins more frequently the object of aging processes, or is it random? Errors in the mitotic machinery that led to multiple abnormalities of regulation of dozens of enzymes increase with age [54]. So, will aging, which has been perceived by some as dishomeostasis, turn out to have the same molecular basis as disease?

Many other questions are being asked, many no doubt not now askable because the contexts in which they are relevant are unknown. As more and more diseases are given molecular definition, we will surely classify them differently, departing from the current anatomical, organ system, age-related rubrics, moving to more molecular designations. As heterogeneity is laid bare, old classes will go and new ones will come, reflecting a sharp revision in how we will see disease itself. In addition, our language will change. It is likely that we will refer less to genes and more to proteins, so our residual drosophiline language is likely to go, too. Of what use are words such as modifier, epistasis, penetrance, pleiotropy, and the like when visualizing the reality as actions and interactions of proteins, in say, multiunit machines, or even in whole systems [13]? This also suggests that we in medicine will be thinking less in units and more in multiunit devices (proteomes, epigenomes, metabolomes, etc.). Linear thinking may be out as complexity moves in. But maybe the most significant change in our thinking will be compelled by the definitive evidence of human variation and individuality. Typological thinking will give way to population thinking. No doubt there will always be use for the former at one level; that of the value of means and the classic case, but only as a preliminary to the population thinking that perceives the extent and impact of variation on human individuality.

1.4.1.9 Social Impact

The unit step of homeostasis is attaining increasing prominence as a risk factor and signal for preventive action, and medicine has been adapting not only to their potential use but also to their impact on their possessors' lives. These concerns are well known to readers of PPMG; they have been the subject of many articles, books, and committee reports and they touch on counseling, ethics, legal matters, and psychological impact

[14,15,55–57]. They are mentioned here because of the potential uses of such risk factors in prevention. If the HGP fulfills its promise, there is the possibility to know the protein products of all genes known to participate in pathogenesis. Many scenarios as to the use of these markers have been offered. Only time will tell which, if any, is practical, but we would be wise to continue our study and preparation. How to use information, available at birth, about many variant genes, perhaps dozens in single individuals, and known to be associated with diseases all across the life span is something entirely new in medicine. Is it consonant with good medicine? Is it acceptable to the public? How do we prepare the public to make rational decisions about it? How do we prepare individuals to accept and use constructively such emotionally loaded information? These questions can be answered only in colloquy with the public.

1.4.1.10 As a Source of Coherence

In concentrating on the specificities of the pathogenesis of each disease, reductionist investigation emphasizes the separateness of diseases. It exerts a centrifugal effect that adds to that of our conventional nosology, which divides medicine into specialties across which we interact collaboratively, when at all. But the concept of the unit step of homeostasis as the central focus of all pathogenesis provides a principle of disease that exerts a contrary centripetal force that unifies the thinking about both disease and diseases. It is the difference between analysis and synthesis. Medical thinking, until recently, was mainly synthetic. It dealt with the body as a whole, no doubt because of ignorance of its parts. In contrast, in the thinking of today, the emphasis is on analysis; our attention is directed more to microunitary parts with less attention to the whole. But in acting as units of the mechanisms whereby an open system maintains its adaptation to an indifferent environment, the protein gene product is the effective link between those proximate and remote causes portrayed in the Mayr model. And that link is no less evident when unit step and environment are incongruent than when congruent. In the Mayr model, the proximate causes are consequent on the DNA and pose questions prefaced by how, whereas the remote causes lead up to the DNA and pose questions prefaced by why. In this summary of the role of the protein gene product, we have dealt with “how” questions. In the next section, we examine the “why” questions and the principles they illustrate.

1.5 THE “WHY” QUESTIONS

Just as the questions preceded by “how” are answered by reference to proximate causes, so are the “why” questions answered by reference to remote causes. And the answers to the why questions contribute no less to the specificity of identity than the proximate. Indeed, they are the enablers, the ultimate arbiters of that specificity. So we are what we are by virtue of endowment, experience, development, maturation, and aging, but the bounds of what we can be and the precise description of what we are, are determined by how the remote causes came to be what they are and how they were sampled in the making of the individual. So the answers to the why questions are likely to probe more widely and deeply than those preceded by “how.” In fact, they begin with the latter to give them specificity. And yet we spend most of our energies and money on seeking proximate causes. The imperatives of treatment and prevention require it, yet it should be observed that it is the “why” questions that are most often asked by the public, particularly by those affected by disease. Medical education would do well to include them.

1.5.1 Why Do We Have Disease?

Why is it that after all these eons of evolution, species have not evolved to perfect attunement with the environment? First, it must be said that we are remarkably well adapted. Increasing longevity suggests that in the developed world we are moving toward the ideal rectangular survival curve [58]. That is, we are moving in the direction, at least, of some probably unattainable minimal amount of disease compatible with some necessary degree of genetic diversity. In addition, the environment changes, requiring new adaptation, and then observation suggests that nature never reaches for perfection but for some compromise that ensures perpetuation of species usually at the expense of individuals. For example, the fecundity of our own species is presumed to be of the order of only 25%, and not all of that makes it to maturity [59]. But that a principal hazard to the global ecology is a surplus of human beings is testimony to nature’s way of doing business.

Because evolution proceeds by the intervention of natural selection, there must be something to select, and again, although there are mechanisms of astonishing precision to ensure the accuracy of replication of DNA, they are not perfect either, and so, after the removal of

individuals unlikely ever to survive or reproduce, we are still left with sufficient variation among individuals to adapt to the randomness of change in the environment. This includes genes that are either neutral under all conditions or contingent, that is, adaptive under some conditions but nonadaptive and conducive to incongruence and disease under others. So disease must be presumed to be a byproduct of the necessity to have enough variability for all conditions.

1.5.1.1 Why This Disease?

Patients often ask this question. “Why,” they ask, “should I have diabetes or cancer?” or “Why should my baby have this bizarre disease I’ve never even heard of?” The media have spread the news of genetics, so their question may be, “Is it in my genes, and if so where in the world did such genes come from?” So the questions are directed to both proximate and remote causes. First to the proximate, which include the genetic endowment received by the patient at conception and which, together with the kinds, duration, and intensities of experiences of the environment, have created a developmental and maturational matrix that is an expression of individual potential from day to day. Then, the parental contribution to this matrix reflects the specificity of their genes, themselves representative of one or more gene pools, each with its own variable composition and history. The contributions of experiences are representative of qualities of the society the patient inhabits, qualities that vary with cultural history. And finally, the contribution of development and maturation is that of a trajectory whose specificity is derived from both endowment and experiences as they create and characterize the evolving matrix within which the incongruence and disease are engendered. So the answer to the question of “why this disease” requires us to acknowledge that while the disease is a consequence of incongruence between proximate causes, it is the remote causes that account for the existence, availability, and particularity of those proximate factors.

1.5.1.2 Why This Person?

If patients are baffled by the disease they experience, they are angered by its apparent choice of themselves. “Why me?” is their injured cry. Of course, the reasons reside in the origins and specificities of genes, experiences, and development given in answer to the previous question, but “Why me?” is a profoundly different

question because, no matter how specific the genes, the experiences, and the development, so long as it is only the disease itself that commands our attention, there is the possibility that we miss the full impact of the individuality of the patient, its multifariousness, its history, and its uniqueness. There is always a healthy side to a sick patient with its own proximate and remote causes and its diversity. And it is in noting those qualities that some clinicians are distinguished from others to whom the particularity of the patient lies only in his or her variant molecules. In fact, it is in appealing to the particularity of the whole patient, molecules included, that the physician is able to help the patient to discover and to mobilize resources that may make the difference between a timely and a delayed recovery, or even between life and death.

1.5.1.3 Why At This Time?

Perhaps the least understood by its victims is disease's apparent caprice in choosing when it strikes: the infant, blooming and full of promise, who wilts and dies; the robust, active college student who dies in 1 or 2 days of meningococcal meningitis; the busy, tireless 50-year-old who is felled by cancer. No doubt in old age disease and death are less anomalous, more expected, and yet we must ask why one person is privileged to die at 90 years, free of dementia, whereas others have died untimely. But we know that there are reasons for ages at onset of diseases and that they are accommodated within the nominalist definition of disease. These reasons are embodied in human mortality curves that, in the developed world, are U-shaped, declining sharply after birth, reaching a nadir at adolescence, and beginning to rise again in young adult life. If we were to include life before birth, the postnatal decline would be seen to be the end of a steep drop through intrauterine life.

Table 1.1 lists and contrasts qualities of the diseases experienced on the two sides of the U. How do we explain the differences? Again, appeal to remote causes gives the answers. All differences are those expected if the heritability of disease were to fall throughout life. The table tells us that indeed it does decline; the incidence of monogenic disease drops sharply before the nadir of the U, which suggests strong selection against disorders that imperil reproduction, leaving postpubertal disease to be associated mainly with the kind of genetic variation that is contingent, implicated in disease only in the presence of nongenetic proximate cause, and representative

of the kind of variation successful species exhibit. But, although the distribution of mortalities is U-shaped, the principle of continuity is not defied; monogenic disease continues to occur even in late life, and diseases of complex origin, such as congenital heart disease, are known in childhood, even *in utero* (Fig. 1.1).

Another way of perceiving in more detail the decline in prominence of the genetic impulse in disease is to express it as a decline in a gradient of selective effect. The gradient is not of genes but of phenotypes, and the weight of selection is heaviest *in utero* and least in old age. Burden is measured in risk to life, curtailment of reproduction, and permanent disability. These are biological burdens such that loss of life *in utero*, even before implantation, is more burdensome than the death of a bread-winning parent at age of 40 years with all its social upheaval. The apparent discrepancy in these burdens is a consequence of a natural ambiguity in human life in which we have two selves: one biological and the other social. It is as if we live two lives, one in obedience to

TABLE 1.1 Differences in Prepubertal and Postpubertal Diseases

	Prepubertal	Postpubertal
Mode of inheritance	Monogenic	Multifactorial
Age at onset	Early	Late
Frequency	Rare	Frequent
Latency	Short	Long
Affected relatives	Numerous	Few
Diagnostic specificity	High	Low
Number of diseases	Very many	Fewer
Burden	Great	Less
Sex differences	Occasional	Frequent
Influence of migration	No	Yes
Secular change	No	Yes
Effects of SES	Some	More
Success in treatment	Some	More
Heritability	High	Low
Predictability	High	Low

SES, socioeconomic status.

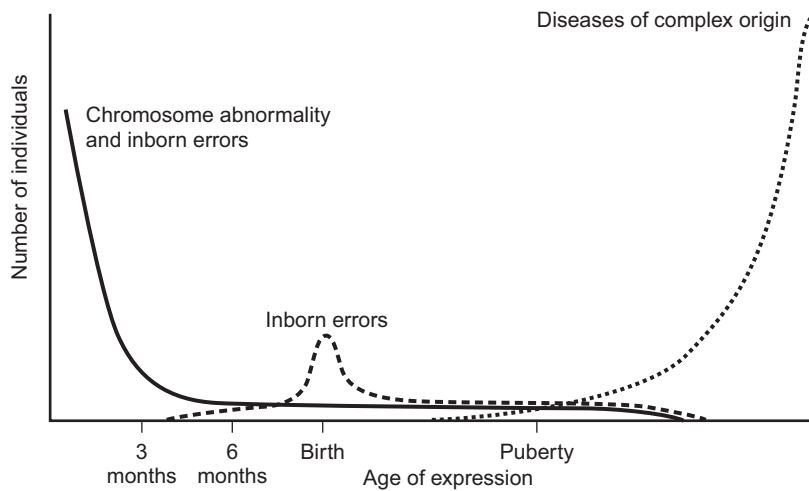


Figure 1.1 Continuity of disease across the lifetime.

the biological imperative to survive to reproduce and the other to exploit endowment and opportunity to have a fulfilling social life. Obviously these lives are intertwined, but when in opposition, there is a schism that may lead to disease.

The gradient is at its peak *in utero* where for reasons being elucidated, the majority of conceptuses are found wanting and die. Because intrauterine life is protected from the outside, most of this mayhem is likely to be genetic; for example, known early losses are most frequently associated with chromosomal anomalies [59]. But there must be inborn errors as well, involving among others, proteins specific to development. Then there are many known inborn errors with onset early in postnatal life and constituting a significant proportion of disease in newborns. In keeping with the biological desirability of a population of reproducers unencumbered by such mutants, 90% of single-gene diseases will have been disclosed by puberty and 99% by the end of reproduction [60]. As teenage passes into early adult life and that into midlife, the residual monogenic disease consists mainly of the cases of earliest onset of complex diseases, the most severe, life-threatening and resistant to treatment, whereas those with later onset, milder, more responsive to treatment are of complex origin, often irregularly familial or sporadic, a continuity commensurate with a high degree of heterogeneity of proximate cause. At the lowest end of the gradient are degrees of health expressed in resistance, completing the continuity that began with the mayhem *in utero*. Resistance to disease

is little noticed in medicine where appeals to the doctor are only made when the patient is sick, but it is well known in infections.

In keeping with the concept of declining heritability, we expect an increasing contribution of nongenetic variation. Most cancer takes its origin from mutation, but it is mainly somatic and so may be counted as of environmental origin, perhaps as an aspect of aging. Aging is perceived as dyshomeostasis, a product of cell loss and dysfunction due to denaturation of proteins. So while gerontologists, no doubt correctly, decline to see aging as disease, its expressions are associated with failure of the same gene products we associate with disease.

The decline in heritability is not monotonic but rises and falls according to developmental phase (see Fig. 1.1) [60]. For example, the genetically determined intrauterine disease has an early onset, whereas fetal disease late in pregnancy is more likely to be of maternal origin; prematurity, placenta previa, hypertension, premature placental detachment, and the like. Then there is a fresh spate of genetic disease in infancy, with deaths in childhood being more likely to be due to poisons, infections, accidents, and, in some places, homicide. In adult life, early-onset cases of complex disorders are strongly conditioned by genetic variation, whereas late-onset cases are likely to be nonfamilial [61].

Reasons for this phenomenon are conjectural, but it makes sense if the effects of new sets of genes are exposed to selection at the beginning of each developmental phase. It is an observation that needs study.

Another kind of continuity exhibited by the gradient of selective effect is the overlap of disorders characteristic of different developmental phases (see Fig. 1.1). For example, most chromosomal anomalies are lethal *in utero*, but for some there is an increase in live-born babies. But not all malformations are due to chromosomal aberration; some are of complex origin and some are harmless and may be recognized only at autopsy or emerge only under conditions of unusual stress, such as urinary tract infection in a patient with a single kidney, perhaps in adult life. Prepubertal life is characterized by monogenic disease, but type 1 diabetes, asthma, inflammatory bowel disease, and other disorders of complex origin sometimes have onset then, too. When a disease has onset over a broad range of ages, we must always wonder whether all victims to whose disorder we give the same name actually have the same disease. So in complex disorders, the problem of naming will be with us until we are able to sort out the heterogeneity. In the end, it seems likely that the number of diseases of adult life will be no less than those of the prepubertal years. No doubt genomics will identify the genes, and the issue of names will be resolved.

One reason why this description of the decline in heritability, with its stops and starts and overlaps, has been given prominence here is to illustrate the principle of the continuity of life, a continuity often ignored in both concept and organization. Hospitals, medical schools, and medical education are arranged around age and organ system for perfectly good reasons, and the system has proved its worth. But molecular genetics provides a continuity that brings the specialties together in taking a longitudinal view of human life.

1.6 PREVENTION AND TREATMENT

One might suppose a priori that the Mayr model could not accommodate prevention and treatment of human disease. Biology is concerned with what is and does not accommodate intervention; the minute one intervenes, what is no longer is. And what could be more unnatural than government agencies that regulate how we deal with nature, or surgical transplantation of organs as a treatment, to say nothing of the idea of designer drugs for molecular defects. But we cannot escape our biological heritage, nor do we wish to, so perhaps the model is apposite after all.

All organisms are capable of adapting to, or otherwise defending themselves against, uncongenial environments. Some call on homeostatic flexibility when under stress, whereas others move to evade it; evolution has seen to these self-protective capabilities. For example, many organisms have molecular mechanisms to withstand stress; heat shock proteins are one [62]. Others are enzymes that detoxify foreign substances, and a third is up- and down-regulation of metabolic systems. Some organisms remove themselves from threats. In addition, animals that can, choose surroundings appropriate for their physiology and improve them, too [63]. So, in the sense that they know what experiences to avoid, animals practice prevention, and when attacked by disease or hurt in the course of the day's work, they fall back on natural mechanisms by way of treatment. The difference between them and us is that we consciously intervene in both. But in the degree to which we seek out proximate causes intending to alter them by prevention or treatment, and in the degree to which we try consciously to influence social and cultural conditions with an eye to changing remote causes, we do fulfill the expectation of the model.

1.6.1 Prevention

Constituted separately from medicine and described as improvement in diet, housing, and other living conditions, prevention has saved more lives than treatment [64]. And when microorganisms were identified as proximate causes, they became the target of prevention by quarantine and immunization; the latter remains a staple of medical care. These preventive measures were and are in the hands mainly of local government, whereas progress in fostering the ideas and promoting education in preventive medicine were and are the work of university-based schools of public health and hygiene. After the 1950s, when antibiotics reduced the mortality of infections, other diseases such as diabetes, cancer, and kidney disorders came into focus. Then the ideas of preventive medicine and epidemiology, which had been restricted to the control of infectious diseases in populations, began to include prevention of noninfectious disease in individuals, leading to the establishment of organizations that included patients, their families, and the public, and that were devoted to education and counseling of patients and relatives as well as the general public with the intention to prevent and to learn enough to treat these common disorders. The American Cancer Society, the American

Diabetes Society, and The March of Dimes come readily to mind. In the 1960s and 1970s as more and more inborn errors were described, this principle was also applied to the generation of a multitude of disease-related societies, each dedicated to education, treatment, and prevention of one disease, the latter in the form of reproductive counseling, antenatal diagnosis, and sometimes abortion. Then, as the molecular basis of these disorders was discovered, newborn screening for inborn errors was offered by many state health departments and intensive studies were undertaken of every aspect of this form of preventive medicine including screening, counseling, and issues both ethical and legal [55,65–69]. The question then arose of testing relatives of patients with inborn errors with an eye to reproductive advice, and the triumph of Tay–Sachs testing is one result [14,55]. And now that rapid progress is being made in unraveling the genomics, epigenomics, and proteomics of complex disease, time will give us more risk factors in the way of variant genes and proteins. These developments are reviewed here in this detail to call attention to the movement of the focus of prevention away from populations to individuals, and now to the molecular emphasis in both prevention and treatment. Just as the discovery of genes associated with disease suggests the possibility of cure, so does it suggest prevention by testing of relatives and populations. Indeed, the logic of prevention is even more powerful than that of cure. That is, unlike treatment, which is always after the fact and is occasionally as threatening as the disease it is designed to combat, prevention spares the organism such rigors even while far less disruptive of social and economic life. On the other hand, in keeping with the principle of continuity, the two are sometimes indistinguishable.

So, may we expect miracles of prevention now that we can identify proximate causes? Readers of PPMGG know that we may not [14,15,56,57]. It is a matter of the continuity of the gradient of selective effect. At one end, the virtual elimination of Tay–Sachs disease among Jews and the prevention of a few other inborn errors by the same means represent successes of the high technology promoted by Lewis Thomas [70]. At the other end are healthy centenarians who attribute their robust health to some idiosyncratic behavior. But in between are those genes and their variant products whose virtues are a sometime thing, depending on, on the one hand, the specificities of experiences over the lifetime, and on the other, their support, or reinforcement in failure, by the variant products of other genes. So the same variant gene product may be

adequate in one person and fail in another even in the same family. Or it may be within the same person adequate under one circumstance and insufficient under another. Thus, as a predictor, a gene may be of only limited use to an individual even while accepted as a significant risk in a population. This is a frequent problem of epidemiologically designated risk factors; it is not always clear to whom among their possessors the trait is actually risky—to say nothing of gradations in risk. It is the problem also of evidence-based medicine, which, however, valuable in increasing the rigor of diagnosis and treatment provides recommendations suitable for populations, not individuals [71–73]. It is a matter of typological, as opposed to population, thinking. Of course, the HGP has added greatly to the list of our genes and their proteins so that the exact identity of all of the units in pathways and other homeostatic devices will be known, improving thereby the predictive value of various combinations of variants [74]. And, assuming increasing identification of exterior proximate causes, the accuracy and usefulness of preventive predictions may improve remarkably. The necessity for the advancement of knowledge of nongenetic proximate causes cannot be exaggerated. We need projects of similar scope and ambition to that of the HGP. In the meantime, we should do what we can where we can, and for the rest, fall back on an aspect of medicine that may have become unfashionable in modern times but that is perhaps more than ever needed: helping patients to live with uncertainty.

A far less likely, but more effective, means of health promotion is the control of remote causes. The virtue of such an approach is clearly indicated in the Mayr model wherein the relationship of the two kinds of causes and the two kinds of biology is so lucidly stated. To influence by law the distributions of genes is both unconstitutional and in strong opposition to “liberty and the pursuit of happiness,” but to influence the organization of society and culture for the betterment of health is not only possible but also already the aim of numerous government agencies, the even more numerous private disease-related societies, as well as physicians who have advised their patients to practice healthy ways. But what has not been emphasized, at least in the United States, is the power of corporate action, of putting the weight of the whole medical profession socially and politically on the side of health promotion. Today in the United States this is impossible, and some will say it is undesirable, but the point to be made here is the logic of the position. On one side is the

power of remote causes in the origin of disease (let us think here of the evolution, growth, organization, penetration, and political powers of the tobacco industry) and on the other is the power of societies to organize themselves to influence those remote causes. In the United States, it was public opinion followed by legal action that began the descent of the authority of tobacco. There are other social conditions likely to retard the advance of prevention. One of these is the combination in the public mind of a superficial grasp of progress in biology and genetics and an unreasoning belief in the limitless potential of that progress. But for a recipient to respond realistically to the offer of prevention requires the ability [1] to differentiate between personal and populational probabilities [2], to grasp its potential for success or failure, and [3] to participate constructively with a knowledgeable and sympathetic physician in greeting either success or failure. We do not often think of the evolution of education, or of the public grasp and acceptance of advances in medicine as remote causes of success or failure, but we should.

1.6.2 Treatment

The essence of Lewis Thomas' concept of high technology in treatment lay in the discovery of the exact point or points in the machine that were broken and that could be repaired by a single, simple, straightforward maneuver [70]. One of his examples of such a treatment was the use of steroid in the adrenogenital syndrome. And after the 1950s, amid the rapid accumulation of newly described inborn errors, there was optimism that such diseases would be brought under control [75]. But results so far have been something less. In the 1980s, Hayes tabulated successes and failures in 65 inborn errors, a part of a larger randomized sample of monogenic diseases taken from MIM 5 [76]. Table 1.2 shows that for 12%, treatments were successful in rendering the patient normal or

essentially so, whereas in 40% there was some improvement, often not very impressive, and about 48% showed no success at all. A further examination of success or failure in treatment of the same 65 diseases 10 years later was reported by Treacy, with results shown in Table 1.2 [77]. There was no increase in the number of very successful treatments, but some of the previously resistant disorders now yielded in some degree, often to such rigorous therapies as tissue and organ transplantation. These, Thomas saw as middle technologies, sometimes effective, but expensive and perhaps hard on the patient. A third look in 1999, this time including 517 inborn errors listed in the seventh edition of MMBID, gave much the same results (see Table 1.2) [78].

Given the qualities of the inborn errors that were the object of treatments, the record is perhaps not surprising. Most are at or near the top of the gradient of selective effect, some are lethal, and some are permanently crippling. And all are heterogeneous, some as to loci, all as to alleles. And we now know that we must expect equal heterogeneity among those shadowy modifiers we presume to exist [17]. So these disorders are simply the most intractable. But farther down the gradient, the diseases are more amenable, not necessarily to cure, but certainly to management. This seems to be telling us that there is a relationship between heritability and success in therapy. When the heritability is high in a population, we expect less of treatment than when it is low (i.e., fewer patients are likely to respond satisfactorily). The history of treatment of rickets with vitamin D is exemplary. When, after the mid-1940s rickets almost disappeared, nearly all that was left were several different kinds of monogenic vitamin D-resistant rickets [19]. This experience seems to furnish medicine, especially preventive medicine, an aim, even a motto. We work to drive the heritability of disease toward 1.0. And we fervently hope that the gene therapists will confound the motto by inventing high-technology treatments that subdue even the most "genetic" and even the most refractory of those disorders that continue to resist every effort to contain them. As for the complex disorders, they are resistant, too, but in a different way. If every individual has his or her own set of proximate causes, the complexity is of a high degree. The problem is one of discovery of the nongenetic causes and trying to eliminate them, as well as discovering which sets of genetic causes pose vulnerability to the threats of those nongenetic influences, both in general and in each affected individual. No easy job,

TABLE 1.2 Effectiveness of Treatment of Inborn Errors

	1983 Hayes	1993 Treacy	2000 Treacy
Fully beneficial	12 ^a	12	12
Partially beneficial	40	57	54
No benefit	48	31	34

^aPercentage of total.

but we are up against a wily opponent. Many years ago, Max Delbrück observed that “any living cell carries with it the experiences of a billion years of experimentation by its ancestors. You cannot expect to explain so wise an old bird in a few simple words” [79]. We are definitely embarked on an effort to expose that wisdom. The next two or three editions of PPMGG should show how wise the old bird is.

Certainly the progress in the past few years in genetically modifying human T-cells to attack various malignancies, the therapy of cystic fibrosis directed at specific mutations in *CFTR*, and the insertion of a normal gene to cure blindness are great examples.

1.7 CONCLUSION

The explanatory principles of disease begin with the capacity of the species for genetic variability, a capacity that is required for survival of species and that is experienced randomly resulting in genes whose products have, through time, conferred on their recipients a status of congruence with equally variable environments. But sometimes the result is incongruence, which can lead to disease. It is in the gene products, the protein unit steps of homeostasis, that this species variability is expressed in congruence or incongruence in health or disease. This expression occurs within a biochemical and molecular cellular matrix conditioned by interaction between such protein products and experiences of the environment through development, maturation, and aging. Accordingly, analysis of pathogenesis must be pursued in three timescales all at once: that of phylogeny whence the genes and their products were derived; that of ontogeny, maturation and aging, which condition the ever-changing matrix; and that of the moment representing the impact of today's events. This principle, embracing the three timescales, also incorporates two kinds of causes, proximate and remote, which are expressed in the uniqueness of individuals. The incongruent proximate causes, variable protein unit steps of homeostasis and varying kinds, amounts and durations of experiences of the environment, account for the expressions of disease phenotypes that are subjected to selection and incur the social stigmas that complicate the lives of their victims. Remote causes are composed of the evolution and dynamics of both biological and social milieu that account for the nature and local availability of proximate causes, their unique assembly as genotypes and availability as experiences, to form

combinations favorable for disease. And it is this particularity that determines who gets which disease at what time in life. The qualities of diseases are expressions of unique and variable human genomes arranged in a gradient of selective effect, a representation of the removal in early life of those unlikely to reproduce, and in post-reproductive life, a less-intense test of survival in a variable environment. It is in the latter part of the range of the gradient that both prevention and treatment are likely to be most effective; prevention because changes in environment can be effective in avoiding disease, and treatment because the homeostasis can be characterized as inefficient and in need of a boost, rather than broken. The logic of prevention is more powerful than that of treatment, but we need both a more comprehensive knowledge of nongenetic proximate causes and, in time, to learn, understand, and adjust to the social dislocations any sudden spate of preventions could bring. But it is in part in the grasp of the possibility and plausibilities of both prevention and treatment, and in part in understanding the meaning in medicine of individuality and the virtues of population thinking in relation to it, that we may be able at once to pursue the reductionist path we have so successfully traversed and return to embrace the integration, the humanity, of patients who appeal to us for relief of both the consequences of their molecular incongruities and the injury of the disease to that integrated humanity.

REFERENCES

- [1] De Solla Price DJ. Little science big science. New York: Columbia University Press; 1963.
- [2] The Online Metabolic and Molecular Basis of Metabolic Disease. <https://ommbid.mhmedical.com/book.aspx?bookID=971>.
- [3] Rosenberg CE. Banishing risk: or the more things change the more they remain the same. *Perspect Biol Med* 1995;39:28–42.
- [4] Lander ES. Initial impact of the sequencing of the human genome. *Nature* 2011;470:187–97.
- [5] Childs B. The entry of genetics into medicine. *J Urban Health* 1999;76:497–508.
- [6] Cohen H. The evolution of the concept of disease. In: Lush B, editor. *Concepts of medicine*. Oxford: Oxford University Press; 1960.
- [7] King L. *Medical thinking*. Princeton: Princeton Press; 1982.
- [8] Temkin O. The scientific approach to disease: specific entity and individual sickness. In: Crombie AC, editor. *Historical studies in intellectual, social and technical*

- conditions for scientific discovery. New York: Basic Books; 1963. p. 629–47.
- [9] Vogel F, Motulsky AG. Human genetics. Preface. 3rd ed. New York: Springer-Verlag; 1998.
 - [10] Mayr E. Cause and effect in biology. *Science* 1961;134:1501–6.
 - [11] Mayr E. The growth of biological thought. Boston: Harvard University Press; 1982. p. 45–7.
 - [12] Burley SK, Almo SC, Bonnano JB, et al. Structural genomics: beyond the human genome project. *Nat Genet* 1999;23:151–7.
 - [13] Lander ES, Weinberg RA. Genomics: journey to the center of biology. *Science* 2000;287:1777–82.
 - [14] Andrews LB, Fullarton JE, Holtzman NA, Motulsky AG. Assessing genetic risks. Washington, DC: National Academy Press; 1994.
 - [15] Motulsky AG. If I had a Gene Test, What would I have and who would I tell? *Lancet* 1999;354:SI35–7.
 - [16] Scriver CR. Changing heritability of nutritional disease: another explanation for clustering. In: Simopoulos A, Childs B, editors. Genetic variation and nutrition. New York: Karger; 1989. p. 60–71.
 - [17] Scriver CR. Monogenic traits are not simple. *Trends Genet* 1999;15:3–8.
 - [18] Hill AVS. Genetics and genomics of infectious disease susceptibility. *Br Med Bull* 1999;55:401–13.
 - [19] Scriver CR, Childs B. Garrod's inborn factors in disease. New York: Oxford Press; 1989.
 - [20] Fox-Keller E. Language and science: genetics, embryology and the discourse of gene action. In: Fox-Keller E, editor. Refiguring life. New York: Columbia University Press; 1995.
 - [21] Beadle GW. Genes and chemical reactions in Neurospora. *Science* 1956;129:1715–9.
 - [22] Garrod AE. The incidence of alkaptonuria: a study in chemical individuality. *Lancet* 1902;2:1616–20.
 - [23] Pauling L, Itano HA, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. *Science* 1949;110:543–8.
 - [24] Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 1998;92:291–4.
 - [25] Gelbert WM. Databases in genomic research. *Science* 1998;282:659–61.
 - [26] Kohler RE. Lords of the fly. Chicago: University of Chicago Press; 1994.
 - [27] Capecchi MR. Hox genes and mammalian development. Cold Spring Harbor Symp Quant Biol 1997;62:273–8.
 - [28] Kornberg TB, Krosnow MA. The Drosophila genome sequence: implications for biology and medicine. *Science* 2000;287:2218–20.
 - [29] Stent GS. Strengths and weaknesses of the genetic approach to the development of the nervous system. In: Cowan WM, editor. Studies in developmental neurobiology. New York: Oxford University Press; 1981.
 - [30] Waddington CH. The strategy of the genes. London: Allen and Unwin; 1957.
 - [31] Marmot MG. Early life and adult disorder. *Br Med Bull* 1997;53:3–9.
 - [32] Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999;22:231–8.
 - [33] Halushka MK, Fan JB, Bentley K, et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood pressure homeostasis. *Nat Genet* 1999;22:239–47.
 - [34] Bearn AG. Archibald Garrod and the individuality of man. Oxford: Oxford University Press; 1993.
 - [35] Garrod AE. Inborn errors of metabolism. Oxford: Oxford University Press; 1909.
 - [36] Fruton JS. Molecules and life. New York: Wiley-Interscience; 1972.
 - [37] Barker LF. Heredity in the clinic. *Am J Med Sci* 1927;173:597–605.
 - [38] Scott-Moncrieff R. The classical period in chemical genetics. Notes and records. *R Soc London* 1981–1983;36–37:125–54.
 - [39] Wright S. Color inheritance in mammals. *J Hered* 1917;8:224–35.
 - [40] Cori GT, Cori CF. Glucose-6-Phosphate of the liver in glycogen storage disease. *J Biol Chem* 1952;199:661–7.
 - [41] Isselbacher KJ, Anderson EP, Kurahashi K, Kalckar HM. Congenital galactosemia, a single enzyme block in galactose metabolism. *Science* 1956;123:635–6.
 - [42] Yanofsky C. Gene structure and protein structure. *Harvey Lect* 1965–1966;61:145–67.
 - [43] Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037–48.
 - [44] Maynard Smith J, Burian R, Kauffman S, et al. Developmental constraints and evolution. *Q Rev Biol* 1985;60:265–87.
 - [45] Mayr E, Provine WB. The evolutionary synthesis. Cambridge: Harvard University Press; 1980.
 - [46] Rosenberg CR. Toward an ecology of knowledge. In: The organization of knowledge in modern America. Baltimore: The Johns Hopkins Press; 1979. p. 440–55.
 - [47] Schroedinger E. What is life?. Cambridge: Cambridge University Press; 1967. p. 23.
 - [48] Garrod AE. Inborn errors of metabolism. 2nd ed. Oxford: Oxford University Press; 1926.
 - [49] Nelkin D, Lindee MS. The DNA mystique. New York: WH Freeman; 1995.
 - [50] Collins FS. Shattuck lecture: medical and societal consequences of the human genome project. *N Engl J Med* 1999;341:28–37.
 - [51] Jiminez-Sanchez G, Childs B, Valle D. The effect of mendelian disease on human health. In: Scriver CR,

- Beaudet AL, Sly WS, Valle D, editors. The metabolic and molecular basis of inherited disease. 8th ed. New York: McGraw-Hill; 2000. p. 167–74.
- [52] Kirschner M, Gerhart J. Evolvability. *Proc Natl Acad Sci USA* 1998;95:8420–7.
- [53] Rutherford S, Lindquist S. Hsp 90 as a capacitor for morphological evolution. *Nature* 1998;396:336–46.
- [54] Ly DH, Lockhart DJ, Lerner RA, Shultz PG. Mitotic misregulation and human aging. *Science* 2000;287:2486–92.
- [55] Anon. Genetic screening. Washington, DC: National Academy of Sciences; 1975.
- [56] Holtzman NA. Proceed with caution. Baltimore: The Johns Hopkins Press; 1989.
- [57] Holtzman NA, Watson MS. Promoting safe and effective genetic testing in the United States. Baltimore: Johns Hopkins University Press; 1998.
- [58] Fries JF, Crapo LM. Vitality and aging. New York: WH Freeman; 1981.
- [59] Jacobs PA. The role of chromosome abnormalities in reproductive failure. *Reprod Nutr Dev* 1990;30(Suppl):63s–74s.
- [60] Costa T, Scriver CR, Childs B. The effect of mendelian disease on human health: a measurement. *Am J Med Genet* 1985;21:231–42.
- [61] Childs B, Scriver CR. Age at onset and causes of disease. *Perspect Biol Med* 1986;29:437–60.
- [62] Hoffmann AA, Parsons PA. Evolutionary genetics and environmental stress. New York: Oxford University Press; 1991.
- [63] Lewontin RC. Gene, organism and environment. In: Bendell DS, editor. *Evolution from molecules to men*. New York: Columbia University Press; 1983. p. 273–86.
- [64] McKeown J. The role of medicine. London: Nuffield Trust; 1976.
- [65] Bergsma D. Ethical, social and legal dimensions of screening for human genetic disease. *Birth Defects Orig Artic Ser* 1974;10(No. 6):1–272.
- [66] Burnham JC. America and Medicine's golden age. What happened to it? *Science* 1982;215:1474–9.
- [67] Hsia YE, Hirshhorn K, Silverberg RL, Godmillow L. *Counseling in genetics*. New York: A.R. Liss; 1979.
- [68] Knoppers BM, Labarge CM. Genetic screening: from newborns to DNA typing. Amsterdam: Excerpta Medica; 1990.
- [69] Lubs HA, de la Cruz F. Genetic counseling. New York: Raven Press; 1977.
- [70] Thomas L. The future impact of science and technology on medicine. *Bioscience* 1974;24:99–105.
- [71] Mant D. Can randomized trials inform clinical decisions about individual patients? *Lancet* 1999;353:743–6.
- [72] Sweeney KG, MacAuley D, Gray DP. Personal significance: the third dimension. *Lancet* 1998;351:134–7.
- [73] Tonelli MR. The philosophical limits of evidence based medicine. *Acad Med* 1998;73:1234–40.
- [74] Van Omenn GJB, Bakker E, den Dunnen JT. The human genome product and the future of diagnostics, treatment and prevention. *Lancet* 1999;354(Suppl):-si5–10.
- [75] Scriver CR. Treatment in medical genetics. In: Crow JF, Neel JV, editors. *Proceedings of the 3rd international congress on genetics*. Baltimore: Johns Hopkins Press; 1967. p. 45.
- [76] Hayes A, Costa T, Scriver CR, Childs B. The effect of mendelian disease on human health. II: response to treatments. *Am J Med Genet* 1985;21:243–55.
- [77] Treacy E, Childs B, Scriver CR. Response to treatment in hereditary metabolic disease: 1993 survey and 10 Years comparison. *Am J Hum Genet* 1995;56:359–67.
- [78] Treacy EP, Valle D, Scriver CR. Treatment of genetic disease. In: Scriver CR, Beaudet AL, Sly WS, Valle D, editors. *The metabolic and molecular basis of inherited disease*. 8th ed. New York: McGraw-Hill; 2000.
- [79] Delbruck M. A physicist looks at biology. In: Cairns J, Stent GS, Watson JD, editors. *Phage and the origins of molecular biology*. Cold Spring Harbor, NY: Cold Spring Harbor Press; 1966. p. 9–22.

Foundations and Application of Precision Medicine

Geoffrey S. Ginsburg, Susanne B. Haga

Center for Applied Genomics & Precision Medicine, Duke University School of Medicine,
Durham, NC, United States

The goal of precision medicine is to optimize disease prevention, diagnosis, and treatment decision-making based on comprehensive information that incorporates traditional clinical measures, and data with DNA variation (genome), gene expression (RNA or transcriptome), proteins (proteome), metabolites (metabolome), methylation (epigenetics), and/or microbial composition (microbiome). These patient data can provide increased accuracy in assessment of disease risk, diagnosis, prognosis, and drug response. The growth of precision medicine has been possible, in part, due to the convergence of a societal shift toward patient-centered care, transition to electronic medical records (EMRs), and development of decentralized digital health technologies. The integrated use of these technologies, such as linking digital health tools to EMRs, improved patient engagement through online patient portals, and improved portability of medical data, further enables rapid implementation and utilization of precision medicine. In particular, the successful integration of -omics information and technologies into clinical practice is reliant on a robust and secure clinical infrastructure to store and analyze large data sets.

Although much attention has focused on the individual, the field of precision medicine can benefit population (or public) health through the combination of individual-level data and population-based interventions, implementation of genetically targeted approaches such as newborn screening or follow-up testing for some inherited cancers, and application of new technologies to existing public health efforts such as infectious

disease surveillance [1]. Preventive medicine can move beyond standard recommendations for health behaviors and work toward tailored lifestyle recommendations to patients' needs and circumstances [2,3]. With the move beyond traditional academia-based initiatives into health systems and exploration of the use of genome technologies in healthy populations, the union of precision medicine and population health stands poised to affect large populations in practical clinical settings [4].

Much work lies ahead to understand the clinical significance of the multiple types of data that can now be generated for a given person, to validate new testing platforms, and to develop evidence-based practice guidelines. We have already witnessed the development and implementation of several new interventions and treatments [5]. In this chapter, we will provide an overview and examples to date that have contributed to the rise of precision medicine and highlight areas where effort is needed to realize the full potential of these applications.

2.1 OVERVIEW OF PRECISION MEDICINE

Since the announcement of the Human Genome Project in the 1990s and its completion in 2003, there have been significant advances in medicine, particularly regarding our understanding of risk factors for chronic disease and their molecular basis. Before the sequencing of the human genome and development of a reference human genome sequence [6,7], the process of identifying the genetic basis of disease was a laborious and time-consuming process. The completion of a reference human

genome sequence was made possible with substantial developments in sequencing technology, data storage, and bioinformatics and analytical capabilities (which continue to grow to this day [8–10]). The availability of new, rapid, and less costly sequencing and other genomic technologies has led to an ongoing wave of discoveries about the causes of and contributors to rare and complex diseases, respectively, drug targets, and clinical diagnostics [11].

Genomics and precision medicine has focused on the study of large groups or populations with complex, multifactorial conditions aiming to molecularly sub-stratify them, whereas the field of human genetics has primarily focused on families with rare inherited conditions. While one of the major goals of precision medicine is to provide more individualized care, precision medicine may also benefit population health through improved screening interventions and disease prevention for healthy populations [4,12,13]. However, greater efforts are needed to generate comprehensive assessments of the impact of gene–environment interactions and examination of other molecular data sets in order to develop better risk predictors and diagnostic and prognostic markers and to inform treatment decisions [1,14,15]. Large-scale national initiatives such as the All of Us Research Program in the United States [16] and the 100,000 Genomes Project in the UK Project [17] will enable more comprehensive data collection and analysis to gain a better understanding of risk factors as well as to raise public awareness and widespread incorporation of genomics into health and medicine.

2.2 PRECISION MEDICINE APPLICATIONS ACROSS THE LIFESPAN AND CLINICAL SPECIALTIES

Precision medicine applications are not targeted to a single disease or patient population but rather can improve well-being and health outcomes for both affected and ill patients across the lifespan (see Fig. 2.1). Complex diseases have long been understood as the culmination of gene and environmental factors. Thus, several types of data sets generated from more than one of these technologies will be essential to understanding disease risk and health outcomes [18,19]. Indeed, the combination of static (genome) and dynamic (RNA, proteins, metabolites, and microbiome) measures has enabled a comprehensive analysis of the interactions between host and

microbiome, environmental exposures, and the resulting impact on gene expression and the proteome. The “exposome” or “exposotypes” may only be interpreted through multiple –omic analyses in order to comprehensively evaluate the impact of an exposure on human physiology and health risk [20]. For example, an analysis of several –omics data sets captured over a 14-month period from a single individual highlighted their temporal dynamics associated with changes in environmental exposures (including infection) and lifestyle habits [21]. Another study collected –omics data, clinical tests, and lifestyle data for 108 individuals over a 9-month period and tailored health coaching based on each patient’s comprehensive risk assessment [22].

Preconceptional and Prenatal Screening. Today, genetic and genomic testing is commonly used in obstetrics and pediatrics. With the rise of in vitro fertilization, preimplantation genetic diagnosis (PGD) provides couples with an option to selectively implant unaffected embryos [23]. Prenatal diagnosis has also been revolutionized with early noninvasive screening of fetal cells isolated from maternal blood for chromosomal anomalies, enabling much earlier results in pregnancy [24,25]. Known as noninvasive prenatal testing (NIPT), the test is recommended for high-risk pregnancies to screen for chromosomal 13, 18, and 21 aneuploidies [26,27]. However, evidence suggest that NIPT performs well in routine (low-risk) pregnancies [24,28], motivating some obstetrics practices to offer NIPT to all women. Although the number of invasive procedures (chorionic villus sampling and amniocentesis) has decreased [29,30], invasive procedures are still necessary to confirm positive NIPT results and in cases with abnormal ultrasonography results to evaluate for other chromosomal abnormalities [31]. In a few cases, NIPT has led to the diagnosis of maternal cancer due to sample contamination with tumor cfDNA [32,33]. Studies are under way to assess the use of fetal cfDNA for single gene analysis, targeted microarray analysis [34], and whole exome or genome sequencing [35,36]. For prenatal diagnosis, chromosomal microarray (CMA) analysis (instead of conventional karyotyping) is recommended for fetuses with a structural abnormality or developmental delay observed by ultrasound examination [37].

Pediatrics. Genetic screening begins at birth with newborn screening tests for a suite of conditions whose course can be substantially affected by early interventions. Whole genome sequencing has also been shown

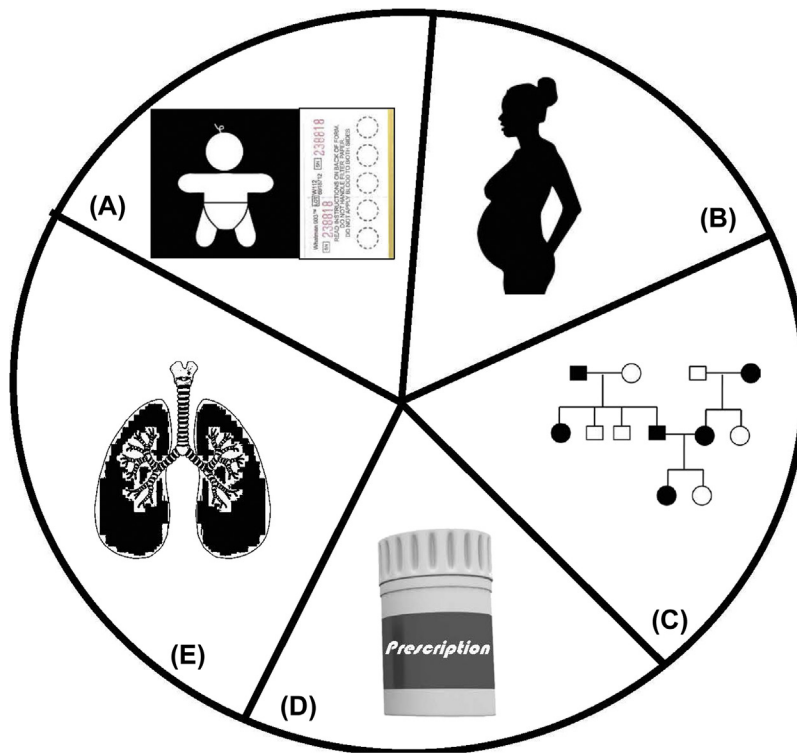


Figure 2.1 Examples of life stages and conditions for which precision medicine applications are used: (A) newborn screening, (B) prenatal screening and diagnosis, (C) family health history, (D) medication use, and (E) lung cancer.

to be useful in establishing a diagnosis for newborns affected with severe congenital malformations or other undiagnosed health issues requiring admission to a neonatal intensive care unit. In these urgent care circumstances, sequencing can be completed in less than 2 days to inform treatment and care decisions [38] and has been shown to be cost effective [39]. For other common conditions that develop during early childhood, CMA is recommended for developmental delay/intellectual disability, congenital anomalies, and syndromic conditions [40–42], and several types of genetic tests have been developed, including panel testing for epilepsy [43–45].

Risk Assessment and Family Health History. The clinical benefits of a comprehensive and accurate family health history (FHH) have been well demonstrated, but use of these data has been low and uneven [46]. Barriers to collection and use include time, lack of confidence to interpret, and inexperience with ordering genetic testing if indicated by the FHH [47]. A variety of new tools and approaches to improve FHH collection and utilization

are under investigation, including online FHH tools [48] and virtual counseling [49].

Use of online FHH tools can improve the accuracy and comprehensiveness of FHH [50], in turn altering patient risk perceptions and health behaviors [51], improving patient care, and increasing identification of at-risk family members of patients affected with inherited conditions. Cascade screening involves offering testing to family members at risk for a hereditary condition [52]. Cascade screening has often been performed for hereditary cancer syndromes such as Lynch syndrome [53] and hereditary breast and ovarian cancer, familial hypercholesterolemia [54,55], and other rare conditions such as long QT syndrome [56]. However, there is a lack of consensus on cascade screening practices, and other barriers, such as family communication [57] or other strategies to notify family members [58,59], and a low rate of follow-up care of at-risk family members [60], pose challenges to promptly identifying and providing care to at-risk family members.

The use of whole genome or exome screening in healthy populations is under active investigation in both adult and pediatric populations [61–63]. Other projects are focused on evaluating the feasibility and utility of sequencing various patient populations, such as the MedSeq project for primary care and cardiovascular patients [64], the ClinSeq project for patients affected with atherosclerotic heart disease [65], and infants in neonatal intensive care units [66,67]. Outside of the clinical setting, direct-to-consumer testing companies provide genetic testing for several genetic risks [68]. For at-risk populations, several groups have generated genetic risk scores based on the analysis of multiple genetic variants for myocardial infarction [69], prostate cancer [70], and obesity [71,72]. The combination of genetic risk score and FHH can further improve likelihood estimates [73,74].

Oncology. In oncology, for both children and adults, the growing list of single-gene tests is being replaced by large gene panels. Furthermore, comprehensive somatic genotyping or sequencing of tumor tissue compared to normal tissue has enabled identification of key mutations relevant to disease stages and potentially informative for treatment decisions [75,76]. In addition, to DNA-based gene profiles for cancer susceptibility [77,78], gene expression profiling (polymerase chain reaction [PCR] based) has been used to generate risk recurrence scores in breast cancer patients [79] and determine origin of cancer of unknown primary [80]. As discussed later in the section on drug development, the majority of targeted drugs are for oncology, and new personalized cellular therapies warrant care genetic interrogation to appropriately inform treatment selection and management.

Cardiovascular Diseases. For cardiovascular diseases and the multitude of risk factors at play, the addition of genetics provides another source of risk prediction, but the gene–environment interactions remain ambiguous. Gene expression profiling for heart transplant patients (Allomap) has enabled the prediction of acute cellular rejection that was heretofore detected by an invasive procedure (percutaneous transvenous endomyocardial biopsy) after damage had occurred [81–83]. In addition, increase in Allomap scores between 6 and 9 months after transplantation has been associated with long-term survival [84]. Another gene expression profiling test (Corus CAD) for coronary heart disease (CAD) in nondiabetic patients has been shown to be effective in identifying

patients with obstructive CAD and in reducing unnecessary invasive procedures for patients with nonobstructive CAD symptoms [85,86]. Several types of gene variant panels are under investigation to predict risk or recurrence of myocardial infarction [87], atrial fibrillation [88–90], lipoprotein levels [91], and coronary heart disease [92–94]. As evidence of the utility of genetic risk scores increases [95], their use in the clinical setting has begun to be evaluated [96].

Pharmacogenetic Testing. Pharmacogenetic (PGx) testing, or the analysis of genes known to affect drug response, is considered one of the leading applications of precision medicine [97]. Because adverse drug response is one of the leading causes of hospitalizations [98], a better understanding of a patient's PGx risk of severe drug response may reduce the cost and burden of hospitalizations [99,100]. However, drug response is a complex phenotype and, as such, many factors can influence the outcome [101]. The expressivity of a genetic variant may be altered by diet, age, comorbidities, or other genetic variants. This may account for some of the equivocal trial findings published regarding the clinical benefits of PGx testing [102,103]. Thus, with the exception of some highly penetrant variants, PGx test results should be combined with other information in decisions regarding treatment selection and dosing. Advances in understanding the impact of PGx variants on medications used in several clinical specialties have been made, particularly in psychiatry [104,105], infectious disease [106], pain management [107,108], and cardiology [109,110] (we excluded oncology here as it often involves testing for drug targets discussed elsewhere). In addition, certain patient populations may gain greater benefits from PGx testing, such as polypharmacy patients [111,112].

Integration of PGx testing into clinical care may occur in more than one approach [113]. Testing may be ordered at the point-of-care when treatment is needed. The specific test for that specific medication or medication class is then ordered. Alternatively, to avoid delay in obtaining PGx test results at the time treatment is needed, ordering testing preemptively will enable the results to be available when treatment is needed. With advances in testing technology, a range of several types of PGx tests are commercially available that include genes known to affect commonly prescribed medications. Thus, a comprehensive PGx panel test may be ordered preemptively and the results stored in the

patient's medical record for future use. Several medical centers have begun PGx preemptive testing programs to assess feasibility, implementation barriers, and provider utilization [114,115]. A hybrid approach may be for a physician to order a comprehensive panel test at the point-of-care, which would inform current and future treatments. Provider knowledge, perceived utility, uneven reimbursement coverage, and lack of clinical guidelines among other factors have hamstrung routine use of PGx testing [116–118].

2.3 PRECISION DRUG DEVELOPMENT

Developing and prescribing safe and effective treatments remain major challenges. Drug trials have evolved in light of the voluminous genomic and other data sets now available to improve characterization of disease phenotype and host response [119,120]. Genome-wide studies, particularly in cancer, have identified key mutations associated with growth and survival, providing new targets for drug developers. Use of genetically defined targets could improve the success rate and reduce cost of clinical trials [121]. Adaptive study designs whereby research participants may switch between intervention arms based on accumulating response data [122,123] and basket trials where participants undergo molecularly screening to determine study eligibility are increasingly being used [124,125]. Evidence generated from such novel trial designs has contributed to drug approvals [126].

In the past two decades, numerous drugs have been approved for the treatment of conditions characterized by specific genetic indications (Tables 2.1 and 2.2). Starting with the approval of trastuzumab in 1998, for women with *HER2/Neu*-positive breast cancer, and imatinib mesylate in 2001, for patients with *BCR-ABL*-positive chronic myeloid leukemia, the number of targeted drugs has steadily increased as the clinically relevant genes in tumor growth have been identified. In some cases, multiple generations of targeted drugs, such as crizotinib (2011), alectinib (2015), and ceritinib (2017), have reached the market through an accelerated approval path, all for *ML4-ALK*-positive non-small cell lung cancer (NSCLC) patients, with the latter two as second-generation drugs for patients intolerant to crizotinib [127]. In 2018, the first drug, olaparib (Lynparza), a poly(ADP)ribose polymerase (PARP) inhibitor, was approved by the US Food and Drug Administration

(FDA) for treatment of women with a *BRCA* mutation and advanced breast cancer (previously approved for advanced ovarian cancer) [128]. As a result, the management of many cancers now entails a combination of targeted and traditional cytotoxic chemotherapeutic agents [129].

With greater understanding of shared mutations in different cancers, targeted treatments based on genetic mutation instead of tumor types are under investigation [130] and have already led to a handful of approved drugs. For example, mutations in the epidermal growth factor receptor (EGFR) are found in head and neck cancers [131] and NSCLC [132–134], yet the recommended treatment guidelines differ. In 2017, the first tumor-agnostic drug was approved (Keytruda), with a clinical indication defined by the tumor's molecular profile [135]. However, targeted therapies may not be effective for all cancers exhibiting the specific mutation, as response may be affected by the genetic burden and potentially the causative factors initiating cancer [136].

Several medications for nononcology indications have also been approved, such as ivacaftor for cystic fibrosis and maraviroc for human immunodeficiency virus (HIV). Monoclonal antibody inhibitors of pro-protein convertase subtilisin/kexin type 9 (PCSK9), alirocumab and evolocumab, were approved by the FDA in 2015. These drugs have been shown to be effective treatments to reduce low-density lipoprotein (LDL) cholesterol levels [137,138] through reduced activity of PCSK9, a regulatory protein that affects LDL receptors and is linked to lipid and cholesterol metabolism [139]. Gain-of-function mutations in PCSK9 cause autosomal dominant familial hypercholesterolemia.

With the growing number of targeted drugs, development of companion diagnostics, or tests intended to be used to identify genetic aberrations associated with the targeted drugs, has become an expanding market. In some cases, testing existed long before development of a targeted drug (such as for *BCR-ABL*) and no companion diagnostic test was codeveloped along with the drug. In other cases, such as for the drugs vemurafenib and olaparib, the Cobas 4800 *BRAF* V60 Mutation test and Myriad Genetics' BRACAnalysis CDx were approved as companion diagnostics, respectively. In addition to identifying mutations in drug targeted genes, understanding of a tumor's mutational load and gene expression profiles may further inform appropriate treatment selection [140,141]. Use of comprehensive

TABLE 2.1 List of Genes, Evidence Level, and Type of Information Included in FDA-Approved Package Insert

Gene	Drug	CPIC Level	PharmGKB Level of Evidence	PGx on FDA Label
<i>CFTR</i>	Ivacaftor	A	1A	Testing required
<i>CYP2C19</i>	Amitriptyline	A	1A	
<i>CYP2C19</i>	Clopidogrel	A	1A	Actionable PGx
<i>CYP2C19</i>	Voriconazole	A	1A	Actionable PGx
<i>CYP2C19</i>	Citalopram	A	1A	Actionable PGx
<i>CYP2C19</i>	Escitalopram	A	1A	Actionable PGx
<i>CYP2C9</i>	Phenytoin	A	1A	Actionable PGx
<i>CYP2C9</i>	Warfarin	A	1A	Actionable PGx
<i>CYP2D6</i>	Amitriptyline	A	1A	Actionable PGx
<i>CYP2D6</i>	Codeine	A	1A	Actionable PGx
<i>CYP2D6</i>	Fluvoxamine	A	1A	Actionable PGx
<i>CYP2D6</i>	Nortriptyline	A	1A	Actionable PGx
<i>CYP2D6</i>	Ondansetron	A	1A	Informative PGx
<i>CYP2D6</i>	Paroxetine	A	1A	Informative PGx
<i>CYP2D6</i>	Tropisetron	A		
<i>CYP2D6</i>	Clomipramine	B	1A	Actionable PGx
<i>CYP2D6</i>	Desipramine	B	1A	Actionable PGx
<i>CYP2D6</i>	Doxepin	B	1A	Actionable PGx
<i>CYP2D6</i>	Imipramine	B	1A	Actionable PGx
<i>CYP2D6</i>	Trimipramine	B	1A	Actionable PGx
<i>CYP3A5</i>	Tacrolimus	A	1A	
<i>CYP4F2</i>	Warfarin	A	1B	
<i>DPYD</i>	Capecitabine	A	1A	Actionable PGx
<i>DPYD</i>	Fluorouracil	A	1A	Actionable PGx
<i>DPYD</i>	Tegafur	A	1A	
<i>G6PD</i>	Rasburicase	A	1A	Testing required
<i>HLA-B</i>	Abacavir	A	1A	Testing required
<i>HLA-B</i>	Allopurinol	A	1A	
<i>HLA-B</i>	Carbamazepine	A	1A	Testing required
<i>HLA-B</i>	Phenytoin	A	1A	Actionable PGx
<i>IFNL3</i>	Peginterferon alfa-2a	A	1A	
<i>IFNL3</i>	Peginterferon alfa-2b	A	1A	Actionable PGx
<i>IFNL3</i>	Ribavirin	A	1A	
<i>SLCO1B1</i>	Simvastatin	A	1A	
<i>TPMT</i>	Azathioprine	A	1A	Testing recommended
<i>TPMT</i>	Mercaptopurine	A	1A	Testing recommended
<i>TPMT</i>	Thioguanine	A	1A	Actionable PGx
<i>UGT1A1</i>	Atazanavir	A	1A	
<i>VKORC1</i>	Warfarin	A	1A	Actionable PGx

Data obtained from <https://cpicpgx.org/genes-drugs/>.

TABLE 2.2 Approved Oncology Drugs With a Specific Genetic Indication/Target or Known Risk of Adverse Events Associated With a Genetic Variant

Indication	Drug (Target Genes)
Approved for Treatment of Single Condition	
Breast cancer	<ul style="list-style-type: none"> Abemaciclib (<i>ESR, ERBB2</i>) Ado-trastuzumab emtansine (<i>ERBB2</i>) Anastrozole (<i>ESR, ERBB2</i>) Exemestane (<i>ESR, PGR</i>) Fulvestrant (<i>ERBB2 ESR, PGR</i>) Lapatinib (<i>ERBB2, ESR, PGR, HLA-DQA1, HLA-DRB1</i>) Letrozole (<i>ESR, PGR</i>) Neratinib (<i>ERBB2, ESR, PGR</i>) Palbociclib (<i>ESR, ERBB2</i>) Ribociclib (<i>ESR, PGR, ERBB2</i>) Pertuzumab (<i>ERBB2, ESR, PGR</i>) Tamoxifen (<i>ESR, PGR, F5, F2</i>)
Non-small cell lung cancer (NSCLC)	<ul style="list-style-type: none"> Afatinib (<i>EGFR</i>) Alectinib (<i>ALK</i>) Brigatinib (<i>ALK</i>) Ceritinib (<i>ALK</i>) Crizotinib (<i>ALK, ROS1</i>) Erlotinib (<i>EGFR</i>) Gefitinib (<i>EGFR</i>) Osimertinib (<i>EGFR</i>)
Acute promyelocytic leukemia	<ul style="list-style-type: none"> Arsenic trioxide (<i>PML-RARA</i>) Tretinoin (<i>PML-RARA</i>)
Acute lymphoblastic leukemia (ALL)	<ul style="list-style-type: none"> Blinatumomab (<i>BCR-ABL1</i>) Mercaptopurine (<i>TMPT</i>) Inotuzumab ozogamicin (<i>BCR-ABL1</i>)
AML	<ul style="list-style-type: none"> Enasidenib (<i>IDH2</i>) Thioguanine (<i>TPMT</i>)
Cutaneous T-cell lymphoma	<ul style="list-style-type: none"> Denileukin diftitox (<i>IL2RA [CD25 antigen]</i>)
Peripheral T-cell lymphoma	<ul style="list-style-type: none"> Belinostat (<i>UGT1A1</i>)
Colon/colorectal cancer	<ul style="list-style-type: none"> Irinotecan (<i>UGT1A1</i>) Panitumumab (<i>EGFR, RAS</i>)
Chronic myelogenous leukemia (CML)	<ul style="list-style-type: none"> Bosutinib (<i>BCR-ABL1</i>) Busulfan (<i>BCR-ABL1</i>) Nilotinib (<i>BCR-ABL1</i>) Omacetaxine (<i>BCR-ABL1</i>)
Chronic lymphocytic leukemia (CLL)	<ul style="list-style-type: none"> Venetoclax (<i>17p</i>)
Hyperuricemia associated with malignancy	<ul style="list-style-type: none"> Rasburicase (<i>G6PD, CYB5R</i>)
Melanoma	<ul style="list-style-type: none"> Cobimetinib (<i>BRAF</i>)
Neuroblastoma	<ul style="list-style-type: none"> Dinutuximab (<i>MYCN</i>)
Ovarian cancer	<ul style="list-style-type: none"> Rucaparib (<i>BRCA, CYP2D6, CYP1A2</i>)
Soft tissue sarcoma	<ul style="list-style-type: none"> Olaratumab (<i>PDGFRA</i>)
Approved for Treatment of Multiple Conditions	
<ul style="list-style-type: none"> -Urothelial carcinoma 	Atezolizumab (<i>CD274 [PD-L1]</i>)
<ul style="list-style-type: none"> -NSCLC 	
<ul style="list-style-type: none"> -Merkel cell carcinoma 	Avelumab (<i>CD274 [PD-L1]</i>)
<ul style="list-style-type: none"> -Urothelial carcinoma 	

Continued

TABLE 2.2 Approved Oncology Drugs With a Specific Genetic Indication/Target or Known Risk of Adverse Events Associated With a Genetic Variant—cont'd

Indication	Drug (Target Genes)
<ul style="list-style-type: none"> Anaplastic large cell lymphoma 	Brentuximab vedotin (<i>ALK</i>)
<ul style="list-style-type: none"> Hodgkin lymphoma Mycosis fungoides Renal cell carcinoma 	Cabozantinib (<i>RET</i>)
<ul style="list-style-type: none"> Thyroid cancer Breast cancer 	Capecitabine (<i>DPYD</i>)
<ul style="list-style-type: none"> Colorectal cancer Colorectal cancer Head and neck cancer 	Cetuximab (<i>EGFR RAS</i>)
<ul style="list-style-type: none"> Bladder cancer Ovarian cancer Testicular cancer 	Cisplatin (<i>TPMT</i>)
<ul style="list-style-type: none"> Melanoma NSCLC Thyroid cancer 	Dabrafenib (<i>BRAF, G6PD, RAS</i>)
<ul style="list-style-type: none"> ALL CML NSCLC 	Dasatinib (<i>BCR-ABL1</i>)
<ul style="list-style-type: none"> Urothelial carcinoma Breast cancer 	Durvalumab (<i>CD274 [PD-L1]</i>)
<ul style="list-style-type: none"> Neuroendocrine tumors Renal cell carcinoma Breast cancer 	Everolimus (<i>ERBB2, ESR</i>)
<ul style="list-style-type: none"> Colon and rectal cancer Gastric cancer Pancreatic cancer 	Fluorouracil (<i>DPYD</i>)
<ul style="list-style-type: none"> Chronic graft-versus-host disease (refractory) CLL/small lymphocytic lymphoma Mantle cell lymphoma (MZL) Waldenström macroglobulinemia 	Ibrutinib (<i>17p, 11q</i>)
<ul style="list-style-type: none"> ALL CML Gastrointestinal stromal tumors 	Imatinib (<i>KIT, BCR-ABL1, PDGFRB, FIP1L1-PDGFR</i>)
<ul style="list-style-type: none"> AML Mast cell leukemia 	Midostaurin (<i>FLT3, NPM1, KIT</i>)
<ul style="list-style-type: none"> Ovarian, fallopian tube, or primary peritoneal cancer Colorectal cancer 	Niraparib (<i>BRCA</i>) Nivolumab (<i>BRAF, CD274 [PD-L1], microsatellite instability, mismatch repair</i>)
<ul style="list-style-type: none"> Head and neck cancer Hepatocellular carcinoma Hodgkin lymphoma Melanoma NSCLC Renal cell cancer 	Obinutuzumab (<i>MS4A1 [CD20 antigen]</i>)
<ul style="list-style-type: none"> CLL Follicular lymphoma Breast cancer Ovarian cancer 	Olaparib (<i>BRCA</i>)

Continued

TABLE 2.2 Approved Oncology Drugs With a Specific Genetic Indication/Target or Known Risk of Adverse Events Associated With a Genetic Variant—cont'd

Indication	Drug (Target Genes)
<ul style="list-style-type: none"> Renal cell carcinoma Soft tissue sarcoma Gastric cancer 	Pazopanib (<i>UGT1A1, HLA-B</i>) Pembrolizumab (<i>BRAF, CD274</i> [PD-L1], microsatellite instability, mismatch repair)
<ul style="list-style-type: none"> Head and neck cancer Hodgkin lymphoma Melanoma NSCLC Urothelial carcinoma ALL CML CLL, Non-Hodgkin lymphomas Melanoma NSCLC Thyroid cancer Breast cancer Gastric cancer Melanoma Erdheim-Chester disease 	Ponatinib (<i>BCR-ABL1</i>) Rituximab (<i>MS4A1</i> [CD20 antigen]) Trametinib (<i>BRAF, G6PD, RAS</i>) Trastuzumab (<i>ERBB2, ESR, PGR</i>) Vemurafenib (<i>BRAF, RAS</i>)

Data source: <https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm>.

gene panels [142] or whole genome sequencing of paired tumor/normal tissue [143] has been explored and developed to inform treatment decision and prognosis in cancer patients. Yet, the options for genetic testing in oncology are numerous, creating challenges for providers, patients, and insurers to determine which is most appropriate for a given patient [144]. In 2018, the first comprehensive test panel (FoundationOne CDx) was approved to inform treatment decisions for multiple types of cancer.

Along with targeted drugs, promising advances with biologics, such as immunotherapeutics (immune checkpoint inhibitors) [145], chimeric antigen receptors T-cell (CAR-T) therapy [146], and gene therapy [147] and gene editing, have yielded a new suite of tools and interventions. In 2017, two CAR-T interventions (for acute lymphocytic leukemia and diffuse large B-cell lymphoma), in addition to the first adeno-associated viral (AAV) gene therapy for inherited retinal dystrophy, were approved by the FDA. Also in 2017, the FDA approved the first gene therapy

that delivers copies of the *RPE65* gene through a viral vector in patients with a rare genetic disorder, retinal blindness. Continued development and evaluation of engineered interventions focus on improving the delivery methods (viral type, ex vivo vs. in vivo), target specificity, reducing systemic toxicities, and expanding applications to other diseases [148].

2.4 PRECISION MEDICINE RESEARCH

Several technologies are under intense investigation, adding layers of analysis of various molecular entities and expanding understanding of development, gene—environment interactions, and risk markers. With the layered data sets for a given individual, the complexity of data analysis is increased, warranting new tools and algorithms. Although much emphasis has focused on the study of DNA-based data and technologies, other developments are under investigation to gain a fuller assessment of various biomolecular entities and response to internal and external factors.

The Epigenome. The field of epigenetics has also garnered substantial attention due to the development of new technologies (whole genome bisulfite sequencing) and focus on the study of regulatory elements of the genome [149–151]. DNA methylation is recognized as a critical factor in numerous cellular processes including development and differentiation, cell proliferation, and tumorigenesis. Epigenetics and the importance of methylation status in gene regulation were limited to gene-by-gene analysis of methylation patterns. Today, whole genome analysis can generate comprehensive snapshots of methylation status (“methyloome”) in a range of tissues, developmental stages, and diseases [152]. Large-scale studies have characterized methylation patterns in pregnancy [153], gametogenesis [154], and diseases such as cancer [155,156] and diabetes [157]. The combined analysis of DNA methylation and transcriptome from a single cell can enable correlative analysis [158].

The Microbiome. With the development of new sequencing technologies, it became feasible to characterize microbial communities residing on various human tissues, which was not possible with traditional culturing methods. In 2008, the US National Institutes of Health launched the Human Microbiome Project (HMP), with the goal of characterizing the microbiome in 300 healthy individuals across multiple tissues (<https://hmpdacc.org/hmp/>) [159–161]. With a national initiative and newly developed sequencing and analytical capabilities, many studies have been conducted of microbiota on a range of human tissues, from healthy and affected patients, and from samples collected at regular intervals. The association of microbiota composition (and potential perturbations) to human development and disease has become an area of intense study for a wide range of diseases, including cancer, dermatological conditions [162], obesity, gastroenterological conditions [163], and diabetes [164]. The impacts of microbiome manipulation through diet (prebiotics and probiotics) [165], antibiotics [166], drug response [167,168], and fecal transplantation [169] are under investigation to better understand the role of local microbial communities to disease susceptibility, onset, outcome, and treatment. Related analyses, such as the mycobiome, are also under way [170,171].

Data Sciences. While the singular use of various –omics technologies can provide great insight into particular level of cellular operations, a more comprehensive picture can be achieved through analysis of

multiple data sets generated from a single individual. In addition to the sheer size of combined data sets, the interpretation of these large data sets will remain a key challenge. For example, the combination of the microbiome and metabolome can begin to elucidate the physiological impacts of certain microbiota compositions and perturbation of the microbiome [172,173]. The unprecedentedly large multi-omic data sets generated per individual patient, referred to as “panomics” [174], multidimensional integrative genomics [175], or integrated –omics [176], will greatly benefit from sophisticated analytical capabilities to assess temporal changes in the data as well as other changes associated with aging, lifestyle, or environmental exposures. With the availability of large heterogeneous data sets, including medical records and genomic data, machine learning approaches have begun to be used to develop predictive or diagnostic algorithms [177–179]. Substantial effort has been devoted to artificial intelligence (AI) or machine learning of large, heterogeneous patient data sets to assess phenotypes [180] and identify pathogenic variants [181], drug response [182], and risk or outcome predictors.

2.5 THE PRACTICE OF PRECISION MEDICINE

The implementation and integration of precision health applications into routine clinical practice are the subjects of numerous research programs supported by the National Human Genome Research Institute of the NIH. In particular, Newborn Sequencing in Genomic Medicine and Public Health (NSIGHT; <https://www.genome.gov/27558493/newborn-sequencing-in-genomic-medicine-and-public-health-nsight/>), Clinical Sequencing Evidence-Generating Research (CSER; <https://www.genome.gov/27546194/clinical-sequencing-exploratory-research-cser/>), the Electronic Medical Records and Genomics (eMERGE; <https://www.genome.gov/27540473/>) network, and the Implementing GeNomics In PracTice (IGNITE; <https://www.genome.gov/27554264/>) network are programs whose goals are to develop the insights and tools required for the use of diverse genome-based precision medicine technologies in day to day health care in diverse settings.

In addition to demonstrating the effectiveness of new interventions, the practice of precision medicine

faces several obstacles that will require provider and patient support including educational resources, an EMR system that can provide adequate storage and facilitate easy look-up of laboratory reports, and positive coverage determinations. The impact of personal genomic risk data continues to be under investigation in order to assess patients' informational needs, delivery approaches, and likelihood of adopting healthy behaviors or complying with screening or treatment recommendations [183,184]. Coinciding with other developments in precision medicine is greater patient engagement through online patient portals and self-collection of health information through mobile health technologies. Such barriers may limit access or cause uneven access to these interventions [185,186]. We highlight a few of these areas next.

Digital Health Tools. With the expanding use of digital health technologies such as wearables to measure and record health-related behaviors, health monitoring has become increasingly convenient and easy [187]. Particularly for patients with chronic conditions, evidence has demonstrated greater patient engagement, symptom monitoring, and/or adherence, and behavioral changes between clinic visits with use of mobile health applications [188–194]. The quality of data from consumer wearables may be inconsistent between brands and vary between features [195], though they may still sufficiently motivate healthy behaviors and patient engagement [196]. Other digital health applications include wearable biosensors for vital sign monitoring [197], ingestible biosensors for medication compliance monitoring [198], and smart homes [199]. Patients may also increase the likelihood of establishing and maintaining healthy behaviors with a variety of support from health coaches, support groups or social networks, and/or daily reminders through mobile apps and text messaging. Future patient medical records will likely include data collected from self-monitored measures to provide a more comprehensive picture of a patient's lifestyle and environment.

Clinical Decision Support. With the transition to EMRs, the provision of point-of-care clinical decision support is now feasible, improving the ability to interpret and generate evidence-based recommendations based on self-reported data, FHH, laboratory testing, and other clinical information. One example of the application of digital health tools is the online collection and analysis of FHH. Long captured through

paper-based formats with reportedly limited use, FHH has been revitalized with the development of online FHH tools [200]. Provider use of FHH is limited by time for collection and interpretation and beliefs that detailed FHH (three generations) fall outside the scope of general practice [47]. Similarly, several patient barriers exist to the comprehensive collection of FHH, understanding about FHH, knowledge about family members' health, family dynamics, cultural norms, and privacy concerns [201–207]. Online collection of FHH can help overcome challenges of timely patient reporting and education as well as immediate generation of recommendations based on clinical guidelines that may be reviewed during the office visit. In some cases, a virtual provider can promote family history-taking [49].

MeTree is one example of a patient-facing Web-based FHH driven risk assessment application [48]. The app is integrated into clinical practices and provides clinical decision support to patients and their primary care providers about risk level for 30 different conditions and evidence-based recommendations for how to manage that risk. Such electronic FHH collection tools may greatly improve the comprehensiveness of information collected from patients [208], spur family health information-sharing [209], particularly if educational support is available to help patients understand what type of information to report and about which family members [210], and the ability to update it outside of a clinic visit. The impact of MeTree on patients and providers and the type of care patients has been piloted [211] and is currently being assessed in a large multi-institutional study of patients in five national healthcare settings [212]. For providers that have not implemented electronic FHH tools, software for specific conditions has been developed to generate risk assessments compatible with publicly available FHH tools [213].

Coverage and Reimbursement. Routine use of new clinical applications is unlikely to occur without positive coverage determinations based in large part on evidence of clinical utility and cost-effectiveness. While several studies have assessed the cost-effectiveness of a wide range of tests [214–217], evidence from large clinical trials is lacking. Furthermore, clinical guidelines are lacking regarding the use of many genetic and genomic tests, with the one exception being PGx testing [218]. Insurers and government review bodies

have reviewed many genetic and genomic tests, but many cite inconclusive evidence as the basis for negative coverage determination. As the costs of genotyping and sequencing technologies continue to decline, testing may become cost-effective [219] or more affordable as an out-of-pocket expense for some patients or health systems. In many areas, laboratory testing has moved beyond single-gene tests to gene panels without significant price increases given the reduced costs of multigene testing. Such tests may be used for risk prediction, carrier status, or prognosis or likelihood of disease recurrence. Although the cost of adding more genes to a test may be negligible, the value of test panels compared with single-gene testing or whole exome or genome is still being investigated [220–222]. As evidence of the utility of genetic risk scores increases for conditions such as cardiovascular disease [95], their use in the clinical setting has begun to be evaluated [96].

Provider Education. Unlike human genetics, primarily practiced by providers specially trained in genetics were needed, giving rise to the development of its own clinical specialty (American Board of Medical Genetics and Genomics) and genetic counseling, precision medicine is anticipated to be practiced across all specialties and, most likely, not involve a geneticist expert. However, one of the major barriers to integrating precision medicine is providers' limited knowledge and/or experience with either traditional genetic tests or more recently developed genetic and genomic tests [223–226]. Other reported barriers include the lack of clinical guidelines [227], inconsistent reimbursement coverage, concerns about insurance discrimination [228], and lack of time and educational resources to discuss testing with patients [229]. Despite their lack of knowledge, providers have expressed their enthusiasm and interest in the new applications [223]. For example, although information is available about the impact of PGx variants on the metabolism of several medications, more clear and concise information is needed at the point-of-care to assist health providers to integrate PGx testing [230].

Delivering precision medicine will require the storage and integration of large data sets, sophisticated EMRs [231,232] and analytical software, clinical decision support, portability and access to patient data, and patient educational resources. While many health apps and other eHealth tools have been developed, additional

research, development, and evaluation are required in order to achieve widespread use [233]. In particular, clinical decision support and point-of-care educational tools are essential for the appropriate use and integration of new applications [234–238].

Patient Education and Support. Overall, the general public and patients have expressed strong interest in genetic and genomic testing for various purposes [239]. However, lack of knowledge [240,241] and some concerns have been repeatedly identified and, therefore, health providers need to be prepared to discuss these issues with patients and family members to promote understanding and informed decision-making. In particular, concerns about the impact of results on family members, psychological impact, privacy, and accuracy of testing have been raised [242]. With the limited number of genetic counselors in clinical care, alternative approaches to facilitating patient understanding (pretesting/posttesting) and informed decision-making are needed [243], such as telegenetics [244], patient-friendly test reports [245], and group-based educational and counseling sessions [246].

2.6 CONCLUSION

The era of personalized or precision medicine has arisen from the intersection of several forces, including new discoveries and understanding of the molecular basis for health and disease, the interaction/impact of environmental exposures on health, development of new laboratory technologies (e.g., sequencing), increasing participant engagement, EHRs, and digital and/or mobile health technologies. With the continued availability of a suite of increasingly rapid and less expensive –omics technologies and greater capability of informatics to merge and analyze heterogeneous data sets, the challenges once posed by complex diseases are less formidable and new opportunities for discovery and clinical translation are now possible. The evidence that these novel approaches provide value to the patient, provider, and health care delivery system remains a significant challenge to widespread adoption by clinicians and for coverage by payers. A range of clinical support and educational opportunities are needed to adequately prepare health providers to appropriately deliver and integrate these new tools into practice.

REFERENCES

- [1] Khoury MJ, Galea S. Will precision medicine improve population health? *J Am Med Assoc* 2016;316:1357–8.
- [2] Bland JS, Minich DM, Eck BM. A systems medicine approach: translating emerging science into individualized wellness. *Adv Met Med* 2017;2017:1718957.
- [3] Mutie PM, Giordano GN, Franks PW. Lifestyle precision medicine: the next generation in type 2 diabetes prevention? *BMC Med* 2017;15:171.
- [4] Feero W, Wicklund CA, Veenstra D. Precision medicine, genome sequencing, and improved population health. *J Am Med Assoc* 2018;319:1979–80.
- [5] Ginsburg GS, Phillips KA. Precision medicine: from science to value. *Health Aff (Millwood)* 2018;37:694–701.
- [6] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chisoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubinfeld M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowski J. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
- [7] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers

- YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hattton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yoosheph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science* 2001;291:1304–51.
- [8] Levy SE, Myers RM. Advancements in next-generation sequencing. *Annu Rev Genom Hum Genet* 2016;17:95–115.
 - [9] Organick L, Ang SD, Chen YJ, Lopez R, Yekhanin S, Makarychev K, Racz MZ, Kamath G, Gopalan P, Nguyen B, Takahashi CN, Newman S, Parker HY, Rashtchian C, Stewart K, Gupta G, Carlson R, Mulligan J, Carmean D, Seelig G, Ceze L, Strauss K. Random access in large-scale DNA data storage. *Nat Biotechnol* 2018;36:242–8.
 - [10] Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J, Nagarajan R, Carter AB, Pantanowitz L. Next-generation sequencing informatics: challenges and strategies for implementation in a clinical environment. *Arch Pathol Lab Med* 2016;140:958–75.
 - [11] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* 2014;30:418–26.
 - [12] Ganguli M, Albanese E, Seshadri S, Bennett DA, Lyketsos C, Kukull WA, Skoog I, Hendrie HC. Population neuroscience: dementia epidemiology serving precision medicine and population health. *Alzheimer Dis Assoc Disord* 2018;32:1–9.
 - [13] Vaithinathan AG, Asokan V. Public health and precision medicine share a goal. *J Evid Based Med* 2017;10:76–80.
 - [14] Belsky DW, Moffitt TE, Caspi A. Genetics in population health science: strategies and opportunities. *Am J Public Health* 2013;103(Suppl. 1):S73–83.
 - [15] Meagher KM, McGowan ML, Settersten Jr RA, Fishman JR, Juengst ET. Precisely where are we going? Charting the new terrain of precision prevention. *Annu Rev Genom Hum Genet* 2017;18:369–87.
 - [16] Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793–5.
 - [17] Turnbull C. Introducing whole genome sequencing into routine cancer care: the genomics England 100,000 genomes project. *Ann Oncol* 2018.
 - [18] Cuypers B, Berg M, Imamura H, Dumetz F, De Muylder G, Domagalska MA, Rijal S, Bhattarai NR, Maes I, Sanders M, Cotton JA, Meysman P, Laukens K, Dujardin JC. Integrated genomic and metabolomic profiling of ISC1, an emerging *Leishmania donovani* population in the Indian subcontinent. *Infect Genet Evol* 2018.
 - [19] Grunert M, Dorn C, Cui H, Dunkel I, Schulz K, Schoenhals S, Sun W, Berger F, Chen W, Sperling SR. Comparative DNA methylation and gene expression analysis identifies novel genes for structural congenital heart diseases. *Cardiovasc Res* 2016;112:464–77.
 - [20] Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasil-iou V, Ioannidis JPA, Johnson CH. Beyond genomics: understanding exposotypes through metabolomics. *Hum Genom* 2018;12:4.
 - [21] Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;148(6):1293–307.
 - [22] Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, McDonald DT, Kusebauch U, Moss CL, Zhou Y, Qin S, Moritz RL, Brogaard K, Omenn GS, Lovejoy JC, Hood L. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol* 2017;35:747–56.
 - [23] Rechitsky S, Pakhalchuk T, San Ramos G, Goodman A, Zlatopolsky Z, Kuliev A. First systematic experience of preimplantation genetic diagnosis for single-gene disorders, and/or preimplantation human leukocyte antigen typing, combined with 24-chromosome aneuploidy testing. *Fertil Steril* 2015;103:503–12.
 - [24] Iwarsson E, Jacobsson B, Dagerhamn J, Davidson T, Bernabe E, Heibert Arnlin M. Analysis of cell-free fetal DNA in maternal blood for detection of trisomy 21, 18 and 13 in a general pregnant population and in a high risk population - a systematic review and meta-analysis. *Acta Obstet Gynecol Scand* 2017;96(1):7–18.
 - [25] Renga B. Non invasive prenatal diagnosis of fetal aneuploidy using cell free fetal DNA. *Eur J Obstet Gynecol Reprod Biol* 2018;225:5–8.

- [26] Committee opinion No. 640: cell-free DNA screening for fetal aneuploidy. *Obstet Gynecol* 2015;126:e31–37.
- [27] Chen F, Liu P, Gu Y, Zhu Z, Nanisetti A, Lan Z, Huang Z, Liu SJ, Kang X, Deng Y, Luo L, Jiang D, Qiu Y, Pan J, Xia J, Xiong K, Liu C, Xie L, Shi Q, Li J, Zhang X, Wang W, Drmanac S, Jiang H, Drmanac R, Xu X. Isolation and Whole Genome Sequencing of fetal cells from maternal blood towards the ultimate non-invasive prenatal testing. *Prenat Diagn* 2017.
- [28] McLennan A, Palma-Dias R, da Silva Costa F, Meagher S, Nisbet DL, Scott F. Noninvasive prenatal testing in routine clinical practice—an audit of NIPT and combined first-trimester screening in an unselected Australian population. *Aust N Z J Obstet Gynaecol* 2016;56:22–8.
- [29] Bjerregaard L, Stenbakken AB, Andersen CS, Kristensen L, Jensen CV, Skovbo P, Sorensen AN. The rate of invasive testing for trisomy 21 is reduced after implementation of NIPT. *Dan Med J* 2017;64.
- [30] Johnson K, Kelley J, Saxton V, Walker SP, Hui L. Declining invasive prenatal diagnostic procedures: a comparison of tertiary hospital and national data from 2012 to 2015. *Aust N Z J Obstet Gynaecol* 2017;57:152–6.
- [31] Kane SC, Reidy KL, Norris F, Nisbet DL, Kornman LH, Palma-Dias R. Chorionic villus sampling in the cell-free DNA aneuploidy screening era: careful selection criteria can maximise the clinical utility of screening and invasive testing. *Prenat Diagn* 2017;37:399–408.
- [32] Bianchi DW, Chudova D, Sehnert AJ, Bhatt S, Murray K, Prosen TL, Garber JE, Wilkins-Haug L, Vora NL, Warsof S, Goldberg J, Ziainia T, Halks-Miller M. Noninvasive prenatal testing and incidental detection of occult maternal malignancies. *J Am Med Assoc* 2015;314:162–9.
- [33] Cohen PA, Flowers N, Tong S, Hannan N, Pertile MD, Hui L. Abnormal plasma DNA profiles in early ovarian cancer using a non-invasive prenatal testing platform: implications for cancer screening. *BMC Med* 2016;14:126.
- [34] Schmid M, Wang E, Bogard PE, Bevilacqua E, Hacker C, Wang S, Doshi J, White K, Kaplan J, Sparks A, Jani JC, Stokowski R. Prenatal screening for 22q11.2 deletion using a targeted microarray-based cell-free DNA test. *Fetal Diagn Ther* 2017.
- [35] Best S, Wou K, Vora N, Van der Veyver IB, Wapner R, Chitty LS. Promises, pitfalls and practicalities of prenatal whole exome sequencing. *Prenat Diagn* 2018;38:10–9.
- [36] Hayward J, Chitty LS. Beyond screening for chromosomal abnormalities: advances in non-invasive diagnosis of single gene disorders and fetal exome sequencing. *Semin Fetal Neonatal Med* 2018;23:94–101.
- [37] Committee opinion No.682: microarrays and next-generation sequencing technology: the use of advanced genetic diagnostic tools in obstetrics and gynecology. *Obstet Gynecol* 2016;128:e262–8.
- [38] Fukami M, Miyado M. Next generation sequencing and array-based comparative genomic hybridization for molecular diagnosis of pediatric endocrine disorders. *Ann Pediatr Endocrinol Metab* 2017;22:90–4.
- [39] Farnaes L, Hildreth A, Sweeney NM, Clark MM, Chowdhury S, Nahas S, Cakici JA, Benson W, Kaplan RH, Kronick R, Bainbridge MN, Friedman J, Gold JJ, Ding Y, Veeraraghavan N, Dimmock D, Kingsmore SF. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom Med* 2018;3:10.
- [40] Bartnik M, Wisniewiecka-Kowalik B, Nowakowska B, Smyk M, Kedzior M, Sobiecka K, Kutkowska-Kazmierczak A, Klapecki J, Szczaluba K, Castaneda J, Wlasienko P, Bezniakow N, Obersztyn E, Bocian E. The usefulness of array comparative genomic hybridization in clinical diagnostics of intellectual disability in children. *Dev Period Med* 2014;18:307–17.
- [41] Manning M, Hudgins L. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genet Med* 2010;12:742–5.
- [42] South ST, Lee C, Lamb AN, Higgins AW, Kearney HM. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. *Genet Med* 2013;15:901–9.
- [43] Bevilacqua J, Hesse A, Cormier B, Davey J, Patel D, Shankar K, Reddi HV. Clinical utility of a 377 gene custom next-generation sequencing epilepsy panel. *J Genet* 2017;96:681–5.
- [44] Butler KM, da Silva C, Alexander JJ, Hegde M, Escayg A. Diagnostic yield from 339 epilepsy patients screened on a clinical gene panel. *Pediatr Neurol* 2017;77:61–6.
- [45] Chambers C, Jansen LA, Dhamija R. Review of commercially available epilepsy genetic panels. *J Genet Couns* 2016;25:213–7.
- [46] Berg AO, Baird MA, Botkin JR, Driscoll DA, Fishman PA, Guarino PD, Hiatt RA, Jarvik GP, Millon-Underwood S, Morgan TM, Mulvihill JJ, Pollin TI, Schimmel SR, Stefanek ME, Vollmer WM, Williams JK. National Institutes of health state-of-the-science conference statement: family history and improving health. *Ann Intern Med* 2009;151(12):872–7.
- [47] Saul RA, Trotter T, Sease K, Tarini B. Survey of family history taking and genetic testing in pediatric practice. *J Community Genet* 2017;8:109–15.
- [48] Orlando LA, Buchanan AH, Hahn SE, Christianson CA, Powell KP, Skinner CS, Chesnut B, Blach C, Due B, Ginsburg GS, Henrich VC. Development and validation of a primary care-based family health history and decision support program (MeTree). *N C Med J* 2013;74:287–96.

- [49] Wang C, Bickmore T, Bowen DJ, Norkunas T, Campion M, Cabral H, Winter M, Paasche-Orlow M. Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. *Genet Med* 2015;17:822–30.
- [50] Wu RR, Himmel TL, Buchanan AH, Powell KP, Hauser ER, Ginsburg GS, Henrich VC, Orlando LA. Quality of family history collection with use of a patient facing family history assessment tool. *BMC Fam Pract* 2014;15:31.
- [51] Wu RR, Myers RA, Hauser ER, Vorderstrasse A, Cho A, Ginsburg GS, Orlando LA. Impact of genetic testing and family health history based risk counseling on behavior change and cognitive precursors for type 2 diabetes. *J Genet Counsel* 2017;26:133–40.
- [52] Roberts MC, Dotson WD, DeVore CS, Bednar EM, Bowen DJ, Ganiats TG, Green RF, Hurst GM, Philp AR, Ricker CN, Sturm AC, Trepanier AM, Williams JL, Zierhut HA, Wilemon KA, Hampel H. Delivery of cascade screening for hereditary conditions: a scoping review of the literature. *Health Aff (Millwood)* 2018;37(5):801–8.
- [53] Hampel H. Genetic counseling and cascade genetic testing in Lynch syndrome. *Fam Cancer* 2016;15(3):423–7.
- [54] Knowles JW, Rader DJ, Khoury MJ. Cascade screening for familial hypercholesterolemia and the use of genetic testing. *J Am Med Assoc* 2017;318:381–2.
- [55] Santos RD, Frauches TS, Chacra AP. Cascade screening in familial hypercholesterolemia: advancing forward. *J Atheroscler Thromb* 2015;22:869–80.
- [56] Theilade J, Kanters J, Henriksen FL, Gilsa-Hansen M, Svendsen JH, Eschen O, Toft E, Reimers JJ, Tybjaerg-Hansen A, Christiansen M, Jensen HK, Bundgaard H. Cascade screening in families with inherited cardiac diseases driven by cardiologists: feasibility and nationwide outcome in long QT syndrome. *Cardiology* 2013;126:131–7.
- [57] Lieberman S, Lahad A, Tomer A, Koka S, BenUziyahu M, Raz A, Levy-Lahad E. Familial communication and cascade testing among relatives of BRCA population screening participants. *Genet Med* 2018.
- [58] Allison M. Communicating risk with relatives in a familial hypercholesterolemia cascade screening program: a summary of the evidence. *J Cardiovasc Nurs* 2015;30:E1–12.
- [59] Sturm AC. Cardiovascular cascade genetic testing: exploring the role of direct contact and technology. *Front Cardiovasc Med* 2016;3:11.
- [60] McClaren BJ, Aitken M, Massie J, Amor D, Ukoumunne OC, Metcalfe SA. Cascade carrier testing after a child is diagnosed with cystic fibrosis through newborn screening: investigating why most relatives do not have testing. *Genet Med* 2013;15:533–40.
- [61] Bodian DL, McCutcheon JN, Kothiyal P, Huddleston KC, Iyer RK, Vockley JG, Niederhuber JE. Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: implications for individual genome sequencing. *PLoS One* 2014;9:e94554.
- [62] Linderman MD, Nielsen DE, Green RC. Personal genome sequencing in ostensibly healthy individuals and the PeopleSeq consortium. *J Pers Med* 2016;6.
- [63] Sanderson SC. Genome sequencing for healthy individuals. *Trends Genet* 2013;29(10):556–8.
- [64] Vassy JL, Lautenbach DM, McLaughlin HM, Kong SW, Christensen KD, Krier J, Kohane IS, Feuerman LZ, Blumenthal-Barby J, Roberts JS, Lehmann LS, Ho CY, Ubel PA, MacRae CA, Seidman CE, Murray MF, McGuire AL, Rehm HL, Green RC. The MedSeq project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* 2014;15:85.
- [65] Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, Bouffard GG, Chines PS, Cruz P, Hansen NF, Teer JK, Maskeri B, Young AC, Manolio TA, Wilson AF, Finkel T, Hwang P, Arai A, Remaley AT, Sachdev V, Shamburek R, Cannon RO, Green ED. The ClinSeq project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res* 2009;19:1665–74.
- [66] Petrikin JE, Willig LK, Smith LD, Kingsmore SF. Rapid whole genome sequencing and precision neonatology. *Semin Perinatol* 2015;39(8):623–31.
- [67] van Diemen CC, Kerstjens-Frederikse WS, Bergman KA, de Koning TJ, Sikkema-Raddatz B, van der Velde JK, Abbott KM, Herkert JC, Lohner K, Rump P, Meems-Veldhuis MT, Neerinx PBT, Jongbloed JDH, van Ravenswaaij-Arts CM, Swertz MA, Sinke RJ, van Langen IM, Wijmenga C. Rapid targeted genomics in critically ill newborns. *Pediatrics* 2017;140.
- [68] Ostergren JE, Gornick MC, Carere DA, Kalia SS, Uhlmann WR, Ruffin MT, Mountain JL, Green RC, Roberts JS, Group PGS. How well do customers of direct-to-consumer personal genomic testing services comprehend genetic test results? Findings from the impact of personal genomics study. *Public Health Genom* 2015;18:216–24.
- [69] Krarup NT, Borglykke A, Allin KH, Sandholt CH, Justesen JM, Andersson EA, Grarup N, Jorgensen T, Pedersen O, Hansen T. A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6041 Danish individuals. *Atherosclerosis* 2015;240(2):305–10.
- [70] Szulkin R, Whittington T, Eklund M, Aly M, Eeles RA, Easton D, Kote-Jarai ZS, Amin Al Olama A, Benlloch S, Muir K, Giles GG, Southey MC, Fitzgerald LM, Henderson BE, Schumacher F, Haiman CA, Schleutker

- J, Wahlfors T, Tammela TL, Nordestgaard BG, Key TJ, Travis RC, Neal DE, Donovan JL, Hamdy FC, Pharoah P, Pashayan N, Khaw KT, Stanford JL, Thibodeau SN, McDonnell SK, Schaid DJ, Maier C, Vogel W, Luedeke M, Herkommer K, Kibel AS, Cybulski C, Lubinski J, Kluzniak W, Cannon-Albright L, Brenner H, Butterbach K, Stegmaier C, Park JY, Sellers T, Lin HY, Slavov C, Kaneva R, Mitev V, Batra J, Clements JA, Spurdle A, Teixeira MR, Paulo P, Maia S, Pandha H, Michael A, Kierzek A, Gronberg H, Wiklund F. Prediction of individual genetic risk to prostate cancer using a polygenic score. *Prostate* 2015;75:1467–74.
- [71] Hung CF, Breen G, Czamara D, Corre T, Wolf C, Kloiber S, Bergmann S, Craddock N, Gill M, Holsboer F, Jones L, Jones I, Korszun A, Kutalik Z, Lucae S, Maier W, Mors O, Owen MJ, Rice J, Rietschel M, Uher R, Vollenweider P, Waeber G, Craig IW, Farmer AE, Lewis CM, Muller-Myhsok B, Preisig M, McGuffin P, Rivera M. A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder. *BMC Med* 2015;13:86.
- [72] Prescott J, Setiawan VW, Wentzensen N, Schumacher F, Yu H, Delahanty R, Bernstein L, Chanock SJ, Chen C, Cook LS, Friedenreich C, Garcia-Closas M, Haiman CA, Le Marchand L, Liang X, Lissowska J, Lu L, Magliocco AM, Olson SH, Risch HA, Shu XO, Ursin G, Yang HP, Kraft P, De Vivo I. Body mass index genetic risk score and endometrial cancer risk. *PLoS One* 2015;10:e0143256.
- [73] Chen H, Liu X, Brendler CB, Ankerst DP, Leach RJ, Goodman PJ, Lucia MS, Tangen CM, Wang L, Hsu FC, Sun J, Kader AK, Isaacs WB, Helfand BT, Zheng SL, Thompson IM, Platz EA, Xu J. Adding genetic risk score to family history identifies twice as many high-risk men for prostate cancer: results from the prostate cancer prevention trial. *Prostate* 2016;76:1120–9.
- [74] Macinnis RJ, Antoniou AC, Eeles RA, Severi G, Al Olama AA, McGuffog L, Kote-Jarai Z, Guy M, O'Brien LT, Hall AL, Wilkinson RA, Sawyer E, Ardern-Jones AT, Dearnaley DP, Horwich A, Khoo VS, Parker CC, Huddart RA, Van As N, McCredie MR, English DR, Giles GG, Hopper JL, Easton DF. A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. *Genet Epidemiol* 2011;35:549–56.
- [75] Abida W, Armenia J, Gopalan A, Brennan R, Walsh M, Barron D, Danila D, Rathkopf D, Morris M, Slovin S, McLaughlin B, Curtis K, Hyman DM, Durack JC, Solomon SB, Arcila ME, Zehir A, Syed A, Gao J, Chakravarty D, Vargas HA, Robson ME, Joseph V, Offit K, Donoghue MTA, Abeshouse AA, Kundra R, Heins ZJ, Person AV, Harris C, Taylor BS, Ladanyi M, Mandelker D, Zhang L, Reuter VE, Kantoff PW, Solit DB, Berger MF, Sawyers CL, Schultz N, Scher HI. Prospective genomic profiling of prostate cancer across disease states reveals germline and somatic alterations that may affect clinical decision making. *JCO Precis Oncol* 2017;2017.
- [76] Schrader KA, Cheng DT, Joseph V, Prasad M, Walsh M, Zehir A, Ni A, Thomas T, Benayed R, Ashraf A, Lincoln A, Arcila M, Stadler Z, Solit D, Hyman DM, Zhang L, Klimstra D, Ladanyi M, Offit K, Berger M, Robson M. Germline variants in targeted tumor sequencing using matched normal DNA. *JAMA Oncol* 2016;2:104–11.
- [77] Frey MK, Kim SH, Bassett RY, Martineau J, Dalton E, Chern JY, Blank SV. Rescreening for genetic mutations using multi-gene panel testing in patients who previously underwent non-informative genetic screening. *Gynecol Oncol* 2015;139:211–5.
- [78] Judkins T, Leclair B, Bowles K, Gutin N, Trost J, McCulloch J, Bhatnagar S, Murray A, Craft J, Wardell B, Bastian M, Mitchell J, Chen J, Tran T, Williams D, Potter J, Jammulapati S, Perry M, Morris B, Roa B, Timms K. Development and analytical validation of a 25-gene next generation sequencing panel that includes the BRCA1 and BRCA2 genes to assess hereditary cancer risk. *BMC Cancer* 2015;15:215.
- [79] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817–26.
- [80] Tothill RW, Shi F, Paiman L, Bedo J, Kowalczyk A, Mileskin L, Buella E, Klupacs R, Bowtell D, Byron K. Development and validation of a gene expression tumour classifier for cancer of unknown primary. *Pathology* 2015;47(1):7–12.
- [81] Crespo-Leiro MG, Stypmann J, Schulz U, Zuckermann A, Mohacsi P, Bara C, Ross H, Parameshwar J, Zakliczynski M, Fioocchi R, Hoefer D, Deng M, Leprince P, Hiller D, Eubank L, Deljkich E, Yee JP, Vanhaecke J. Performance of gene-expression profiling test score variability to predict future clinical events in heart transplant recipients. *BMC Cardiovasc Disord* 2015;15:120.
- [82] Deng MC, Elashoff B, Pham MX, Teuteberg JJ, Kfoury AG, Starling RC, Cappola TP, Kao A, Anderson AS, Cotts WG, Ewald GA, Baran DA, Bogaev RC, Shahzad K, Hiller D, Yee J, Valentine HA. Utility of gene expression profiling score variability to predict clinical events in heart transplant recipients. *Transplantation* 2014;97:708–14.

- [83] Pham MX, Teuteberg JJ, Kfoury AG, Starling RC, Deng MC, Cappola TP, Kao A, Anderson AS, Cotts WG, Ewald GA, Baran DA, Bogaev RC, Elashoff B, Baron H, Yee J, Valantine HA. Gene-expression profiling for rejection surveillance after cardiac transplantation. *N Engl J Med* 2010;362:1890–900.
- [84] Fujita B, Prashovikj E, Schulz U, Borgermann J, Sunavsky J, Fuchs U, Gummert J, Ensminger S. Predictive value of gene expression profiling for long-term survival after heart transplantation. *Transpl Immunol* 2017;41:27–31.
- [85] Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, Sager PT, Sehnert AJ, Yau M, Kraus WE, Newby LK, Schwartz RS, Voros S, Ellis SG, Tahirkheli N, Waksman R, McPherson J, Lansky A, Winn ME, Schork NJ, Topol EJ. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. *Ann Intern Med* 2010;153(7):425–34.
- [86] Vargas J, Lima JA, Kraus WE, Douglas PS, Rosenberg S. Use of the corus(R) CAD gene expression test for assessment of obstructive coronary artery disease likelihood in symptomatic non-diabetic patients. *PLoS Curr* 2013;5.
- [87] Boileau A, Lalem T, Vausort M, Zhang L, Devaux Y. A 3-gene panel improves the prediction of left ventricular dysfunction after acute myocardial infarction. *Int J Cardiol* 2018;254:28–35.
- [88] Muse ED, Wineinger NE, Spencer EG, Peters M, Henderson R, Zhang Y, Barrett PM, Rivera SP, Wohlgemuth JG, Devlin JJ, Shiffman D, Topol EJ. Validation of a genetic risk score for atrial fibrillation: a prospective multicenter cohort study. *PLoS Med* 2018;15:e1002525.
- [89] Saracyn M, Kisiel B, Bachta A, Franaszczyk M, Brodowska-Kania D, Zmudzki W, Szymanski K, Sokalski A, Klatko W, Stopinski M, Grochowski J, Paplinski M, Gozdziak Z, Niemczyk L, Bober B, Kolodziej M, Tlustochowicz W, Kaminski G, Ploski R, Niemczyk S. Value of multilocus genetic risk score for atrial fibrillation in end-stage kidney disease patients in a Polish population. *Sci Rep* 2018;8(1):9284.
- [90] Theriault S, Whitlock R, Raman K, Vincent J, Yusuf S, Pare G. Gene expression profiles for the identification of prevalent atrial fibrillation. *J Am Heart Assoc* 2017;6.
- [91] Buscot MJ, Magnussen CG, Juonala M, Pitkanen N, Lehtimäki T, Viikari JS, Kahonen M, Hutri-Kahonen N, Schork NJ, Raitakari OT, Thomson RJ. The combined effect of common genetic risk variants on circulating lipoproteins is evident in childhood: a longitudinal analysis of the cardiovascular risk in young finns study. *PLoS One* 2016;11:e0146081.
- [92] Knowles JW, Zarafshar S, Pavlovic A, Goldstein BA, Tsai S, Li J, McConnell MV, Absher D, Ashley EA, Kiernan M, Ioannidis JPA, Assimes TL. Impact of a genetic risk score for coronary artery disease on reducing cardiovascular risk: a pilot randomized controlled study. *Front Cardiovasc Med* 2017;4:53.
- [93] Kullo IJ, Jouni H, Austin EE, Brown SA, Kruisselbrink TM, Isseh IN, Haddad RA, Marroush TS, Shameer K, Olson JE, Broeckel U, Green RC, Schaid DJ, Montori VM, Bailey KR. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the MI-GENES clinical trial). *Circulation* 2016;133:1181–8.
- [94] Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, Guiducci C, Perola M, Jula A, Sinisalo J, Lokki ML, Nieminen MS, Melander O, Salomaa V, Peltonen L, Kathiresan S. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* 2010;376:1393–400.
- [95] Knowles JW, Ashley EA. Cardiovascular disease: the rise of the genetic risk score. *PLoS Med* 2018;15:e1002546.
- [96] Celestino-Soper PB, Gao H, Lynnes TC, Lin H, Liu Y, Spoonamore KG, Chen PS, Vatta M. Validation and utilization of a clinical next-generation sequencing panel for selected cardiovascular disorders. *Front Cardiovasc Med* 2017;4:11.
- [97] Weinshilboum RM, Wang L. Pharmacogenomics: precision medicine and drug response. *Mayo Clin Proc* 2017;92:1711–22.
- [98] Poudel DR, Acharya P, Ghimire S, Dhital R, Bharati R. Burden of hospitalizations related to adverse drug events in the USA: a retrospective analysis from large inpatient database. *Pharmacoepidemiol Drug Saf* 2017;26:635–41.
- [99] Cardelli M, Marchegiani F, Corsonello A, Lattanzio F, Provinciali M. A review of pharmacogenetics of adverse drug reactions in elderly people. *Drug Saf* 2012;35(Suppl. 1):3–20.
- [100] Phillips KA, Veenstra DL, Oren E, Lee JK, Sadee W. Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. *J Am Med Assoc* 2001;286:2270–9.
- [101] Peck R. Precision medicine is not just genomics: the right dose for every patient. *Annu Rev Pharmacol Toxicol* 2017.
- [102] Kimmel SE, French B, Kasner SE, Johnson JA, Anderson JL, Gage BF, Rosenberg YD, Eby CS, Madigan RA, McBane RB, Abdel-Rahman SZ, Stevens SM, Yale S, Mohler 3rd ER, Fang MC, Shah V, Horenstein RB, Limdi NA, Muldowney 3rd JA, Gujral J, Delafontaine P, Desnick RJ, Ortel TL, Billett HH, Pendleton RC,

- Geller NL, Halperin JL, Goldhaber SZ, Caldwell MD, Califf RM, Ellenberg JH, Investigators C. A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N Engl J Med* 2013;369:2283–93.
- [103] Pirmohamed M, Burnside G, Eriksson N, Jorgensen AL, Toh CH, Nicholson T, Kesteven P, Christersson C, Wahlstrom B, Stafberg C, Zhang JE, Leathart JB, Kohnke H, Maitland-van der Zee AH, Williamson PR, Daly AK, Avery P, Kamali F, Wadelius M, Grp E-P. A randomized trial of genotype-guided dosing of warfarin. *N Engl J Med* 2013;369:2294–303.
- [104] Maciel A, Cullors A, Lukowiak AA, Garces J. Estimating cost savings of pharmacogenetic testing for depression in real-world clinical settings. *Neuropsychiatr Dis Treat* 2018;14:225–30.
- [105] Seripa D, Lozupone M, Stella E, Paroni G, Bisceglia P, La Montagna M, D'Onofrio G, Gravina C, Urbano M, Priore MG, Lamanna A, Daniele A, Bellomo A, Logroscino G, Greco A, Panza F. Psychotropic drugs and CYP2D6 in late-life psychiatric and neurological disorders. What do we know? *Expert Opin Drug Saf* 2017;16:1373–85.
- [106] Maughan A, Ogbuagu O. Pegylated interferon alpha 2a for the treatment of hepatitis C virus infection. *Expert Opin Drug Metab Toxicol* 2018;14:219–27.
- [107] Matic M, de Wildt SN, Tibboel D, van Schaik RHN. Analgesia and opioids: a pharmacogenetics shortlist for implementation in clinical practice. *Clin Chem* 2017;63:1204–13.
- [108] Senagore AJ, Champagne BJ, Dosokey E, Brady J, Steele SR, Reynolds HL, Stein SL, Delaney CP. Pharmacogenetics-guided analgesics in major abdominal surgery: further benefits within an enhanced recovery protocol. *Am J Surg* 2017;213:467–72.
- [109] Turner RM, Pirmohamed M. Cardiovascular pharmacogenomics: expectations and practical benefits. *Clin Pharmacol Ther* 2014;95:281–93.
- [110] Weeke P, Roden DM. Applied pharmacogenomics in cardiovascular medicine. *Annu Rev Med* 2014;65:81–94.
- [111] Elliott LS, Henderson JC, Neradilek MB, Moyer NA, Ashcraft KC, Thirumaran RK. Clinical impact of pharmacogenetic profiling with a clinical decision support tool in polypharmacy home health patients: a prospective pilot randomized controlled trial. *PLoS One* 2017;12:e0170905.
- [112] Sugarman EA, Cullors A, Centeno J, Taylor D. Contribution of pharmacogenetic testing to modeled medication change recommendations in a long-term care population with polypharmacy. *Drugs Aging* 2016;33(12):929–36.
- [113] Haga SB, Moaddab J. Comparison of delivery strategies for pharmacogenetic testing services. *Pharmacogenet Genom* 2014;24:139–45.
- [114] Bielinski SJ, Olson JE, Pathak J, Weinshilboum RM, Wang L, Lyke KJ, Ryu E, Targonski PV, Van Norstrand MD, Hathcock MA, Takahashi PY, McCormick JB, Johnson KJ, Maschke KJ, Rohrer Vitek CR, Ellingson MS, Wieben ED, Farrugia G, Morrisette JA, Kruckeberg KJ, Bruflat JK, Peterson LM, Blommel JH, Skierka JM, Ferber MJ, Black JL, Baudhuin LM, Klee EW, Ross JL, Veldhuizen TL, Schultz CG, Caraballo PJ, Freimuth RR, Chute CG, Kullo IJ. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. *Mayo Clin Proc* 2014;89:25–33.
- [115] Dunnenberger HM, Crews KR, Hoffman JM, Caudle KE, Broeckel U, Howard SC, Hunkler RJ, Klein TE, Evans WE, Relling MV. Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers. *Annu Rev Pharmacol Toxicol* 2015;55:89–106.
- [116] Dong OM, Wiltshire T. Advancing precision medicine in healthcare: addressing implementation challenges to increase pharmacogenetic testing in the clinical setting. *Physiol Genom* 2017;49:346–54.
- [117] Moyer AM, Caraballo PJ. The challenges of implementing pharmacogenomic testing in the clinic. *Expert Rev Pharmacoecon Outcomes Res* 2017;17:567–77.
- [118] Wiltshire T, Dong OM. Clinical pharmacogenetics: how do we ensure a favorable future for patients? *Pharmacogenomics* 2018;19:553–62.
- [119] Dugger SA, Platt A, Goldstein DB. Drug development in the era of precision medicine. *Nat Rev Drug Discov* 2018;17:183–96.
- [120] Wulfkuhle JD, Spira A, Edmiston KH, Petricoin 3rd EF. Innovations in clinical trial design in the era of molecular profiling. *Methods Mol Biol* 2017;1606:19–36.
- [121] Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. *Nat Genet* 2015;47:856–60.
- [122] Boessen R, van der Baan F, Groenwold R, Egberts A, Klungel O, Grobbee D, Knol M, Roes K. Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharm Stat* 2013;12:366–74.
- [123] Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med* 2018;16:29.
- [124] Barroilhet L, Matulonis U. The NCI-MATCH trial and precision medicine in gynecologic cancers. *Gynecol Oncol* 2018;148:585–90.

- [125] Beckman RA, Antonijevic Z, Kalamegham R, Chen C. Adaptive design for a confirmatory basket trial in multiple tumor types based on a putative predictive biomarker. *Clin Pharmacol Ther* 2016;100:617–25.
- [126] Park JW, Liu MC, Yee D, Yau C, van 't Veer LJ, Symmans WF, Paoloni M, Perlmutter J, Hylton NM, Hogarth M, DeMichele A, Buxton MB, Chien AJ, Wallace AM, Boughey JC, Haddad TC, Chui SY, Kemmer KA, Kaplan HG, Isaacs C, Nanda R, Tripathy D, Albain KS, Edmiston KK, Elias AD, Northfelt DW, Puszta L, Moulder SL, Lang JE, Viscusi RK, Euhus DM, Haley BB, Khan QJ, Wood WC, Melisko M, Schwab R, Helsten T, Lyandres J, Davis SE, Hirst GL, Sanil A, Esserman LJ, Berry DA. Adaptive randomization of neratinib in early breast cancer. *N Engl J Med* 2016;375:11–22.
- [127] Patel JN. Cancer pharmacogenomics, challenges in implementation, and patient-focused perspectives. *Pharmgenom Pers Med* 2016;9:65–77.
- [128] Robson M, Im SA, Senkus E, Xu B, Domchek SM, Masuda N, Delaloge S, Li W, Tung N, Armstrong A, Wu W, Goessl C, Runswick S, Conte P. Olaparib for metastatic breast cancer in patients with a germline BRCA mutation. *N Engl J Med* 2017;377:523–33.
- [129] Perl AE. The role of targeted therapy in the management of patients with AML. *Blood Adv* 2017;1:2281–94.
- [130] Hainsworth JD, Meric-Bernstam F, Swanton C, Hurwitz H, Spigel DR, Sweeney C, Burris H, Bose R, Yoo B, Stein A, Beattie M, Kurzrock R. Targeted therapy for advanced solid tumors on the basis of molecular profiles: results from MyPathway, an open-label, phase IIa multiple basket study. *J Clin Oncol* 2018;36:536–42.
- [131] Xu MJ, Johnson DE, Grandis JR. EGFR-targeted therapies in the post-genomic era. *Cancer Metastasis Rev* 2017;36:463–73.
- [132] Schuette W, Schirmacher P, Eberhardt WE, Fischer JR, von der Schulenburg JM, Mezger J, Schumann C, Serke M, Zaun S, Dietel M, Thomas M. EGFR mutation status and first-line treatment in patients with stage III/IV non-small cell lung cancer in Germany: an observational study. *Cancer Epidemiol Biomarkers Prev* 2015;24(8):1254–61.
- [133] Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. *Nat Rev Cancer* 2007;7:169–81.
- [134] Tufman A, Kahnert K, Duell T, Kauffmann-Guerreiro D, Milger K, Schneider C, Stump J, Syunyaeva Z, Huber RM, Reu S. Frequency and clinical relevance of EGFR mutations and EML4-ALK translocations in octogenarians with non-small cell lung cancer. *Onco-Targets Ther* 2017;10:5179–86.
- [135] Lemery S, Keegan P, Pazdur R. First FDA approval agnostic of cancer site - when a biomarker defines the indication. *N Engl J Med* 2017;377:1409–12.
- [136] Afghahi A, Kurian AW. The changing landscape of genetic testing for inherited breast cancer predisposition. *Curr Treat Options Oncol* 2017;18:27.
- [137] Robinson JG, Farnier M, Krempf M, Bergeron J, Luc G, Averna M, Stroes ES, Langslet G, Raal FJ, El Shaway M, Koren MJ, Lepor NE, Lorenzato C, Pordy R, Chaudhari U, Kastelein JJ. Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. *N Engl J Med* 2015;372:1489–99.
- [138] Sabatine MS, Giugliano RP, Wiviott SD, Raal FJ, Blom DJ, Robinson J, Ballantyne CM, Somaratne R, Legg J, Wasserman SM, Scott R, Koren MJ, Stein EA. Efficacy and safety of evolocumab in reducing lipids and cardiovascular events. *N Engl J Med* 2015;372:1500–9.
- [139] Reiss AB, Shah N, Muhieddine D, Zhen J, Yudkevich J, Kasselmann LJ, DeLeon J. PCSK9 in cholesterol metabolism: from bench to bedside. *Clin Sci (Lond)* 2018;132:1135–53.
- [140] Dijkstra KK, Voabil P, Schumacher TN, Voest EE. Genomics- and transcriptomics-based patient selection for cancer treatment with immune checkpoint inhibitors: a review. *JAMA Oncol* 2016;2:1490–5.
- [141] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, Chan TA. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015;348:124–8.
- [142] Campesato LF, Barroso-Sousa R, Jimenez L, Correa BR, Sabbaga J, Hoff PM, Reis LF, Galante PA, Camargo AA. Comprehensive cancer-gene panels can be used to estimate mutational load and predict clinical benefit to PD-1 blockade in clinical practice. *Oncotarget* 2015;6:34221–7.
- [143] Rennert H, Eng K, Zhang T, Tan A, Xiang J, Romanel A, Kim R, Tam W, Liu YC, Bhinder B, Cyrta J, Beltran H, Robinson B, Mosquera JM, Fernandes H, Demichelis F, Sboner A, Kluk M, Rubin MA, Elemento O. Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations, indels and copy-number alterations for precision cancer care. *NPJ Genom Med* 2016;1.
- [144] Borad MJ, LoRusso PM. Twenty-first century precision medicine in oncology: genomic profiling in patients with cancer. *Mayo Clin Proc* 2017;92:1583–91.
- [145] Atkins MB, Larkin J. Immunotherapy combined or sequenced with targeted therapy in the treatment of solid tumors: current perspectives. *J Natl Cancer Inst* 2016;108:djv414.
- [146] Sadelain M, Riviere I, Riddell S. Therapeutic T cell engineering. *Nature* 2017;545:423–31.

- [147] Dunbar CE, High KA, Joung JK, Kohn DB, Ozawa K, Sadelain M. Gene therapy comes of age. *Science* 2018;359.
- [148] Ellebrecht CT, Bhoj VG, Nace A, Choi EJ, Mao X, Cho MJ, Di Zenzo G, Lanzavecchia A, Seykora JT, Cotsarelis G, Milone MC, Payne AS. Reengineering chimeric antigen receptor T cells for targeted therapy of autoimmune disease. *Science* 2016;353:179–84.
- [149] An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
- [150] Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA* 2014;111:6131–8.
- [151] Siggens L, Ekwall K. Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. *J Intern Med* 2014;276:201–14.
- [152] Fouse SD, Nagarajan RO, Costello JF. Genome-scale DNA methylation analysis. *Epigenomics* 2010;2:105–17.
- [153] Lapato DM, Moyer S, Olivares E, Amstadter AB, Kinser PA, Latendresse SJ, Jackson-Cook C, Roberson-Nay R, Strauss JF, York TP. Prospective longitudinal study of the pregnancy DNA methylome: the US pregnancy, race, environment, genes (PREG) study. *BMJ Open* 2018;8:e019721.
- [154] Castillo J, Jodar M, Oliva R. The contribution of human sperm proteins to the development and epigenome of the preimplantation embryo. *Hum Reprod Update* 2018.
- [155] Beekman R, Chapaprieta V, Russinol N, Vilarrasa-Blasi R, Verdaguer-Dot N, Martens JHA, Duran-Ferrer M, Kulis M, Serra F, Javierre BM, Wingett SW, Clot G, Queiros AC, Castellano G, Blanc J, Gut M, Merkel A, Heath S, Vlasova A, Ullrich S, Palumbo E, Enjuanes A, Martin-Garcia D, Bea S, Pinyol M, Aymerich M, Royo R, Puiggros M, Torrents D, Datta A, Lowy E, Kostadima M, Roller M, Clarke L, Flicek P, Agirre X, Prosper F, Baumann T, Delgado J, Lopez-Guillermo A, Fraser P, Yaspo ML, Guigo R, Siebert R, Marti-Renom MA, Puente XS, Lopez-Otin C, Gut I, Stunnenberg HG, Campo E, Martin-Subero JI. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat Med* 2018.
- [156] Lee EJ, Luo J, Wilson JM, Shi H. Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer Lett* 2013;340:171–8.
- [157] Chen Z, Miao F, Paterson AD, Lachin JM, Zhang L, Schones DE, Wu X, Wang J, Tompkins JD, Genuth S, Braffett BH, Riggs AD, Natarajan R. Epigenomic profiling reveals an association between persistence of DNA methylation and metabolic memory in the DCCT/EDIC type 1 diabetes cohort. *Proc Natl Acad Sci USA* 2016;113:E3002–11.
- [158] Hu Y, Huang K, An Q, Du G, Hu G, Xue J, Zhu X, Wang CY, Xue Z, Fan G. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* 2016;17:88.
- [159] Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–14.
- [160] Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N, Patel S, Cutting M, Madden T, Hamilton H, Harris E, Gevers D, Simone G, McInnes P, Versalovic J. The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. *Faseb J* 2013;27:1012–22.
- [161] Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, Young S, Warren WC, Badger J, Crabtree J, Markowitz VM, Orvis J, Cree A, Ferreira S, Fulton LL, Fulton RS, Gillis M, Hemphill LD, Joshi V, Kovar C, Torralba M, Wetterstrand KA, Abouelille A, Wollam AM, Buhay CJ, Ding Y, Dugan S, FitzGerald MG, Holder M, Hostetler J, Clifton SW, Allen-Vercos E, Earl AM, Farmer CN, Liolios K, Surette MG, Xu Q, Pohl C, Wilczek-Boney K, Zhu D. A catalog of reference genomes from the human microbiome. *Science* 2010;328(5981):994–9.
- [162] Yamazaki Y, Nakamura Y, Nunez G. Role of the microbiota in skin immunity and atopic dermatitis. *Allergol Int* 2017;66:539–44.
- [163] Goulet O. Potential role of the intestinal microbiota in programming health and disease. *Nutr Rev* 2015;73(Suppl. 1):32–40.
- [164] Sohail MU, Althani A, Anwar H, Rizzi R, Marei HE. Role of the gastrointestinal tract microbiome in the pathophysiology of diabetes mellitus. *J Diabetes Res* 2017;2017:9631435.
- [165] Tankou SK, Regev K, Healy BC, Tjon E, Laghi L, Cox LM, Kivisakk P, Pierre IV, Lokhande H, Gandhi R, Cook S, Glanz B, Stankiewicz J, Weiner HL. A probiotic modulates the microbiome and immunity in multiple sclerosis. *Ann Neurol* 2018.
- [166] Becattini S, Taur Y, Pamer EG. Antibiotic-induced changes in the intestinal microbiota and disease. *Trends Mol Med* 2016;22:458–78.

- [167] Doestzada M, Vila AV, Zhernakova A, Koonen DPY, Weersma RK, Touw DJ, Kuipers F, Wijmenga C, Fu J. Pharmacomicrobiomics: a novel route towards personalized medicine? *Protein Cell* 2018;9:432–45.
- [168] Swanson HI. Drug metabolism by the host and gut microbiota: a partnership or rivalry? *Drug Metab Dispos* 2015;43:1499–504.
- [169] Mintz M, Khair S, Grewal S, LaComb JF, Park J, Chaner B, Rajapakse R, Bucobo JC, Buscaglia JM, Monzur F, Chawla A, Yang J, Robertson CE, Frank DN, Li E. Longitudinal microbiome analysis of single donor fecal microbiota transplantation in patients with recurrent *Clostridium difficile* infection and/or ulcerative colitis. *PLoS One* 2018;13:e0190997.
- [170] Hoarau G, Mukherjee PK, Gower-Rousseau C, Hager C, Chandra J, Retuerto MA, Neut C, Vermeire S, Clemente J, Colombel JF, Fujioka H, Poulain D, Sendid B, Ghannoum MA. Bacteriome and mycobiome interactions underscore microbial dysbiosis in familial Crohn's disease. *mBio* 2016;7.
- [171] Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA, Muzny DM, Gibbs RA, Ajami NJ, Petrosino JF. The gut mycobiome of the human microbiome project healthy cohort. *Microbiome* 2017;5:153.
- [172] Aw W, Fukuda S. An integrated outlook on the metagenome and metabolome of intestinal diseases. *Diseases* 2015;3:341–59.
- [173] Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. Microbiome and metabolome data integration provides insight into health and disease. *Transl Res* 2017;189:51–64.
- [174] Sandhu C, Qureshi A, Emili A. Panomics for precision medicine. *Trends Mol Med* 2017.
- [175] Arneson D, Shu L, Tsai B, Barrere-Cain R, Sun C, Yang X. Multidimensional integrative genomics approaches to dissecting cardiovascular disease. *Front Cardiovasc Med* 2017;4:8.
- [176] Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* 2018;19:299–310.
- [177] Beam AL, Kohane IS. Big data and machine learning in health care. *J Am Med Assoc* 2018.
- [178] Diao J, Kohane IS, Manrai AK. Biomedical informatics and machine learning for clinical genomics. *Hum Mol Genet* 2018.
- [179] Lin E, Lane HY. Machine learning and systems genomics approaches for multi-omics data. *Biomark Res* 2017;5:2.
- [180] Basile AO, Ritchie MD. Informatics and machine learning to define the phenotype. *Expert Rev Mol Diagn* 2018;18:219–26.
- [181] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;31:761–3.
- [182] Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, Sander C, Cherniack AD, Mina M, Ciriello G, Schultz N, Sanchez Y, Greene CS. Machine learning detects Pan-cancer ras pathway activation in the cancer genome atlas. *Cell Rep* 2018;23:172–80.
- [183] Hay JL, Berwick M, Zielaskowski K, White KA, Rodriguez VM, Robers E, Guest DD, Sussman A, Talamantes Y, Schwartz MR, Greb J, Bigney J, Kaphingst KA, Hunley K, Buller DB. Implementing an internet-delivered skin cancer genetic testing intervention to improve sun protection behavior in a diverse population: protocol for a randomized controlled trial. *JMIR Res Protoc* 2017;6:e52.
- [184] Smit AK, Espinoza D, Newson AJ, Morton RL, Fenton G, Freeman L, Dunlop K, Butow PN, Law MH, Kimlin MG, Keogh LA, Dobbins SJ, Kirk J, Kanetsky PA, Mann GJ, Cust AE. A pilot randomized controlled trial of the feasibility, acceptability, and impact of giving information on personalized genomic risk of melanoma to the public. *Cancer Epidemiol Biomarkers Prev* 2017;26:212–21.
- [185] Childers KK, Maggard-Gibbons M, Macinko J, Childers CP. National distribution of cancer genetic testing in the United States: evidence for a gender disparity in hereditary breast and ovarian cancer. *JAMA Oncol* 2018;4:876–9.
- [186] Dellefave-Castillo LM, Puckelwartz MJ, McNally EM. Reducing racial/ethnic disparities in cardiovascular genetic testing. *JAMA Cardiol* 2018;3:277–9.
- [187] Piwek L, Ellis DA, Andrews S, Joinson A. The rise of consumer health wearables: promises and barriers. *PLoS Med* 2016;13:e1001953.
- [188] Buller DB, Berwick M, Lantz K, Buller MK, Shane J, Kane I, Liu X. Smartphone mobile application delivering personalized, real-time sun protection advice: a randomized clinical trial. *JAMA Dermatol* 2015;151:497–504.
- [189] Lee JA, Choi M, Lee SA, Jiang N. Effective behavioral intervention strategies using mobile health applications for chronic disease management: a systematic review. *BMC Med Inform Decis Mak* 2018;18:12.
- [190] Low CA, Dey AK, Ferreira D, Kamarck T, Sun W, Bae S, Doryab A. Estimation of symptom severity during chemotherapy from passively sensed data: exploratory study. *J Med Internet Res* 2017;19(12):e420.
- [191] Murphy J, Holmes J, Brooks C. Measurements of daily energy intake and total energy expenditure in people with dementia in care homes: the use of wearable technology. *J Nutr Health Aging* 2017;21:927–32.

- [192] Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. *J Med Internet Res* 2017;19:e295.
- [193] Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. *J Med Internet Res* 2016;18:e97.
- [194] Zhan A, Mohan S, Tarolli C, et al. Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. *JAMA Neurol* 2018;75(7):876–80.
- [195] Shcherbina A, Mattsson CM, Waggott D, Salisbury H, Christle JW, Hastie T, Wheeler MT, Ashley EA. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med* 2017;7.
- [196] Mercer K, Giangregorio L, Schneider E, Chilana P, Li M, Grindrod K. Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation. *JMIR Mhealth Uhealth* 2016;4:e7.
- [197] Ajami S, Teimouri F. Features and application of wearable biosensors in medical care. *J Res Med Sci* 2015;20(12):1208–15.
- [198] Chai PR, Rosen RK, Boyer EW. Ingestible biosensors for real-time medical adherence monitoring: MyTMed. *Proc Annu Hawaii Int Conf Syst Sci* 2016;2016:3416–23.
- [199] Karunanithi M, Zhang Q. An innovative technology to support independent living: the smarter safer homes platform. *Stud Health Technol Inform* 2018;246:102–10.
- [200] Welch BM, Wiley K, Pflieger L, Achiangia R, Baker K, Hughes-Halbert C, Morrison H, Schiffman J, Doerr M. Review and comparison of electronic patient-facing family health history tools. *J Genet Counsel* 2018.
- [201] Ashida S, Kaphingst KA, Goodman M, Schafer EJ. Family health history communication networks of older adults: importance of social relationships and disease perceptions. *Health Educ Behav* 2013;40:612–9.
- [202] Ashida S, Schafer EJ. Family health information sharing among older adults: reaching more family members. *J Community Genet* 2015;6:17–27.
- [203] Chen LS, Li M, Talwar D, Xu L, Zhao M. Chinese Americans' views and use of family health history: a qualitative study. *PLoS One* 2016;11:e0162706.
- [204] Hughes Halbert C, Welch B, Lynch C, Magwood G, Rice L, Jefferson M, Riley J. Social determinants of family health history collection. *J Community Genet* 2016;7:57–64.
- [205] Koehly LM, Ashida S, Goergen AF, Skapinsky KF, Hadley DW, Wilkinson AV. Willingness of Mexican-American adults to share family health history with healthcare providers. *Am J Prev Med* 2011;40:633–6.
- [206] Thompson T, Seo J, Griffith J, Baxter M, James A, Kaphingst KA. The context of collecting family health history: examining definitions of family and family communication about health among African American women. *J Health Commun* 2015;20:416–23.
- [207] Underwood SM, Kelber S. Enhancing the collection, discussion and use of family health history by consumers, nurses and other health care providers: because family health history matters. *Nurs Clin N Am* 2015;50:509–29.
- [208] Conway-Pearson LS, Christensen KD, Savage SK, Huntington NL, Weitzman ER, Ziniel SI, Bacon P, Cacioppo CN, Green RC, Holm IA. Family health history reporting is sensitive to small changes in wording. *Genet Med* 2016;18:1308–11.
- [209] Wang C, Sen A, Plegue M, Ruffin MT, O'Neill SM, Rubinstein WS, Acheson LS. Impact of family history assessment on communication with family members and health care providers: a report from the Family Healthware Impact Trial (FHITr). *Prev Med* 2015;77:28–34.
- [210] Beadles CA, Rynne Wu R, Himmel T, Buchanan AH, Powell KP, Hauser E, Henrich VC, Ginsburg GS, Orlando LA. Providing patient education: impact on quantity and quality of family health history collection. *Fam Cancer* 2014;13:325–32.
- [211] Wu RR, Orlando LA, Himmel TL, Buchanan AH, Powell KP, Hauser ER, Agbaje AB, Henrich VC, Ginsburg GS. Patient and primary care provider experience using a family health history collection, risk stratification, and clinical decision support tool: a type 2 hybrid controlled implementation-effectiveness trial. *BMC Fam Pract* 2013;14:111.
- [212] Wu RR, Myers RA, McCarty CA, Dimmock D, Farrell M, Cross D, Chenevere TD, Ginsburg GS, Orlando LA, Family Health History N. Protocol for the “Implementation, adoption, and utility of family history in diverse care settings” study. *Implement Sci* 2015;10:163.
- [213] Feero WG, Facio FM, Glogowski EA, Hampel HL, Stopfer JE, Eidem H, Pizzino AM, Barton DK, Biesecker LG. Preliminary validation of a consumer-oriented colorectal cancer risk assessment tool compatible with the US Surgeon General's my family health portrait. *Genet Med* 2015;17:753–6.
- [214] Eccleston A, Bentley A, Dyer M, Strydom A, Vereecken W, George A, Rahman N. A cost-effectiveness evaluation of germline BRCA1 and BRCA2 testing in UK women with ovarian cancer. *Value Health* 2017;20(4):567–76.
- [215] Kazi DS, Garber AM, Shah RU, Dudley RA, Mell MW, Rhee C, Moshkevich S, Boothroyd DB, Owens DK, Hlatky MA. Cost-effectiveness of genotype-guided and dual antiplatelet therapies in acute coronary syndrome. *Ann Intern Med* 2014;160:221–32.

- [216] Li Y, Arellano AR, Bare LA, Bender RA, Strom CM, Devlin JJ. A multigene test could cost-effectively help extend life expectancy for women at risk of hereditary breast cancer. *Value Health* 2017;20:547–55.
- [217] Yuen T, Carter MT, Szatmari P, Ungar WJ. Cost-effectiveness of genome and exome sequencing in children diagnosed with autism spectrum disorder. *Appl Health Econ Health Policy* 2018.
- [218] Caudle KE, Rettie AE, Whirl-Carrillo M, Smith LH, Mintzer S, Lee MT, Klein TE, Callaghan JT, Clinical Pharmacogenetics Implementation, C. Clinical pharmacogenetics implementation consortium guidelines for CYP2C9 and HLA-B genotypes and phenytoin dosing. *Clin Pharmacol Ther* 2014;96:542–8.
- [219] Plumpton CO, Alfirevic A, Pirmohamed M, Hughes DA. Cost effectiveness analysis of HLA-B*58:01 genotyping prior to initiation of allopurinol for gout. *Rheumatology* 2017;56:1729–39.
- [220] Alfares AA, Kelly MA, McDermott G, Funke BH, Lebo MS, Baxter SB, Shen J, McLaughlin HM, Clark EH, Babb LJ, Cox SW, DePalma SR, Ho CY, Seidman JG, Seidman CE, Rehm HL. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. *Genet Med* 2015;17(11):880–8.
- [221] Dong OM, Li A, Suzuki O, Oni-Orisan A, Gonzalez R, Stouffer GA, Lee CR, Wiltshire T. Projected impact of a multigene pharmacogenetic test to optimize medication prescribing in cardiovascular patients. *Pharmacogenomics* 2018.
- [222] Karakaya M, Storbeck M, Strathmann EA, Vedove AD, Holker I, Altmueller J, Naghiyeva L, Schmitz-Steinkruger L, Vezyroglou K, Motameny S, Alawbathani S, Thiele H, Polat AI, Okur D, Boostani R, Karimiani EG, Wunderlich G, Ardicli D, Topaloglu H, Kirschner J, Schrank B, Maroofian R, Magnusson O, Yis U, Nurnberg P, Heller R, Wirth B. Targeted sequencing with expanded gene profile enables high diagnostic yield in non-5q-spinal muscular atrophies. *Hum Mutat* 2018.
- [223] Carroll JC, Makuwaza T, Manca DP, Sopcak N, Permaul JA, O'Brien MA, Heisey R, Eisenhauer EA, Easley J, Krzyzanowska MK, Miedema B, Pruthi S, Sawka C, Schneider N, Sussman J, Urquhart R, Versaevl C, Grunfeld E. Primary care providers' experiences with and perceptions of personalized genomic medicine. *Can Fam Physician* 2016;62:e626–35.
- [224] Chan WV, Johnson JA, Wilson RD, Metcalfe A. Obstetrical provider knowledge and attitudes towards cell-free DNA screening: results of a cross-sectional national survey. *BMC Pregnancy Childbirth* 2018;18:40.
- [225] Hamilton JG, Abdiwahab E, Edwards HM, Fang ML, Jdayani A, Breslau ES. Primary care providers' cancer genetic testing-related knowledge, attitudes, and communication behaviors: a systematic review and research agenda. *J Gen Intern Med* 2017;32:315–24.
- [226] Johnson LM, Valdez JM, Quinn EA, Sykes AD, McGee RB, Nuccio R, Hines-Dowell SJ, Baker JN, Kesserwan C, Nichols KE, Mandrell BN. Integrating next-generation sequencing into pediatric oncology practice: an assessment of physician confidence and understanding of clinical genomics. *Cancer* 2017;123:2352–9.
- [227] Schully SD, Lam TK, Dotson WD, Chang CQ, Aronson N, Birkeland ML, Brewster SJ, Boccia S, Buchanan AH, Calonge N, Calzone K, Djulbegovic B, Goddard KA, Klein RD, Klein TE, Lau J, Long R, Lyman GH, Morgan RL, Palmer CG, Relling MV, Rubinstein WS, Swen JJ, Terry SF, Williams MS, Khoury MJ. Evidence synthesis and guideline development in genomic medicine: current status and future prospects. *Genet Med* 2014.
- [228] Sperber NR, Carpenter JS, Cavallari LHLJD, Cooper-DeHoff RM, Denny JC, Ginsburg GS, Guan Y, Horowitz CR, Levy KD, Levy MA, Madden EB, Matheny ME, Pollin TI, Pratt VM, Rosenman M, Voils CIKWW, Wilke RA, RYanne Wu R, Orlando LA. Challenges and strategies for implementing genomic services in diverse settings: experiences from the Implementing GeNomics in pracTice (IGNITE) network. *BMC Med Genom* 2017;10:35.
- [229] Gammon BL, Kraft SA, Michie M, Allyse M. "I think we've got too many tests!": prenatal providers' reflections on ethical and clinical challenges in the practice integration of cell-free DNA screening. *Ethics Med Public Health* 2016;2:334–42.
- [230] Filipski KK, Pacanowski MA, Ramamoorthy A, Feero WG, Freedman AN. Dosing recommendations for pharmacogenetic interactions related to drug metabolism. *Pharmacogenet Genom* 2016;26:334–9.
- [231] Hartzler A, McCarty CA, Rasmussen LV, Williams MS, Brilliant M, Bowton EA, Clayton EW, Faucett WA, Ferryman K, Field JR, Fullerton SM, Horowitz CR, Koenig BA, McCormick JB, Ralston JD, Sanderson SC, Smith ME, Trinidad SB. Stakeholder engagement: a key component of integrating genomic information into electronic health records. *Genet Med* 2013;15:792–801.
- [232] Sitapati A, Kim H, Berkovich B, Marmor R, Singh S, El-Kareh R, Clay B, Ohno-Machado L. Integrated precision medicine: the role of electronic health records in delivering personalized treatment. *Wiley Interdiscip Rev Syst Biol Med* 2017;9.

- [233] Spanakis EG, Santana S, Tsiknakis M, Marias K, Sakkalis V, Teixeira A, Janssen JH, de Jong H, Tziraki C. Technology-based innovations to foster personalized healthy lifestyles and well-being: a targeted review. *J Med Internet Res* 2016;18:e128.
- [234] Aronson SJ, Clark EH, Babb LJ, Baxter S, Farwell LM, Funke BH, Hernandez AL, Joshi VA, Lyon E, Parthum AR, Russell FJ, Varugheese M, Venman TC, Rehm HL. The GeneInsight suite: a platform to support laboratory and provider use of DNA-based genetic testing. *Hum Mutat* 2011;32:532–6.
- [235] Caraballo PJ, Bielinski SJ, St Sauver JL, Weinshilboum RM. Electronic medical record-integrated pharmacogenomics and related clinical decision support concepts. *Clin Pharmacol Ther* 2017;102:254–64.
- [236] Freimuth RR, Formea CM, Hoffman JM, Matey E, Peterson JF, Boyce RD. Implementing genomic clinical decision support for drug-based precision medicine. *CPT Pharmacometrics Syst Pharmacol* 2017;6(3):153–5.
- [237] Melton BL, Zillich AJ, Saleem J, Russ AL, Tisdale JE, Overholser BR. Iterative development and evaluation of a pharmacogenomic-guided clinical decision support system for warfarin dosing. *Appl Clin Inform* 2016;7:1088–106.
- [238] Rohrer Vitek CR, Abul-Husn NS, Connolly JJ, Hartzler AL, Kitchner T, Peterson JF, Rasmussen LV, Smith ME, Stallings S, Williams MS, Wolf WA, Prows CA. Healthcare provider education to support integration of pharmacogenomics in practice: the eMERGE Network experience. *Pharmacogenomics* 2017;18:1013–25.
- [239] Vermeulen E, Henneman L, van El CG, Cornel MC. Public attitudes towards preventive genomics and personal interest in genetic testing to prevent disease: a survey study. *Eur J Public Health* 2014;24:768–75.
- [240] Chapman R, Likhanov M, Selita F, Zakharov I, Smith-Woolley E, Kovas Y. New literacy challenge for the twenty-first century: genetic knowledge is poor even among well educated. *J Community Genet* 2018.
- [241] Haga SB, Barry WT, Mills R, Ginsburg GS, Svetkey L, Sullivan J, Willard HF. Public knowledge of and attitudes toward genetics and genetic testing. *Genet Test Mol Biomarkers* 2013;17:327–35.
- [242] Frost CJ, Andrusis IL, Buys SS, Hopper JL, John EM, Terry MB, Bradbury A, Chung WK, Colbath K, Quintana N, Gamarra E, Egleston B, Galpern N, Bealin L, Glendon G, Miller LP, Daly MB. Assessing patient readiness for personalized genomic medicine. *J Community Genet* 2018.
- [243] Buchanan AH, Rahm AK, Williams JL. Alternate service delivery models in cancer genetic counseling: a mini-review. *Front Oncol* 2016;6:120.
- [244] Vrecar I, Hristovski D, Peterlin B. Telegenetics: an update on availability and use of telemedicine in clinical genetics service. *J Med Syst* 2017;41:21.
- [245] Haga SB, Mills R, Pollak KI, Rehder C, Buchanan AH, Lipkus IM, Crow JH, Datto M. Developing patient-friendly genetic and genomic test reports: formats to promote patient engagement and understanding. *Genome Med* 2014;6.
- [246] Gammon BL, Otto L, Wick M, Borowski K, Allyse M. Implementing group prenatal counseling for expanded noninvasive screening options. *J Genet Counsel* 2017;27(4):894–901.

Nature and Frequency of Genetic Disease

Bruce R. Korf¹, Reed E. Pyeritz², Wayne W. Grody³

¹Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, United States

²Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, United States

³UCLA School of Medicine, Los Angeles, CA, United States

3.1 INTRODUCTION

Genes are major determinants of human variation. Genome sequencing studies have shown that an individual genome contains millions of differences from the reference genome, some of which occur in coding or regulatory sequences that may have a phenotypic effect. Any individual is heterozygous for 50–100 variants that have been associated with genetic disorders [1]. The de novo mutation rate in the human genome is on the order of 10^{-8} per nucleotide per generation [2–4]. The phenotypic effect of genetically determined characteristics spans a continuum and may be nil at one extreme or lethal at the other. In determining a trait, one or two alleles may be all important, but more commonly genes interact with one another and with environmental factors. This interaction between genetic and environmental factors is now apparent for numerous conditions and includes many previously believed to have a purely environmental etiology. For example, genetically determined susceptibility has now been identified for several infections [5], for many drug-induced effects, and for several carcinogens (e.g., bladder cancer in aniline dye workers who are slow acetylators [6]). We suspect that such interactions are commonplace and that relatively few conditions are solely environmental in causation. In addition, to the environment, interactions with the microbiome, epigenetic modifications resulting in

altered gene expression, and chance all influence how the genome translates into a phenotype.

3.2 FREQUENCY OF GENETIC DISEASE

For definitions of the types and frequency of genetic disorders, we use the currently available information, with the proviso that these are all the minimum frequencies and based on imperfect categorization. Interpretation of what constitutes a genetic disorder will depend on the situation (e.g., red-green color blindness may be a serious disability to the hunter-gatherer) or on the public perception (e.g., persons with albinism are considered blessed in some populations). Hence, the distinction between normal variation and disease is blurred and variable criteria have been used in different surveys. The situation is further complicated by the continued delineation of new phenotypic subtypes and by population variations in the frequencies of different genetic disorders [7]. Thus, most of the available data pertain to specific conditions either in a sample of the general population or in specific populations (e.g., the learning disabled); extrapolation to the overall population is difficult.

3.2.1 Chromosomal Disorders

A chromosomal disorder is classically defined as the phenotype resulting from visible alteration in the number or structure of the chromosomes. Using routine

light microscopy and a moderate level of chromosome banding, the frequency of balanced and unbalanced structural rearrangements in newborns has been estimated at about 9.2:1000 [8]. Some of those with unbalanced rearrangements will have congenital anomalies and/or intellectual disabilities. A proportion of those with balanced changes will, in adult life, be at increased risk of either miscarriage or having a disabled child. The incidence of aneuploidy in newborns is about 3:1000, but the frequency increases dramatically among stillbirths or in spontaneous abortions [9]. It is estimated that one in two conceptuses has a chromosome abnormality, usually resulting in miscarriage [10]. Different types of chromosomal abnormalities predominate in spontaneous abortions compared with live-born infants. For example, trisomy 16 is the most common autosomal trisomy in abortions [11], whereas trisomies for chromosomes 21, 18, and 13 are the only autosomal trisomies occurring at appreciable frequencies in live-born infants. Monosomy for the X chromosome (45,X) occurs in about 1% of all conceptions, but 98% of those affected do not reach term. Triploidy is also frequent in abortions but is exceptional in newborns. The high frequency of chromosomally abnormal conceptions is mirrored by results of chromosome analysis in gametes, which reveal an approximate abnormality rate of 4%–5% in sperm [12] and 12%–15% in oocytes [13]. When a family experiences three or more miscarriages, one parent is identified with an autosomal chromosome abnormality in 8.5% of analyses [14].

Routine light microscopy cannot resolve small amounts of missing or additional material (less than 4 Mb of DNA). The advent of cytogenomic microarray analysis has revealed a high frequency of submicroscopic deletions and duplications and other copy number variations, including both apparently benign and pathological changes [15]. Multiple microdeletion and microduplication disorders have been defined in recent years, and undoubtedly more await discovery. Such microdeletions, which epistemologically link “chromosomal disorders” with single-gene disorders, account for a proportion of currently unexplained learning disability and multiple malformation syndromes [16,17].

3.2.2 Single-Gene Disorders

By definition, single-gene disorders arise as a result of variation in one or both alleles of a gene on an autosome

or sex chromosome or in a mitochondrial gene. There have been many investigations into the overall frequency of single-gene disorders. Many early estimates were misleadingly low due to underascertainment, especially of late-onset disorders (e.g., familial hypercholesterolemia, adult polycystic kidney disease, and Huntington disease). Carter [18] reviewed the earlier literature and estimated an overall incidence of autosomal dominant traits of 7.0 in 1000 live births, of autosomal recessive traits of 2.5 in 1000 live births, and of X-linked disorders of 0.5 in 1000 live births. This gave a combined frequency of 10 in 1000 live births (1%). At that time, approximately 2500 single-gene disorders had been delineated. The number of recognized Mendelian phenotypes has since more than doubled, and these new entities include several particularly common conditions (e.g., familial breast cancer syndromes, with a combined estimated frequency of five in 1000; hereditary nonpolyposis colon cancer syndromes, with a combined frequency of five in 1000). In addition, new technologies for DNA analysis have revealed a higher-than-expected frequency of generally asymptomatic people with one or two variant alleles at a locus [1]. For example, up to 1% of the population has an allele for von Willebrand factor, but many of these people have few or no symptoms, so again there is the problem of the imprecise and variable boundary between a harmless variant and a clinically important one. Furthermore, DNA analysis has shown that for some disorders, such as fragile X syndrome, relatives of an affected individual may harbor a premutation that has the potential for expansion to a full deleterious mutation in an offspring. The prevalence of such premutation carriers may be as high as one in 178 females for fragile X syndrome [19], and female heterozygotes may present with premature ovarian failure.

The frequencies of many single-gene disorders show population variation. Geographic variation may be explained by selection, by founder effects, or attributed to random genetic drift. Selection has resulted in a carrier frequency of one in three for sickle cell anemia in parts of equatorial Africa, and the Afrikaners of South Africa have a high frequency of variegate porphyria and familial hypercholesterolemia due to a founder effect. The carrier frequency for mutations in *HFE*, one of the genes responsible for hemochromatosis, is one in 10 in individuals of Celtic ancestry. As another example, heterozygosity for a mutation in *F11* associated with

deficiency of coagulation factor 11 causes a bleeding disorder of variable severity. The prevalence of heterozygous mutations varies markedly among populations, with Ashkenazi Jews having 9%–12%, East Asians having 0.45%, and Finns having 0.03% [20]. Undoubtedly, more single-gene disorders are going to be delineated. In theory, at least one per locus will eventually be recognized (about 20,000) minus those with no or a mild phenotype and minus those incompatible with establishment/continuance of a pregnancy. Increasing the total are those loci for which different mutations cause entirely different phenotypes. For example, mutations in *LMNA* can result in at least 13 distinct disorders [21]. There is also overlap with the multifactorial category. For example, many patients with acute intermittent porphyria are asymptomatic in the absence of an environmental trigger, and epistatic involvement of other genes is believed to contribute to intrafamilial phenotypic variation for patients with the same mutation. As more gene–environment and gene–gene interactions are identified, the boundary between single-gene disorders and multifactorial disorders will become further blurred. Among children diagnosed with a developmental disorder, 40% had a confirmed genetic diagnosis. Consanguineous parents are more likely to have affected children (34% vs. 22%) [22].

3.2.3 Mitochondrial Disorders

Metabolic defects in the respiratory chain can be due to mutations in autosomal or X-linked genes or to mutations in the genes encoded by the mitochondrial chromosome (mtDNA). Because mitochondria are inherited from females, all their offspring are at risk for the defect in oxygenation present in mother. If all of mother's mtDNA carries the mutation (homoplasmy), then all of the offspring will as well. More frequently, only a fraction of female mtDNA carries the mutation (heteroplasmy), so the offspring will inherit variable proportions of mutant mtDNA and their clinical features will vary in severity [23].

3.2.4 Multifactorial Disorders

Multifactorial disorders result from an interaction of one or more genes with one or more environmental factors. Thus, in effect, the genetic contribution predisposes the individual to the actions of environmental agents. Such an interaction is suspected when conditions show an increased recurrence risk within families that does

not reach the level of risk or pattern seen for single-gene disorders and when identical twin concordance exceeds that for nonidentical twins but is less than 100%. For most multifactorial disorders, however, the nature of the environmental agent(s) and the genetic predisposition are currently unclear and are the subject of intensive research efforts. The ability to conduct genome-wide association studies (GWAS) has accelerated progress in this area [24]. A relevant example is asthma.

Multifactorial disorders are believed to account for approximately one-half of all congenital malformations and to be relevant to many common chronic disorders of adulthood, including hypertension, rheumatoid arthritis, psychoses, and atherosclerosis (complex common disorders). The former group had an estimated frequency of 46.4 per 1000 in the British Columbia Health Surveillance Registry [25]. In addition, a multifactorial etiology is suspected for many common psychological disorders of childhood, including dyslexia (5%–10% of the population), specific language impairment (5% of children), and attention-deficit/hyperactivity disorder (4%–10% of children). Hence the multifactorial disorder category represents the most common type of genetic disorder in both children and adults. Spontaneous preterm birth is contributed to by maternal genetic predispositions, whose deleterious effects can be stimulated by external factors [26]. Multifactorial disorders also show considerable ethnic and geographic variation. For example, talipes equinovarus is about six times more common among Maoris than among Europeans [27], and neural tube defects were once 10 times more frequent in Ireland than in North America [28].

Often ignored, both intellectually and in research, are genotypes that reduce susceptibility to potentially harmful environmental factors. An understanding of alleles that provide protection from disease or increase longevity will yield insight into pathogenesis as well as novel approaches to therapy and prevention. Genome-wide association study (GWAS) is one approach to deciphering the genetic contributions to overt diseases as well as benign variations [29].

3.2.5 Somatic Cell Genetic Disorders

Somatic cell mutation is a natural developmental process in the immune system, but it is also responsible for a significant burden of genetic disease. This includes somatic or germline mosaicism for single-gene disorders [30], as well as mutations that give rise to cancer. Cancer cells tend

to have accumulated multiple mutations; the first step in the cascade of mutations may be inherited (i.e., involving germ cells and all somatic cells). Carcinogens are important causes of noninherited mutations, and genetic susceptibility is suspected to account for individual variation in risk on exposure. Somatic cell genetic disorders might also be involved in other clinical conditions, such as autoimmune disorders and the aging process.

3.3 MORBIDITY AND MORTALITY DUE TO GENETIC DISEASE

The same general difficulties that pertain to the frequency estimates for genetic disorders also apply to estimates of the contribution of the various types of genetic disorders to morbidity and mortality during pregnancy, in childhood, and in adulthood. Hence these figures should be taken as minimum estimates.

3.3.1 Conception and Pregnancy

One in 15 recognized pregnancies spontaneously miscarry, and a higher percentage (up to 50%) of conceptions are lost before recognition of the pregnancy [10]. The majority of these losses are caused by numerical chromosomal abnormalities.

3.3.2 Childhood

Since the turn of the century in many developed countries, advances in medicine and public health have resulted in a gradual decline in the contribution of environmental factors to childhood morbidity and mortality. The result of these changes has been to raise genetic disorders to greater prominence. By contrast, in developing countries, nongenetic causes of childhood mortality continue to predominate. An idea of the contribution of genetic disease to morbidity can be judged from the prevalence of such diseases among pediatric inpatients. In reviewing 4115 inpatients, Hall and colleagues [31] found multifactorial disease in 22.1%, a single-gene disorder in 3.9%, and a chromosomal disorder in 0.6%. Thus, more than one in four pediatric inpatients has a genetic disorder in one of these categories, compared with the general population frequency estimate of one in 20 by age 25 years. This does not include morbidity that does not lead to inpatient admission. An updated survey by McCandless and colleagues [32] revealed that 71% of children admitted to the hospital had a disorder with a significant genetic component. Stevenson and Carey [33] found that 34.4% of deaths among children hospitalized in a tertiary care

center could be attributed to congenital anomalies, of which 16.7% were due to chromosomal abnormality and 11.7% were due to a recognized malformation syndrome.

3.3.3 Adulthood

In developed countries, the most common causes of death are cancer and cardiovascular disease. All cancers are now known to have a cumulative somatic cell genetic basis, and there is evidence for a major genetic contribution to cardiovascular disease. Single-gene disorders causing diabetes or high blood pressure are relatively uncommon, but multifactorial inheritance accounts for a large proportion of patients with premature vascular disease and systemic hypertension. Similarly, there is a growing recognition of the importance of multifactorial inheritance for many other common disorders of adulthood responsible for both morbidity and mortality.

REFERENCES

- [1] Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491:56–65.
- [2] de Ligt J, Veltman JA, Vissers LE. Point mutations as a source of de novo genetic disease. *Curr Opin Genet Dev* 2013;23:257–63.
- [3] Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
- [4] Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet* 2012;13:565–75.
- [5] Mozzi A, Pontremoli C, Sironi M. Genetic susceptibility to infectious diseases: current status and future perspectives from genome-wide approaches. *Infect Genet Evol* 2017. Sep 22. pii: S1567-1348(17):30334–9. <https://doi.org/10.1016/j.meegid.2017.09.028>. [Epub ahead of print].
- [6] An Y, Li H, Wang KJ, Liu XH, Qiu MX, Liao Y, Huang JL, Wang XS. Meta-analysis of the relationship between slow acetylation of N-acetyl transferase 2 and the risk of bladder cancer. *Genet Mol Res* 2015;14:16896–904.
- [7] Verma IC, Puri RD. Global burden of genetic disease and the role of genetic screening. *Semin Fetal Neonatal Med* 2015;20:354–63.
- [8] Jacobs PA, Browne C, Gregson N, Joyce C, White H. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J Med Genet* 1992;29:103–8.

- [9] Hassold T, Hunt P. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet* 2001;2:280–91.
- [10] Boue J, Boue A, Lazar P. Retrospective and prospective epidemiological studies of 1500 karyotyped spontaneous human abortions. 1975. *Birth Defects Res A Clin Mol Teratol* 2013;97:471–86.
- [11] Boue JG, Boue A. Chromosomal anomalies in early spontaneous abortion. (Their consequences on early embryogenesis and in vitro growth of embryonic cells). *Curr Top Pathol* 1976;62:193–208.
- [12] Templado C, Vidal F, Estop A. Aneuploidy in human spermatozoa. *Cytogenet Genome Res* 2011;133:91–9.
- [13] Rosenbusch B. The incidence of aneuploidy in human oocytes assessed by conventional cytogenetic analysis. *Hereditas* 2004;141:97–105.
- [14] Ayed W, Messaoudi I, Belghith ZHammami W, Chemkhi I, Abidi N, Guermani H, Obay R, Amouri A. Chromosomal abnormalities in 163 Tunisian couples with recurrent miscarriages. *Pan Afr Med J* 2017;28:99–104.
- [15] Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010;86:749–64.
- [16] Kirov G. CNVs in neuropsychiatric disorders. *Hum Mol Genet* 2015;24:R45–9.
- [17] Mikhail FM. Copy number variations and human genetic disease. *Curr Opin Pediatr* 2014;26:646–52.
- [18] Carter CO. Monogenic disorders. *J Med Genet* 1977;14:316–20.
- [19] Hantash FM, Goos DM, Crossley B, Anderson B, Zhang K, Sun W, Strom CM. FMR1 premutation Carrier frequency in patients undergoing routine population-based Carrier screening: insights into the prevalence of fragile X syndrome, fragile X-associated tremor/ataxia syndrome, and fragile X-associated primary ovarian insufficiency in the United States. *Genet Med* 2011;13:39–45.
- [20] Asselta R, Paraboschi EM, Rimoldi V, Menegatti M, Peyvandi F, Salomon O, Duga S. Exploring the global landscape of genetic variation in coagulation factor XI deficiency. *Blood* 2017;130:e1–6.
- [21] Capell BC, Collins FS. Human laminopathies: nuclei gone genetically awry. *Nat Rev Genet* 2006;7:940–52.
- [22] Best S, Rosser E, Bajaj M. Fifteen years of genetic testing from a London developmental clinic. *Arch Dis Child* 2017;102:1014–8.
- [23] Davis RL, Liang C, Sue CM. Mitochondrial disease. *Hanb Clin Neurol* 2018;147:125–41.
- [24] O’Rielly DD, Rahman P. Clinical genetic research 2: genetic epidemiology of complex phenotypes. *Methods Mol Biol* 2015;1281:349–67.
- [25] Baird PA, Anderson TW, Newcombe HB, Lowry RB. Genetic disorders in children and young adults: a population study. *Am J Hum Genet* 1988;42:677–93.
- [26] Strauss 3rd JF, Romero R, Gomez-Lopez N, Haymond-Thornburg H, Modi BP, Teves ME, Pearson LN, York TP, Schenkein HA. Spontaneous preterm birth: advances toward the discovery of genetic predisposition. *Am J Obstet Gynecol* 2017;32484–5. S0002-9378(17).
- [27] Cartlidge I. Observations on the epidemiology of club foot in Polynesian and Caucasian populations. *J Med Genet* 1984;21:290–2.
- [28] McDonnell RJ, Johnson Z, Delaney V, Dack P. East Ireland 1980-1994: epidemiology of neural tube defects. *J Epidemiol Community Health* 1999;53:782–788.
- [29] Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* 2018;19(2):110–124. <https://doi.org/10.1038/nrg.2017.101>. [Epub 2017 Dec 11. Review].
- [30] Youssoufian H, Pyeritz RE. Mechanisms and consequences of somatic mosaicism in humans. *Nat Rev Genet* 2002;3:748–58.
- [31] Hall JG, Powers EK, McIlvaine RT, Ean VH. The frequency and financial burden of genetic disease in a pediatric hospital. *Am J Med Genet* 1978;1:417–36.
- [32] McCandless SE, Brunger JW, Cassidy SB. The burden of genetic disease on inpatient care in a children’s hospital. *Am J Hum Genet* 2004;74:121–7.
- [33] Stevenson DA, Carey JC. Contribution of malformations and genetic disorders to mortality in a children’s hospital. *Am J Med Genet A* 2004;126A:393–7.

Genome and Gene Structure*

Madhuri R. Hegde^{1,2}, Michael R. Crowley^{1,2}

¹Department of Human Genetics, Emory University, Atlanta, GA, United States

²The Department of Genetics, The University of Alabama at Birmingham, Birmingham, AL, United States

The past decade of biological research has focused heavily on the human genome, and the Human Genome Project has had a significant impact on biomedical research. Our genetic material is encoded in two genomes: nuclear and mitochondrial. Both genomes reflect the molecular evolution of humans, which started about 4.5 billion years ago. The function of the human genome is to transfer information reliably from parent to daughter cells and from one generation to the next. At particular developmental times and in specific tissues, the transcriptional machinery initiates programmed patterns of gene expression that are dictated by chromatin structure and the activities of transcriptional regulatory factors. Gene expression is followed by the processes of splicing, translation, and protein localization, ultimately leading to synthesis of the protein and RNA molecules that mediate cellular function. Variability in genome structure, including single nucleotide differences and larger-scale variations at the genome level between humans, dictates the traits we manifest as well as the diseases to which individuals are predisposed.

4.1 INTRODUCTION: COMPOSITION OF THE NUCLEAR HUMAN GENOME

The present assembly of human DNA sequence contains approximately 3.1 billion bp, which covers most of the nonheterochromatic portions of the genome and

contains about 250 gaps. The 3.2 Mbp are packed into 22 pairs of chromosomes and two sex chromosomes, X and Y. The human chromosomes are not equal sizes; the smallest, chromosome 21, is 54 Mbp long, and the largest, chromosome 1, is about 249 Mbp long (Fig. 4.1 and Table 4.1). From a functional point of view, the genomic sequences are distinguished by genes, pseudogenes, and noncoding DNA, and only a minute fraction of the sequences code for proteins (approximately 2%). There are many pseudogenes (0.5%) but most of the genome consists of introns and intergenic DNA. Almost half of intergenic sequences consist of transposons. Gene clusters have evolved from several duplication events in deep evolutionary time, and these include clusters such as the *HOX* and *globin* clusters. The chemical structure of genetic material as well as the storage, processing, and transfer of genetic information from one generation to the next are similar in all living organisms. Thus, it was expected that the complexity of the human phenotype would be explained by a significantly higher number of genes in humans compared with simpler organisms. Surprisingly, instead of the predicted 50,000–150,000 human genes, the sequence of the human genome revealed about 20,000–25,000 genes, similar to the number of genes in many other organisms. However, analysis of the genome and its products has revealed complexity in the form of exquisite temporal and spatial regulation of gene alternative transcripts,

* This article is a revision of the previous edition article by David H. Cohen, vol. 1, pp. 61–80, © 2007, Elsevier Ltd.

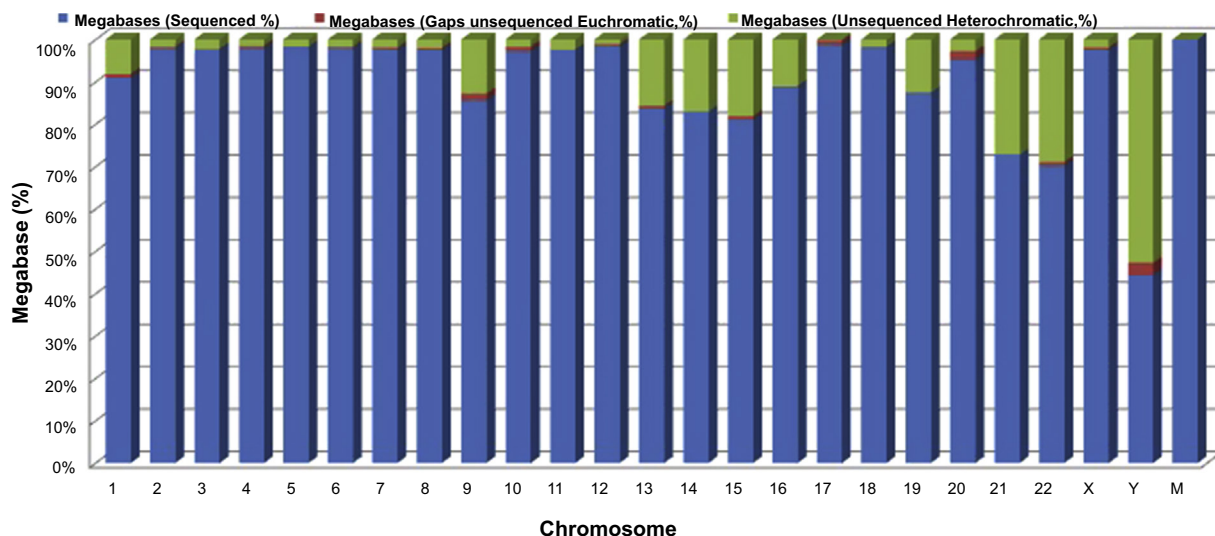


Figure 4.1 The percent proportion of the sequenced (blue), unsequenced gaps from euchromatic regions (red), and unsequenced gaps from heterochromatic regions (green) of the human genome, listed by chromosome numbers. Statistics are from the NCBI Build 36.1, UCSC assembly of March 2006, Assembly hg18. (Data from <http://genome.cse.ucsc.edu/goldenPath/stats.html#hg18>.)

TABLE 4.1 Physical Sizes (Megabasespairs, Mbp) of Human Chromosomes

Chromosome	Size (Mbp)
1	249
2	237
3	192
4	183
5	174
6	165
7	152
8	135
9	132
10	132
11	132
12	123
13	108
14	105
15	99
16	84
17	81
18	75
19	69
20	63
21	54
22	57
X	141
Y	60

and their expression derived from a single locus, which ranges from 35% to 60%, but there remains an uncertainty in determining the extent to which these reflect functional splice variants or splice errors (including tissue-specific variants) for most genes. The complex posttranslational modifications of proteins also create a yet-undiscovered endless diversity in gene products and their functions.

Since the initial sequence of the human genome was determined, we have gained tremendous new insights into genome structure such as how the sequence and structure define the complex functions of human cells and how genome architecture can be altered to produce disease. Additionally, how our genome compares with the sequences of the genomes of closely and distantly related organisms has provided insights into evolutionary conservation of gene and protein functions as well as our origins as a species. Technological innovations have facilitated defining the genomic sequences of many individual humans, particularly the differences that distinguish individuals, allowing a fuller description of the history of our species and the traits we manifest. We are also at the point where we can conceive of understanding at the molecular level the genetic contributions to disease across the human population, facilitating targeted medical intervention based on these findings.

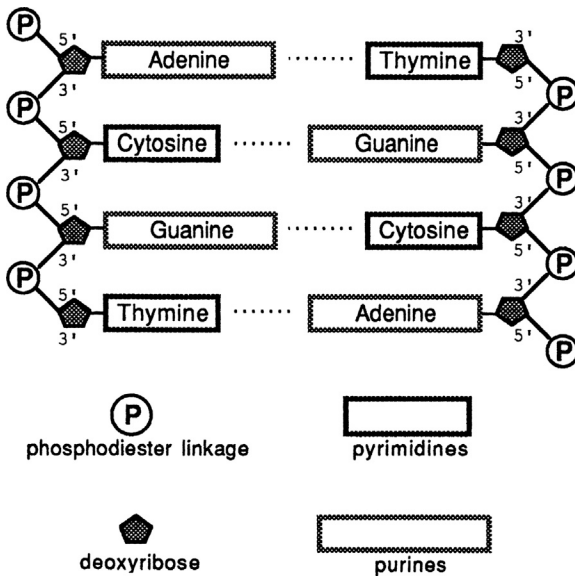
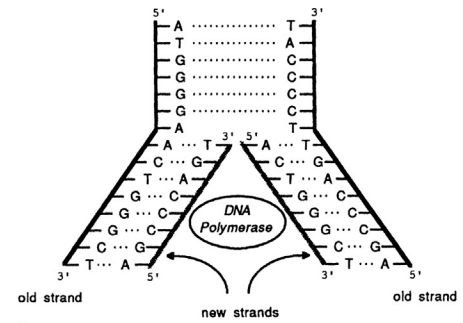


Figure 4.2 Complementary structure of double-stranded DNA.

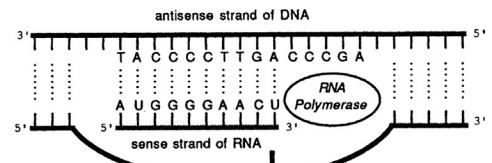
4.2 DOUBLE HELIX STRUCTURE, DNA REPLICATION, TRANSCRIPTION, AND MEIOTIC RECOMBINATION

4.2.1 Double Helix

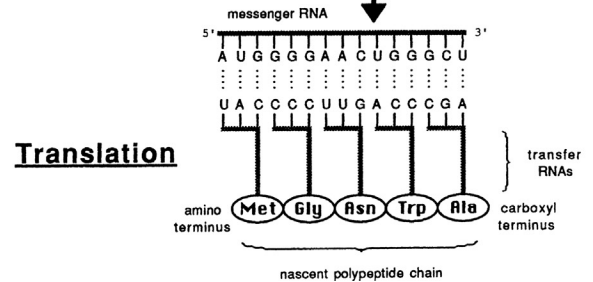
The function of the human genome is to transfer information reliably from parent cells to daughter cells and from one generation to the next. This is carried out in a semiconservative manner. One of the two parental DNA strands of a double helix remains intact in every cell division, serving as a template for copying the sequence. The two DNA strands form the double helix by hydrogen bonding between the nitrogenous bases: guanine (G) pairs with cytosine (C) and adenine (A) pairs with thymine (T) (Fig. 4.2). The hydrogen bonds formed between these pyrimidine–purine pairs (guanine and adenine are purines; cytosine and thymine are pyrimidines) stabilize the double helix and ensure that the two complementary strands remain together and in register. The strands are oriented antiparallel to each other, meaning that they run in opposite directions: One strand is oriented in a 5′–3′ direction, whereas the other is in a 3′–5′ direction. These opposite strands are often referred to as the Watson strand and the Crick strand after the two investigators who initially described the DNA double helix, James Watson and Francis Crick,



Replication



Transcription



Translation

Figure 4.3 Flow of genetic information.

with X-ray diffraction images produced by Rosalind Frankland.

4.2.2 Replication

Genetic information is preserved and transmitted via DNA replication, a process that produces two identical copies of the DNA. During this process, the two parental strands separate, and each serves as a template for synthesis of a new complementary strand by an enzyme called DNA polymerase (Fig. 4.3). As a consequence, each daughter cell inherits one strand of the parental duplex. Every DNA molecule thus contains a “young” strand that was synthesized in the parental cell during DNA replication and an “old” strand that was inherited from the parental cell and synthesized in the grandparental cell. This semiconservative manner of replication

guarantees transmission of intact information from one generation to the next. Remarkably, the genome copies itself through millions of cell divisions during an individual's life with amazing precision. The error rate of about 1×10^{-8} per bp per generation means that replication of 3×10^9 bp comprising the human genome leads to about 60 new single base mutations per individual.

4.2.3 Transcription

Only about 1%–1.5% of the genome is reflected in the population of mature protein-coding transcripts. Protein-coding genes are transcribed from DNA into messenger RNA (mRNA) (see Fig. 4.3). A single gene can give rise to multiple transcripts through alternative splicing and alternative sites of transcription initiation and termination, generating functional diversity. During transcription, the DNA duplex unwinds, and one of the strands serves as the template for the synthesis of a complementary RNA strand. RNA is distinguished from DNA by the presence of uracil instead of thymine, ribose instead of deoxyribose, and a different three-dimensional folding pattern. The mRNA molecules are single stranded and function as the vehicle for translating genomic information into a protein.

In eukaryotes, genes are transcribed by one of the three different RNA polymerases (I, II, and III, respectively). RNA polymerase I transcribes ribosomal RNAs (rRNAs, a structural noncoding RNA) (except for 5S rRNA); RNA polymerase II transcribes mRNAs, microRNAs (miRNAs), and many long noncoding RNAs (lncRNAs); and RNA polymerase III transcribes 5S rRNA, transfer RNA (tRNA), and other small RNAs. In addition to RNA polymerase, initiation of gene transcription requires other proteins, so that multiple factors form the complex responsible for transcriptional initiation. This complex gets attached to the initiation site of transcription at the 5' end of the gene (the promoter) and determines which genes are transcribed in different cell types or during different developmental stages. The transcription factors (TFs), along with the activities of *cis*-acting enhancer and inhibitor sequences, also determine the level of gene expression (Fig. 4.4). The enhancers and inhibitors can be located near the promoter of a gene, at the 5' or 3' side of the promoter, or at significant distances away from the transcription start site. Such sequences are commonly found within first introns of many mammalian genes. These regions of the genome also contain the “ultraconserved elements”

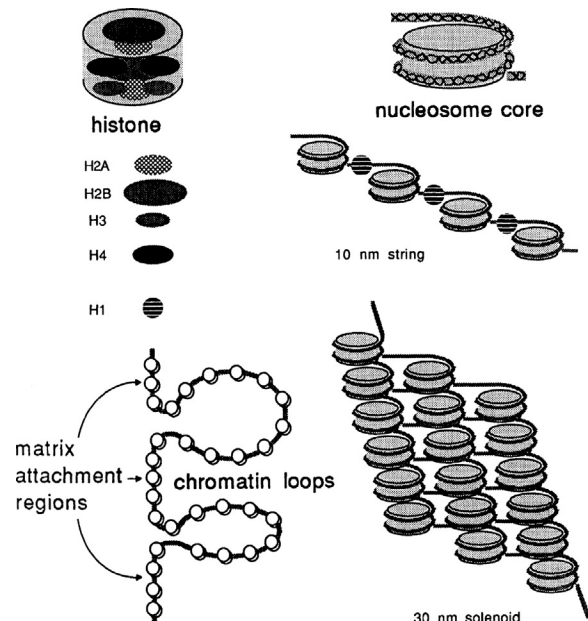


Figure 4.4 Packaging of DNA into chromatin.

(UCEs), which are extraordinarily highly conserved between evolutionary distant species. The human genome contains 481 such regions, which are >200 bp in length and are 100% invariable between human, rat, and mouse sequences.

The DNA strand that is similar to the transcribed mRNA sequence is referred to as the sense strand. The DNA sequence that serves as the transcriptional template is referred to as the antisense strand. However, in recent years it has become more apparent that transcription occurs from both strands of the DNA. There is a substantial amount of transcription of the genome, between 65% and 80% depending on the study. Yet, as stated here, only 1.0%–1.5% of the genome encodes for proteins. The remainder of the transcription is of miRNA, small structural and regulator RNAs, and long noncoding RNAs.

4.2.4 Meiotic Recombination

Meiotic recombination [1] refers to the reciprocal physical exchange of chromosomal DNA between the parental chromosomes and occurs at meiosis during spermatogenesis and oogenesis, serving to ensure proper chromosome segregation. During the four-strand stage of meiosis, two duplex DNA molecules (one from each parent) form a hybrid, and a single strand of

one duplex is paired with its complement from the other duplex. Single-stranded DNA is exchanged between the homologous chromosomes, and the process involves DNA strand breakage and resealing, resulting in the precise recombination and exchange of DNA sequences between the two homologous chromosomes. This process is highly efficient and does not usually result in mutations at the sites of recombination. Recombination thus shuffles genetic material between homologous chromosomes, generating much of the genetic diversity that characterizes differences between individuals, even within the same family.

The frequency of recombination between two loci along a chromosome is proportional to the physical distance between them, and historically, this provided the basis for defining the genetic distance between loci, allowing genetic maps to be constructed. The genetic proximity of two loci is measured by the percentage of recombination between them; a map distance of 1 centimorgan (cM) indicates 1% recombination frequency between the two loci. The human genome sequence has made it possible to compare genetic and physical distances and to analyze variations in recombination frequency in different chromosomal regions. On average, 1 million bp (1 Mb) correspond to 1 cM (1% recombination frequency). However, there is a tremendous local variation between individual chromosomes and among particular chromosomal regions. For example, the average recombination rate is higher in the short arms of chromosomes and at the distal segments of the arms but overall is suppressed near the centromeres. There is also a significant variation in the recombination rates between the sexes, with 1.6-fold more recombination on average in females relative to males. On average, female recombination is higher at the centromeres and male recombination is higher at the telomeres [2].

4.2.5 DNA and RNA Synthesis

Each chromosome in the human cell consists of a continuous double-helical DNA strand; an average chromosome contains about 4–5 cm of DNA. The polarity of a single-stranded nucleic acid is defined by the position of the phosphodiester bonds, which connect the 3' hydroxyl group of one nucleoside to the 5' hydroxyl group of the next (see Fig. 4.2). A nucleoside is composed of a purine or pyrimidine base and a deoxyribose (in DNA) or ribose (in RNA), and a nucleotide is composed of a nucleoside and one or more phosphate

groups. The phosphate groups in a single nucleoside triphosphate are attached to the ribose or deoxyribose moiety via the 5' hydroxyl residue, and two of the three phosphates are removed during the incorporation of each nucleoside triphosphate into DNA or RNA. When the new DNA strand is synthesized during replication or copied during transcription, the polymerase enzymes add new nucleotides to the 3' hydroxyl group of a growing polynucleotide chain. The new strand of DNA or RNA is thus synthesized in the 5'–3' direction, so the parental or template DNA strand is read in the 3'–5' direction. If two proteins are encoded by adjacent genes that lie on different strands along the chromosome, those genes are said to have opposite transcriptional orientations.

4.3 ORGANIZATION OF GENOMIC DNA

The 3 billion bp that constitute the human genome are packaged into 22 pairs of autosomes and the X and Y sex chromosomes. The chromosomal DNA can be divided into regions of heterochromatin and euchromatin: heterochromatin represents the “tightly” packed regions of chromosomal DNA and euchromatin represents the “loose” regions, which are generally the actively transcribed DNA regions. Using cytogenetic staining methods, differently packed chromosomal regions can be viewed as G (Giemsa staining)-bands, with the banding pattern characteristic of each individual chromosome providing the basis for the cytogenetic identification of each human chromosome. From early on, the dark G-bands were considered to reflect “gene-poor and GC-poor” regions and are comparatively more condensed and more AT-rich, and they replicate later than the DNA within the lighter staining bands; the sequence of the human genome has proved this concept to be accurate. The light staining bands correspond to the R-bands via an alternative staining technique. However, human genome sequence information has revealed that there can be a tremendous variation in the GC content across chromosomal regions.

Each of the 23 pairs of human chromosomes contains a single DNA duplex extending between the two telomeres. When the DNA in the human genome is stretched from one end to the other, its length would be longer than 1 m (approximately 3 ft) long! Remarkably, compacting the DNA by greater than 100,000-fold, which is required to fit the chromosomes into the

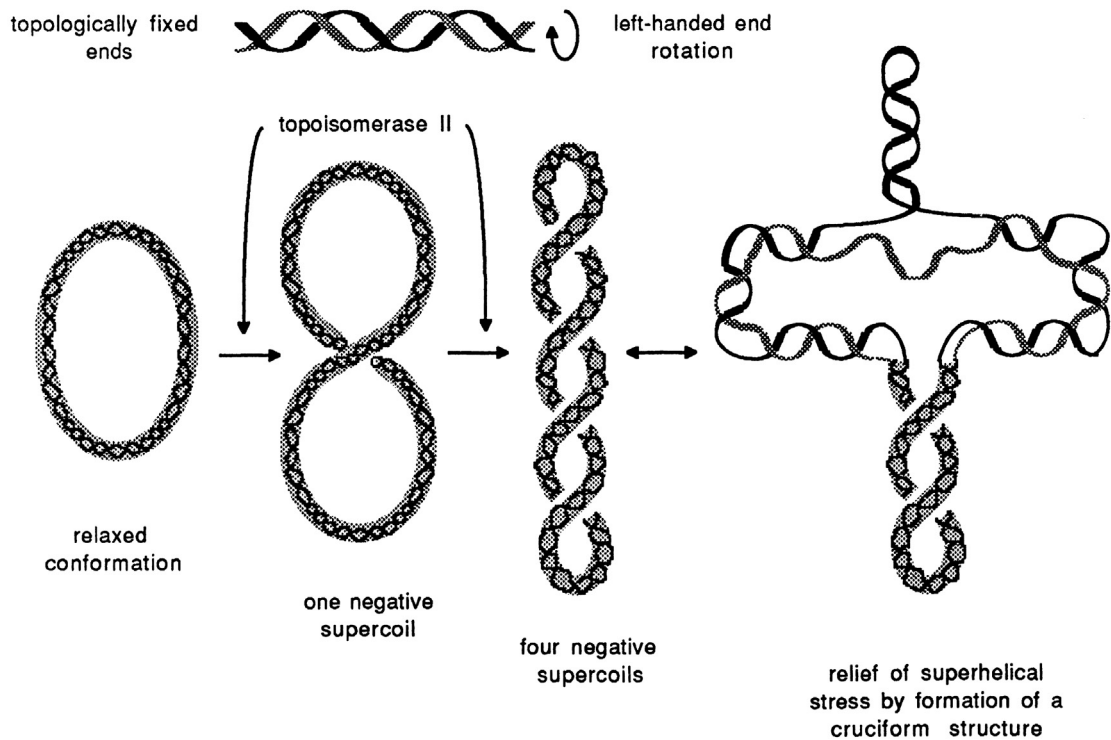


Figure 4.5 Superhelical turns in DNA.

nucleus, is achieved by coiling and folding the double helix into a series of progressively shorter and thicker structures (Fig. 4.5). Proteins that bind to DNA help direct and organize this folding, and the folded complex of DNA and protein is referred to as chromatin.

4.3.1 Nucleosomes and Higher Order Chromatin Structure

In addition to compacting the genetic material to fit into the nucleus, chromatin condensation can regulate accessibility of the DNA for transcription and other processes. The simplest level of chromatin structure is the organization of DNA and histones into nucleosomes [2]. Each nucleosome is a 147-bp-long segment of DNA tightly wrapped almost two times around an octamer histone core. This octamer core contains two molecules each of the histones H2A, H2B, H3, and H4. Nucleosomes are the fundamental feature of all eukaryotic DNA, and the sequences of the core histones are well conserved among even, distantly related species. A fifth histone, H1, binds to a particular modification on the core histone H3 at lysine 9 and its sequence is less well conserved. A region

of linker DNA, about 60 bp in length in humans, usually separates adjacent nucleosomes, so that nucleosomes are for the most part regularly spaced.

Nucleosomes represent the first level in the packaging of naked DNA into chromatin and appear in the electron microscope as strings of 11-nm “beads.” Previous models have the next level in packaging as the coiling of the nucleosomes to form a 30-nm structure with a solenoid conformation. However, recent data suggest the 30-nm solenoid structure is an artifact of X-ray diffraction studies of chromatin *in vitro* and does not reflect chromatin organization in the nucleus. Indeed, studies by Maeshima and colleagues [3] failed to find the 30-nm structure using small-angle X-ray scattering on purified nuclei or chromosomes in the absence of contaminating ribosomes. More recently, a novel method of decorating the chromatin with photoactivatable long polymers of diaminobenzidine (DAB), staining with OsO_4 , and visualization by multi-tilt scanning electron microscopy tomography reveals *in situ* chromatin to be disordered polymers of 5 and 24 nm with no evidence of the 30-nm fibers (Fig. 4.6) [4]. Epigenetic modifications,

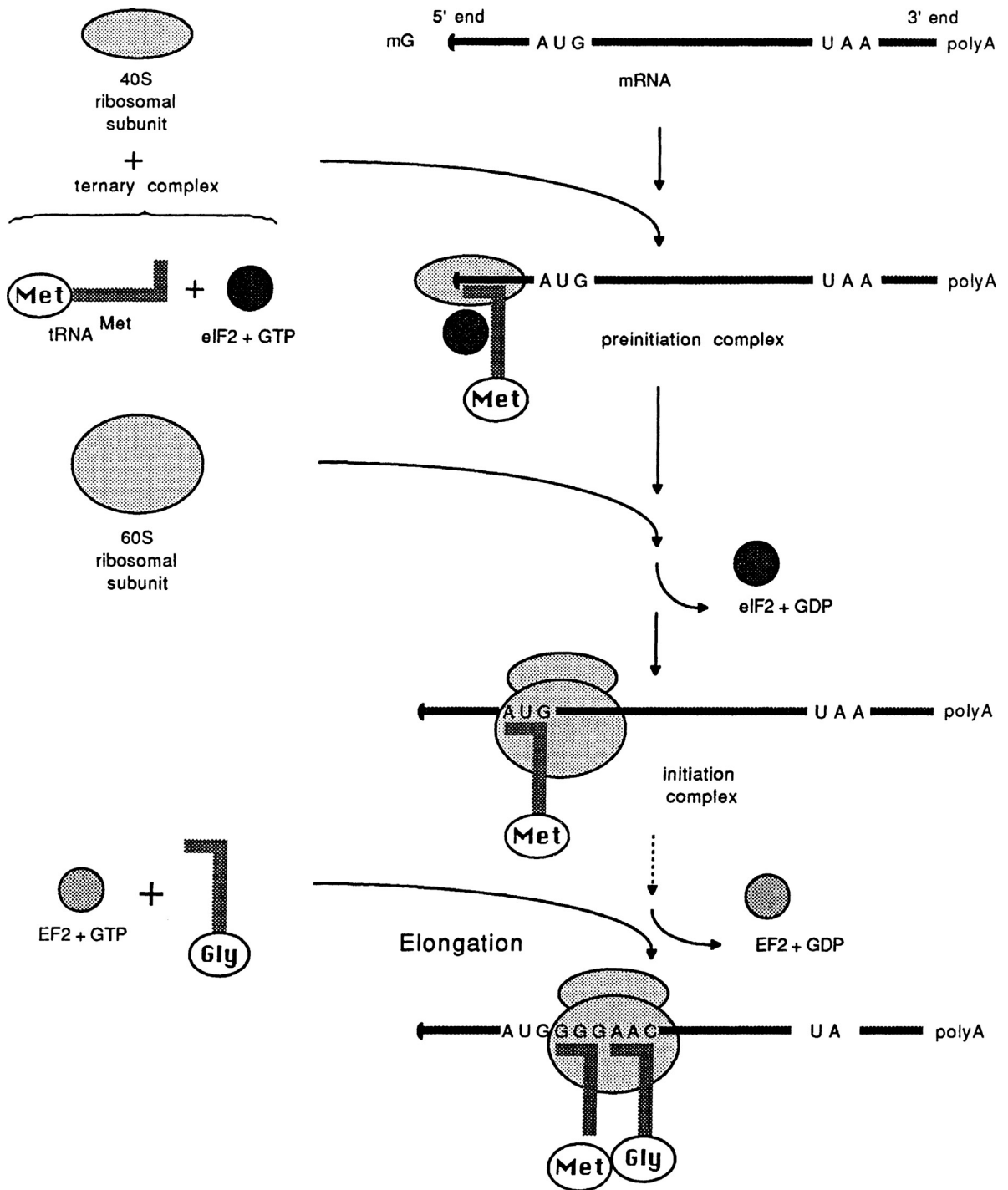


Figure 4.6 Eukaryotic gene structure and the pathway of gene expression.

primarily DNA methylation, and histone modifications also regulate the structure and dynamics of chromatin folding. Additional levels of folding can compress DNA into 300-nm fibers and a nearly 1000-nm metaphase chromatid.

Following separation of sister chromatids at meiosis II, or sister chromosomes at mitosis, the DNA decompacts as the cell returns to interphase or the G₁ phase of the cell cycle. Chromosomes do not, however, distribute randomly throughout the nucleus, nor does the DNA of every chromosome simply fill the nucleus. Instead the chromosomes reside in discrete territories, which can be observed by staining chromosomes with interphase spectral karyotyping (SKY karyotyping or “painting” each chromosome with a unique color) [5].

The human genome is also compartmentalized into large (>300-kb) segments of DNA that are homogeneous in base composition and are referred to as isochores, based on sequence analysis and compositional mapping. L1 and L2 are GC-poor (“light”) isochore families, which represent 62% of the genome. The heavy H1, H2, and H3 isochores are GC-rich. This also corresponds to the G-bands, which are composed of GC-poor isochores, whereas the R-bands are composed of the GC-rich isochores and some GC-poor isochores.

4.3.2 Euchromatin and Heterochromatin

During metaphase, the entire chromosome is highly condensed, but at other times, most chromatin is organized into one of two separate compartments within the nucleus. Some portions of the genome remain highly condensed throughout the entire cell cycle, associate with the nuclear periphery and the nucleolus, are gene poor, are enriched in long interspersed nuclear element (LINE)-1-type repetitive elements, and are late replicating during S phase, which are termed heterochromatin. Many areas of heterochromatin are located close to chromosome centromeres and at the telomeres of acrocentric chromosomes (chromosomes with the centromere at one end), contain highly repetitive or simple-sequence DNA instead of genes, and may play a structural role in chromosome organization. In addition, heterochromatin can be subdivided into facultative and constitutive heterochromatin. Constitutive heterochromatin is found at highly repetitive features, such as α -satellite DNA and transposons. Constitutive heterochromatin is also identified by its high compaction, the absence of histone acetylation, and the presence of

histone hypermethylation, especially histone 3-lysine 9 monomethylation (H3K9me). Facultative heterochromatin, in contrast, results from tissue- or cell-specific downregulation of transcription of specific genes, such as during cell or tissue differentiation. Facultative heterochromatin is found in the inactive X chromosome in female cells, which contains genes but generally does not express them. This is partially due to a high degree of chromatin condensation and histone modification that does not allow access to the DNA by the transcriptional machinery. Inactive X heterochromatin remains highly condensed during the lifetime of somatic cells, but in germ cells, during oogenesis, it becomes active and euchromatic by the time of entrance into meiosis.

Formation of heterochromatin involves proteins that help direct condensation and packaging to achieve assembly of these proteins and DNA into heterochromatin. The molecular mechanisms behind heterochromatin formation have been revealed by studies of inactivation of the second X chromosome in female cells [6]. Heterochromatin spreading is also thought to be involved in the initiation of X-chromosome inactivation that occurs in all female cells because condensation begins from a specific site on the X chromosome, the X-inactivation center, which in humans is located on Xq13. At this center, the X-inactivation-specific transcript (*XIST*) gene encodes a 17- to 19-kb-long noncoding RNA, which is transcribed only from the inactive X chromosome and acts in *cis* to initiate X-inactivation. During early embryonic development, X-inactivation is random, so females are mosaics with respect to whether the maternal or paternal X is active in each cell. Once established in somatic cells, the inactive X is stable through replication and cell division (i.e., the same inactive X will be inactive in all daughter cells).

XIST RNA inactivates genes at a significant distance from the gene that encodes it. Although *XIST* is essential for X-inactivation, regulatory genes including *TSIX* and a variety of TFs control *XIST*. While *XIST* is essential for the initiation of X-chromosome inactivation, it is not required for the maintenance of X-inactivation. The detailed analyses of the process of X-inactivation have informed more general aspects of heterochromatin formation and gene inactivation. A more general form of heterochromatin formation is directed by the removal of acetyl lysine histone marks through the increased association of histone deacetylases and the recruitment of histone methyltransferases [7]. Acetylation of H3

histones adds a positive charge to the histone tail that results in repulsion from the negatively charged DNA strand and is associated with open chromatin and active gene transcription. In contrast, removal of the acetyl group from the H3 histone tails in association with methylation of different H3 lysine residues, specifically H3K9 and H3K27, is associated with histone H1 recruitment, silencing of transcription, and heterochromatin formation.

Euchromatin, on the other hand, has an open chromatin configuration, is gene-rich, is early replicating, is enriched in short interspersed nuclear element (SINE) transposable elements, and is found in the center of the nucleus. There are also several histone modifications that delineate euchromatin including H3K27 acetylation and H3K36 methylation. Indeed, differential histone modifications define chromatin type and hence transcriptional potential. Most gene transcription occurs within the euchromatic regions of the genome.

4.3.3 Centromeres and Telomeres

Special features of the DNA molecule that compose each human chromosome are required at the centromeres and the telomeres. Located close to most centromeres are many copies of a 171-bp α -satellite repeat that forms the core of the centromere. These sequences bind structural proteins that serve as a site for kinetochore formation and spindle attachment during metaphase. Certain alphoid repeats are found close to the centromeres of all chromosomes, while others are specific for one or a small number of chromosomes. The proteins and DNA sequences that make up the centromeres must also ensure that the two daughter chromatids are partitioned to different cells during cytokinesis [8].

As template-directed replication of DNA can be only performed in a 5′–3′ direction, one strand of each duplex cannot be fully replicated at its 3′-terminus by the DNA polymerase. Therefore, the telomeres of each human chromosome contain many copies of a short repeat 5′-TTAGGG-3′, which can be replicated using the enzyme telomerase [9]. This enzyme has an RNA component, which itself serves as a template and can elongate the 5′-TTAGGG-3′ repeat in a manner that does not depend on the DNA strand.

4.3.4 Repeat Content of the Human Genome

Less than 5% of the human genome sequence encodes proteins, whereas repeat sequences account for at least

50% of the sequence. The human genome contains long stretches of DNA sequences of various lengths that also exist in variable copy number [10,11]. The repeats fall into five categories: (1) transposon-derived repeats, often referred to as interspersed repeats; (2) inactive retrotransposed copies of cellular genes (referred to as processed pseudogenes); (3) segmental duplications (SDs) (also known as low copy repeats [LCRs]) consisting of blocks of around 10–300 kb that have been copied from one region of the genome into another; (4) blocks of tandemly repeated sequences such as centromeres and ribosomal gene clusters; and (5) simple-sequence repeats consisting of direct repeats of short sequences such as (CA)_n or (CGG)_n, which have been extremely important for human genetic studies as they have been used as genetic markers (see Section 4.3.6).

Transposable elements in humans, as in all mammals, fall into four types: LINEs, SINEs (including Alu sequences), long terminal repeat (LTR) retrotransposons, and DNA transposons. Both the number and age of transposable elements in the human genome are strikingly different from those in other species. The density of transposable elements is much higher in humans than in other species, and the human genome contains more ancient transposons than do other species. It thus appears that these repeats have survived because of a significant evolutionary advantage although their selective advantage and precise function are not well understood. Some chromosomes are extremely crowded with repeat elements (e.g., a 500-kb region on the short arm of the X chromosome has an overall transposable element density of 89%), whereas other chromosomal regions are nearly devoid of repeats (e.g., the homeobox gene clusters). Several transposable elements, LINEs and SINEs, are still active within the human genome and have been associated with gene disruption, thereby causing disease. The human genome does not contain any active DNA transposons. An important distinction between LINEs and SINEs and the DNA transposons is their mechanism of movement within a genome. LINEs and SINEs are types of retrotransposons and therefore use an RNA intermediate through which to move. This is often referred to as a copy-and-paste mechanism. Notably, SINEs do not encode the proteins required for transposition but rather co-opt the proteins produced from LINEs. DNA transposons (of which there are many active transposons in species other than humans) are excised from their original location, usually during

DNA replication, and reinserted into the genome in a new location. These transposons use a cut-and-paste mechanism for moving through genomes.

Pseudogenes (full and partial) are regions of DNA with many sequence elements of a potential transcriptional unit (e.g., promoter, protein-coding region, splice junctions, etc.), yet do not code for a functional product. They can originate after gene duplication when the duplicated sequence acquires a mutation that prevents its expression. For example, a member of the α -globin gene family, $\psi\zeta$, has all the sequence characteristics of a functional globin gene, but the protein-coding region contains a point mutation that prevents the expression of a full-length globin [12]. A second way in which pseudogenes originate is via the pathway of reverse transcription and integration. If the mRNA of a cellular gene is converted into complementary DNA by reverse transcriptase, a duplex DNA molecule can be formed that lacks introns and contains a poly(A) tract. Pseudogenes with this pattern are commonly found in genomic DNA, showing that cellular mRNAs are occasional substrates for reverse transcriptase and that the DNA products can integrate back into the genome.

Large segmental duplications [13] are especially enriched in pericentromeric and subtelomeric regions of chromosomes. These intrachromosomal or interchromosomal duplications range between 1 and 300 kb, are >90% identical at the sequence level, and are much more common in humans than in yeast, flies, or worms, suggesting a relatively recent origin for these genomic elements. Segmental duplications have been demonstrated to serve as templates for the production of copy number variants (CNVs) within the genome through different mechanisms (see later).

Highly homologous sequences can reside within the CDS of the same gene (intragenic homology; same gene), have homology to functional genes (different genes), and have homology to nonfunctional pseudogenes or homology to sequence regions that are still poorly annotated.

Satellite DNA consists of arrays of simple tandem repeats; microsatellites are composed of repeats primarily of 4 bp or less dispersed throughout the genome [14]. These cover about 0.5% of the genome and exist in the form of dinucleotide repeat CA/TG. Microsatellites are well known for their causative roles in as many as 40 neurological diseases. Certain specific triplet repeats can be unstable, expanding and contracting during meiosis

and/or mitosis. If the repeat becomes excessively long, it can cause diseases such as Huntington disease or spinocerebellar ataxia, both of which are caused by the expanded repeat CAG in the coding region of a gene and result in a long polyglutamine tract within the gene product. Some expanded repeats occur in 5'- and 3'-UTRs and result in a disease due to an inhibitory effect on gene expression (Fig. 4.7).

Effects of microsatellites can occur at different levels of RNA, which includes alternative splicing, structural changes, and working as microRNAs (Fig. 4.7). Minisatellites are tandemly repeated sequences of DNA lengths from 1 to 15 kbp. The telomeric DNA sequences contain 10–15 kb of hexanucleotide repeats—TTAGGG. Macrosatellites are very long arrays of sequences up to hundreds of kilobases of tandemly repeated DNA. The α -satellite DNA constitutes the bulk of the centromeric heterochromatin on all chromosomes.

4.3.5 Gene Families

Genes belong to a family of closely related DNA sequences, which cluster together as families because of similarity in their nucleotide sequence or amino acid sequences. Gene families consist of structurally (and usually functionally) related genes with a common evolutionary origin. Multiple levels of hierarchical subfamily structure are common. A well-studied and extensively described example due to its clinical significance is the gene family consisting of genes that code for α - and β -globin gene clusters, which are assumed to have arisen from a gene duplication event 500 million years ago. This gene family is an example of duplication events due to retrotransposition. These two gene clusters also code for globin chains that are expressed during stages of human developmental stages. Another example of gene family with extensive diversity is the immunoglobulin superfamily.

Some gene family members, such as collagens, are dispersed among different chromosomal locations, but many others, such as the κ or λ variable region genes, are physically linked. Among gene families that exhibit linkage, members are usually oriented in the same direction. In some cases, linkage is thought to be an evolutionary footprint without functional significance, suggesting that evolutionary divergence of the gene family occurred through successive rounds of duplication in tandem arrays. However, in other gene families, such as the immunoglobulins, linkage has been conserved

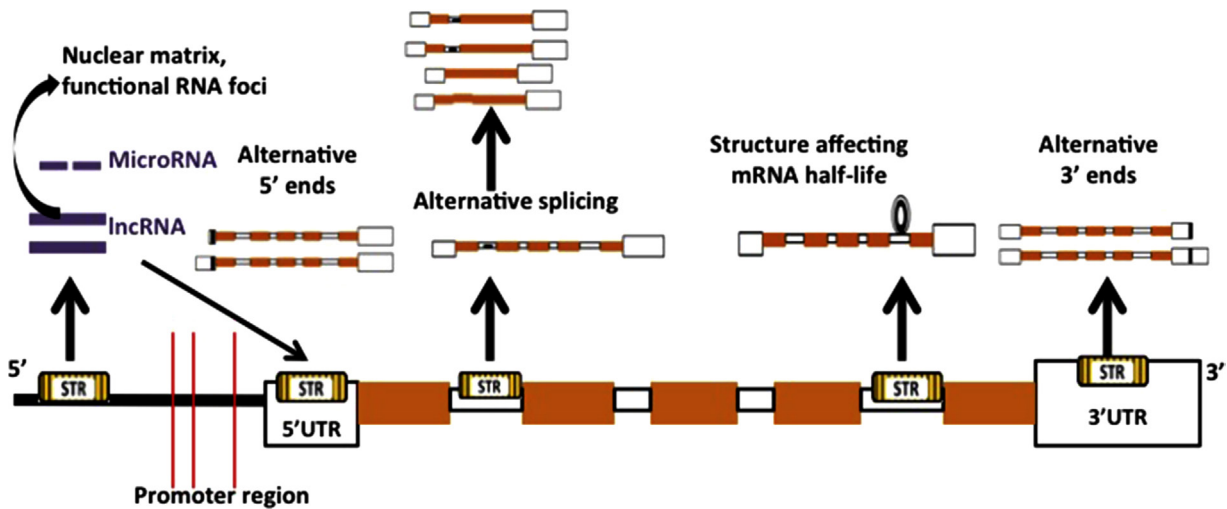


Figure 4.7 Effects of microsatellites at the level of RNA. Long noncoding RNAs (lncRNAs) predominantly consisting of microsatellites have been observed to function in the nuclear matrix and to aggregate into nuclear foci with indications of functional significance. They also may associate with DNA microsatellites (short tandem repeat [STR]) in both UTRs. Microsatellite-dominated microRNAs have been observed, but their function is not yet known. Intronic microsatellites can regulate splicing efficiency that can lead to exon skipping, intron inclusion or new splice site selection. STRs located in the UTRs can influence the locations of the start and end sites of transcription. Microsatellites transcribed can also affect the mRNA half-life, which may be due to formation of secondary structures such as hairpins.

during evolution because it provides a mechanism for coordinated or regulated control of gene expression. Even when dispersed, coordinated expression of members of gene families can be regulated by similar control mechanisms by carrying similar response elements in their 5' regulatory regions.

4.3.6 Interindividual Variations in the Human Genome

The Human Haplotype Map (HapMap) project is a key component in understanding the genetic potential of the Human Genome Project. Sequencing of the human genome has exposed multiple interindividual variations. These variants include both simple-sequence repeat polymorphisms and single nucleotide polymorphisms (SNPs). Repeat polymorphisms can represent di- (mostly CA), tri-, or tetranucleotide repeats, and they form the basis of the genetic map of the human genome. These repeat markers are multiallelic; the alleles differ in the number of repeat units and thus can be used to identify the maternal and paternal alleles of individuals as well as to define recombinations between marker loci. The high degree of length polymorphism among simple-sequence repeats within the human population

is due to frequent slippage by DNA polymerase during replication. These repeats comprise about 3% of the human genome, and there is approximately one such repeat per 2 kb of genomic sequence. The large number and wide distribution of these repeats have facilitated mapping and identification of many genes associated with inherited human disorders solely based on their chromosomal position.

SNPs are also nonrandomly distributed in the human genome. Millions of SNPs have been identified in the genome sequence but only a small fraction is predicted to affect the protein sequence. This limits the extent to which such genetic variations contribute to the structural diversity of human polypeptides, but regulatory effects on gene expression may cause or result in susceptibility to a variety of human phenotypes.

Copy number variation (CNV) describes the variation identified within the population or associated with human diseases in genomic segments larger than the SNP and simple-sequence repeat polymorphisms but smaller than cytogenetically visible chromosomal abnormalities [15]. An appreciation of the number and diversity of such variants has primarily been a product of comparative genomic hybridization studies in

normal individuals (copy number polymorphisms) and in patients with a wide variety of genetic disorders. Although the number of sites that vary is small when compared with SNPs and simple-sequence repeat polymorphisms, the number of base pairs involved may be as much as two orders of magnitude greater [16]. Similar to the SNP variations, CNVs may be associated with susceptibility to particular disorders or may be causative, especially when they arise *de novo* in an individual. There are three prevailing mechanisms for the creation of large CNVs—nonhomologous end joining (NHEJ), nonallelic homologous recombination (NAHR), and fork stalling template switching (FoSTeS)—and, more recently, aberrant firing of replication origins [17] (Maya-Mendoza, *Nature*, June 27, 2018). NHEJ is the cellular process to repair double-strand breaks (DSBs) and proceeds in four basic steps; the DSB is recognized, bridging of both broken DNA ends, preparation of the ends for ligation, and finally, ligation of the two DNA strands. NHEJ can result in small alterations within the genome such as microdeletions at the repair site, which may or may not have an effect on the organism. Moreover, NHEJ can lead to larger rearrangements such as chromosomal translocations, which have been associated with different types of cancer. NHEJ events are nonrecurrent types of CNVs and can, but not always, be found within low CNRs (SDs) (LCRs/SDs). NAHR events, on the other hand, are associated with LCRs/SDs and can be recurrent due to the presence of similar alterations in different individuals between the same LCR/SD. NAHR can occur during meiosis or mitosis and will result in either constitutive deletions/duplications that are associated with genetic abnormalities or sporadic deletion/duplication events that are mosaic or observed in cancers, respectively. Normal homologous recombination occurs between sister chromatids during meiosis to exchange genetic information or between sister chromosomes during mitosis. If the pairing of the different chromosomes occurs through allelic segments of the genome the resulting recombination has no phenotypic effect. If, however, the pairing occurs between two nonallelic segments of the genome, the result can be duplications or deletions if the LCRs/SDs are in a direct orientation or an inversion if the LCRs/SDs are in opposite orientations. Translocations can also occur if the recombination occurs between LCRs/SDs on different chromosomes.

Finally, the last major method involved in the production of CNVs is FoSTeS. The formation of CNVs through this mechanism starts during DNA replication with a stalled replication fork whereby the 3' end of the lagging strand can dissociate from its current strand and anneal, through microhomology, in a nearby replication fork. The nearby replication fork will be close in spatial proximity, not necessarily close in the linear sequence context. DNA synthesis can commence through the newly “primed” site. The newly synthesized DNA can dissociate from the replication fork and reintegrate back into the original fork or could invade and anneal to another replication fork. The process of disengagement and invasion could occur several times, resulting in very complex rearrangements [18].

4.3.7 DNA Looping and TADs

It has become increasingly clear in the past several years that the DNA in the nucleus has a higher order structure in the form of loops. As stated earlier, enhancer elements that lie either upstream or downstream of a gene's promoter can influence developmental timing of gene expression or tissue or cell type-specific gene expression. Enhancer sequences work in a position- and orientation-independent manner to affect gene transcription. Some early studies had suggested that the enhancer and the promoter form a loop in the DNA, bringing the proteins that bind both elements in close proximity to regulate transcription [19]. The locus control region (LCR) of the globin locus is a notable example where the LCR resides 50 kb upstream of the globin genes (α , β , and γ) themselves and regulates the switch between fetal hemoglobin and adult hemoglobin. More recently with the advent of next-generation sequencing (see later), it has become possible to assess contacts between regions of the genome in *cis* and in *trans* on a genome-wide scale. Indeed, through the use of a proximity-based ligation assay termed HiC (a process whereby DNA and proteins are crosslinked), the DNA is digested with a restriction enzyme, and the DNA within the complex is ligated, bringing together normally distant sequences. The ligation products are eventually sequenced with high throughput DNA sequencing; it has been shown that regions of chromatin that are in close proximity to each other have a higher probability to interact with each other, whereas more distant regions of chromatin have a lower probability of interaction. However, there are many regions of the genome that are separated by

great distances on the linear chromatin fiber that preferentially interact with each other. These segments form large DNA loops. The prevailing hypothesis of how the loops form is through the loop extrusion model whereby the ring protein cohesin binds to and draws the DNA through itself until the complex encounters the protein CTCF, which is also bound to the DNA. These proteins demarcate most, although not all DNA loops. They also act as boundary elements restricting access to the loop. The loops have been found to persist between cell types and throughout vertebrate evolution and have been termed topologically associated domains (TADs) [20,21]. One example of how TADs are important for regulating gene transcription was the discovery that individuals with copy number alterations at 2q35-36, which caused malformation of the hands and feet, were associated with the deletion of boundary elements and not with mutation of individual genes [22]. The location of 2q35-36 contains the genes *WNT6*, *IHH* (*Indian hedgehog*), *EPHA4*, and *PAX3* separated into TADs with boundaries between the *WNT6/IHH* and *EPHA4* and between the *EPHA4* and *PAX3* TADs. In other words, a *WNT6/IHH* TAD, an *EPHA4* TAD, and a *PAX3* TAD, each TAD separated by binding sites for the boundary proteins CTCF and cohesin. Importantly, the chromosome structure is conserved between mouse and humans at this locus. Modeling the human CNVs in the mouse resulted in similar phenotypes in the paws of the affected mice. Deletion of the boundary element, and hence the binding sites for the CTCF/cohesin complex, was responsible for the observed phenotypes in both humans and mice [22]. These results provide evidence that alterations in chromatin architecture can lead to phenotypic changes in the organism without mutations to the genes that reside within the loops.

4.3.8 Analysis of Genomes

During the past 10 years there has been an explosion of genomic data due primarily to the use of next-generation sequencing (NextGen or NGS). NGS, at its most basic level, is the process of sequence analysis of DNA (or cDNA) in a massively parallel configuration [23–25]. There are several competing concepts for NGS, but the most prevalent technology in use today is the Illumina Sequencing by Synthesis method. All NGS assays begin with the creation of a sequencing library. The process is basically the same for genomic sequencing and begins with shearing the DNA in some fashion, either by

mechanical means (sonication) or enzymatically (tagmentation with a DNA transposon). Following fragmentation, the ends of the DNA are repaired to blunt ends and phosphorylated, and an adenosine residue attached to provide a more efficient substrate for ligation of platform-specific adaptors. (These steps are not necessary during tagmentation as oligomers are inserted into the DNA by the transposase.) Following adaptor ligation, the library is amplified by PCR and is ready for sequencing [23–25].

Sequencing by synthesis on the Illumina platform begins by creating individual sequencing reactions, known as clusters, on the surface of a flow cell. The prepared DNA libraries are denatured and flowed across the surface of the flow cell in order for the adaptor sequences to bind their complementary sequences that are physically attached to the flow cell surface. Once attached the molecules are amplified in situ by bridge amplification, and the resulting clusters are ready for sequencing. In the Illumina system, nucleotides, with a specific fluorescence for each base and a 3' block, are added to the flow cell with the polymerase and one base is incorporated into the growing strand. The base is “read” by its unique color tag and the nucleotide is recorded. The 3' block is removed from nucleotide and the cycle begins again. This is repeated from 50 cycles up to 600 cycles in either one direction (single end sequencing) or in both directions (paired end sequencing) depending on the instrument and length of chemistry. Once the sequencing is finished, the raw data file is converted to a usable format, the FASTQ file. The FASTQ file format is similar to a FASTA except it has the addition of run quality metrics for each base position within the read. The FASTQ files are used in downstream bioinformatics analysis of the sequencing run.

4.4 STRUCTURE OF GENES (TRANSCRIPTIONAL UNITS): EXONS AND mRNA

Sequences coding for a single eukaryotic mRNA molecule are typically separated by noncoding sequences into noncontiguous segments along the chromosomal DNA strand (Fig. 4.6). The segments that are retained in the mature mRNA are referred to as exons. During transcription, the exons are spliced together from a larger precursor RNA that contains, in addition to the exons, interspersed noncoding segments referred to as introns.

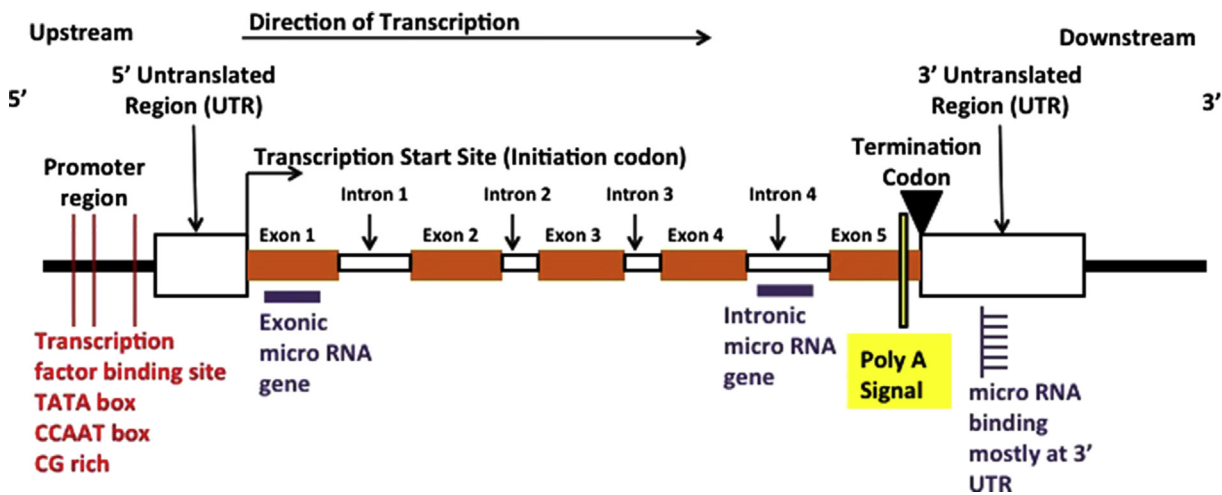


Figure 4.8 Organization of a human gene. At the 5' end (upstream) of each gene lies a promoter region that includes sequences (red lines) responsible for the proper initiation of transcription (TATA, CCAAT), including regulatory elements. At the 5' and 3' end of the gene is the untranslated region (5'-UTR and 3'-UTR; white boxes). The 3'-UTR contains a signal for the addition of polyA tail (yellow) to the end of the mature mRNA. Gene expression can be regulated by binding of specific microRNAs (purple) in microRNA recognition sequences mostly present in the 3'-UTR. Genes (open reading frames) expressing untranslated microRNAs can also be embedded within exons (exonic microRNA; purple) or introns (intronic microRNA) such as that in immunoglobulin lambda variable region gene family (Das S. Mol Biol Evol 2009;26(5):1179–89.). The nucleotide sequences adjacent after the 5'-UTR or before the 3'-UTR provide the molecular “start” (Initiation codon) and “stop” (Termination codon, black arrow) signals respectively for the mRNA synthesis from the gene. Exons (light brown) and intervening introns (white) are the coding and noncoding parts of the gene.

The number of exons coding for a single mRNA molecule depends on the gene and the organism, but ranges from one to more than 100 (Fig. 4.8). The noncoding mRNA sequences are spliced out during mRNA maturation. Human genes tend to have small exons, with a median value of only 167bp and mean equal to 216bp. The shortest exon is only 12bp while the longest is 6609bp. The exons are separated by introns, which can be less than 100bp, but can also exceed 10kb. The size distribution of exons and introns of human genes based on the analyzed sequence information and comparison to worm and fly sequences are provided in Table 4.2. It is important to note that some introns carry significant information and even code for other complete genes (nested genes).

Individual exons may correspond to structural and/or functional domains of the proteins for which they code, such as the signal peptide of secreted polypeptides or the heme-binding domain of globin. For some complex proteins, domains encoded by single exons often appear in apparently unrelated proteins, suggesting that the evolution of these proteins may have

TABLE 4.2 Physical Sizes of Human Chromosomes	
	Size
Median exon	167bp
Longest exon	6609bp
Smallest exon	12bp
Exon number	9
Introns	3300bp
3'-UTR	770bp
5'-UTR	300bp
Average coding sequence	1340bp
Polypeptide	447aa
Overall size	27kb

aa, amino acids; bp, base pairs; kb, kilobase pairs; UTR, untranslated region.

been facilitated by the ability to bring together different protein subdomains by exon shuffling. The origin of intron/exon structure is thought to be extremely ancient and to predate the divergence of eukaryotes and prokaryotes. However, prokaryotes and small eukaryotes

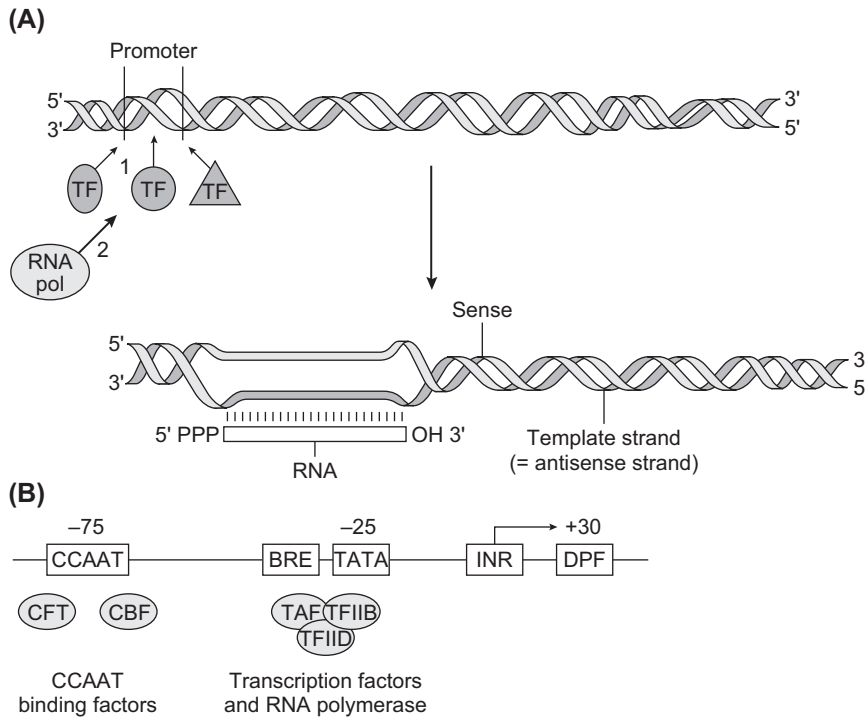


Figure 4.9 (A) Transcription of eukaryotic genes. (B) Basic promoter elements in eukaryotes.

(e.g., yeast) have lost their introns during evolution, perhaps because of the strong selective pressure on these organisms to retain a small genome size. Therefore, exons can be classified as follows, 5'-UTR exons, coding exons, 3'-UTR exons, and all possible combinations of those three main components, including single exons that cover the whole mRNA.

4.4.1 Gene Expression

The expression of individual genes can be regulated at multiple levels. Before a gene sequence gets translated into a polypeptide sequence, multiple events take place: activation of the local DNA structure, initiation and completion of transcription, processing of the primary transcript, transport of the mature transcript to the cytoplasm, and translation of the mRNA. All these steps can be the target of regulation and thus are potential control points for altering gene expression. Some genes are needed in all cell and tissue types as they encode a crucial gene product. Such genes are often referred to as “housekeeping genes.” However, numerous human and mammalian genes show highly restricted cell- or tissue-specific expression

patterns, and this spatial and/or temporal restriction of gene expression can also be regulated at multiple levels.

4.4.2 Transcription

Initiation of transcription happens when the compact DNA structure is loosened and short sequence elements in the 5' end of the gene guide and activate RNA polymerase (Fig. 4.9). A group of such sequences is often clustered upstream of the transcription initiation site to form the promoter. The promoter is a region of DNA at the 5' end of the genes that bind RNA polymerase.

There are different types of promoters for RNA polymerases I, II, and III. RNA polymerases I and III are dedicated to transcribing genes encoding RNA molecules (rRNA and tRNA), which assist in the translation of the polypeptide-coding genes. All RNA polymerases are large proteins and appear as aggregates consisting of 8–14 subunits. Significant amounts of information exist on promoter sequences specific for these polymerases. The basal apparatus, the generic minimal promoter sequence that is sufficient to initiate transcription of any protein-coding gene, contains an RNA polymerase II recognition signal

as well as signals for general TFs needed for the binding of the polymerase by most genes. This minimal promoter contains a consensus sequence (5'-TATA-3', referred to as the TATA box) about 25bp upstream of the site at which transcription begins, surrounded by GC-rich sequences, as well as the B recognition elements (BRE) sequence (TF recognition element), the Inr (initiator) sequence at the start site of transcription, and the DPE (downstream promoter element) at about 30bp 3' from the transcription initiation site. Furthermore, about 50–200bp upstream is the CAAT box, to which several TFs bind. The usual nomenclature of the numerous transcription factors is TF followed by a roman numeral to indicate the associated RNA polymerase. The general TFs, such as TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, facilitate the binding and activation of RNA polymerase II into an activated transcriptional complex.

Genes are constitutively expressed at some basal minimum rate determined by the core promoter. However, transcription can be increased or totally switched off by additional positive or negative elements (enhancers or silencers), which regulate the efficiency and specificity with which a promoter is recognized by the transcriptional apparatus. These *cis*-acting regulatory elements are typically short sequences located at about 200bp upstream from the promoter sequence but may also be placed at more distant locations. Finally, gene expression can be regulated by elements that respond to external stimuli. These response elements are often within 1000bp upstream from the transcription start site (Fig. 4.8). Genes under common control share similar response elements, recognized by regulatory TFs. Some response elements are extremely well characterized, such as heat shock response elements or glucocorticoid response elements.

Promoters do not necessarily have to lie upstream of the transcription initiation site. For example, most promoters for RNA polymerase III, including the 5S RNA promoter, lie downstream of the transcription start site within the coding sequence [26]. This promoter binds the general transcription factor TFIID, a large protein with several zinc fingers, which then, along with the factors TFIIB and TFIIC, binds RNA polymerase III in a manner such that the polymerase is positioned at the exact spot where transcription begins. Although the mechanism of TFIID binding appears to be a common one, promoters that lie in exon sequences may be limited to the special situations in which multicopy genes such as 5S RNA or tRNA are subject to coordinate regulation.

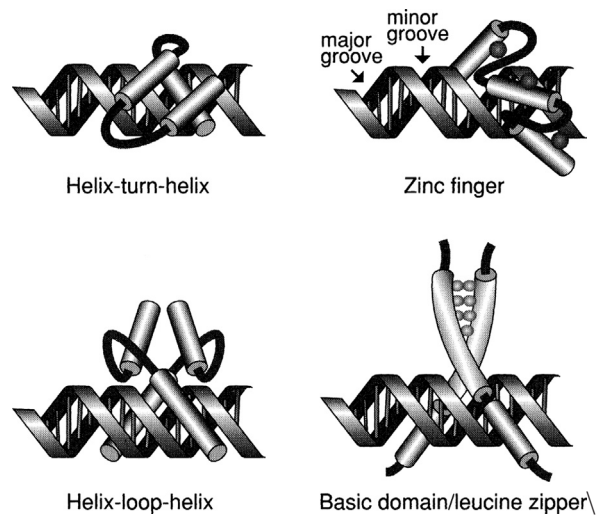


Figure 4.10 Three-dimensional structure of DNA helix bound to TFs.

Comparisons among many TFs have exposed some structural domains characteristic of the DNA-binding character of these proteins (Fig. 4.10). These include zinc finger motifs, helix-turn-helix motifs, helix-loop-helix motifs, and leucine zipper motifs. The zinc finger motif binds a zinc ion with four highly conserved amino acids, two cysteines and two histidines (C_2H_2), to form a finger-like loop [27]. The typical loop is 25 amino acids long, and the finger structure is often tandemly repeated. The helix-turn-helix motif is a common element of homeobox proteins. It consists of two short α -helices separated by a short linker region and confers sequence specificity to DNA binding [28]. The helix-loop-helix motif consists of two α -helices separated by a loop that is flexible enough to allow two helices to pack against each other. The contact of the helix-loop-helix motif with DNA is considered to be looser than other TFs. The leucine zipper motif is a helical stretch of amino acids, with leucine at every seventh amino acid position and occurring once in every two turns of the helix. Characteristic of most TFs is that they recognize and bind a short nucleotide sequence and their binding surfaces have extensive complementarity to the surface of the DNA double helix. Typically, eukaryotic TFs have two functional domains: a DNA-binding domain that binds to the DNA of the target gene and an activation domain that interacts with other proteins, which regulate transcription.

4.4.3 Enhancers and *cis*-Acting Regulatory Elements

Enhancers are defined as *cis*-acting sequences that increase transcriptional initiation but, unlike promoters, are not dependent on their orientation or their distance from the transcriptional start site [29]. They may be found within the introns of the genes they regulate, within adjacent genes or, in extreme cases, 1 Mb or more away. Enhancer sequences are generally short, on the order of 20–30 bp, and bind specific TFs. When there is a mechanistic diversity in enhancer function, many enhancers facilitate the assembly of an activated transcriptional complex at the promoter via a chromatin looping mechanism. Other mechanisms involve recruitment of RNA polymerase II by enhancers or transcription of enhancer sequences to generate long non-coding RNAs that can facilitate transcription. Enhancers have roles in differentiation, tissue specification, and tissue-selective gene expression, playing important roles during development and in specific cell types.

Silencers are another class of *cis*-acting regulatory elements that reduce transcription levels [30]. They are less well characterized than enhancers, and some of them are position dependent while others seem to be position independent. They can bind TFs that act in transcriptional initiation, and many genes contain a combination of both positive and negative upstream regulatory elements that act in concert on a single promoter. This diversity of regulatory elements has the potential to precisely modulate gene expression with regard to cell type, developmental stage, and environmental conditions. Boundary elements are insulators, most of which block or isolate the effects of enhancers or silencers, limiting their action to the target genes.

Variation of gene promoters or enhancers can alter the pattern of gene expression but not the structure of a particular gene product. While such variation is much less frequent than structural variation in genes, they provide insight into the elements of transcriptional regulation. For example, SNVs and partial deletions of the β -globin gene cluster that affect upstream regulatory sequences lead to reduced expression of adult β chains in β -thalassemia and/or increased expression of fetal γ chains in hereditary persistence of fetal hemoglobin [31].

4.4.4 5'-Untranslated Sequences

Shortly after initiation of mRNA transcription, a 7-methylguanosine residue is added to the 5' end of the primary transcript (see Fig. 4.6). This 5' cap is a characteristic of

nearly every mRNA molecule [32]. Many functions have been ascribed to the cap, the most notable of which is protection of the mRNA from degradation by exonucleases. The cap may also promote splicing and nuclear export of the RNA and is recognized by the translational machinery. The 5'-UTR extends from the capping site to the beginning of the protein-coding sequence and can be several hundred base pairs in length. The 5'-UTR regions of most mRNAs contain a consensus sequence, 5'-CCA/GCCAUGG-3', known as a Kozak consensus sequence, involved in the initiation of protein synthesis. In addition, about 5'-UTRs contain upstream AUG codons that can affect the initiation of protein synthesis and thus could serve to control expression of selected genes at the translational level.

4.4.5 Introns and Splice Junctions

The number of introns in a simple transcriptional unit will be one less than the number of exons. More complicated arrangements exist in which an upstream exon can be spliced to any of several different downstream exons, or in which a complete transcriptional unit is nested inside an intron of a second transcriptional unit. In these situations, the same DNA sequence can be used as both exon and intron, depending on the transcriptional unit. Regardless of how the transcriptional unit is organized, the boundaries between potential exons and introns share common features that are important in the splicing process. Beginning from the upstream or 5' exon, these splice junctions have the sequence where the brackets define the exon–intron junctions, the underlined nucleotides represent the splice donor, branch point and splice acceptor sequences, respectively, the uppercase sequences are virtually invariant characteristics of every splice junction, and the lowercase sequences are other conserved bases within the consensus splice sites. These conserved intron sequences serve a critical role in the splicing process, and many inherited diseases are caused by mutations of the consensus splice junctions. Most splicing mutations alter one of the invariant GT or AG nucleotides of the splice donor or splice acceptor [33] and result in abnormal splicing and either loss of the gene product or synthesis of an abnormal polypeptide chain.

A transcribed precursor RNA molecule must have its introns spliced out and its ends modified before export to the cytoplasm as mature mRNA. The spliceosome, which is composed of small nuclear ribonucleoproteins

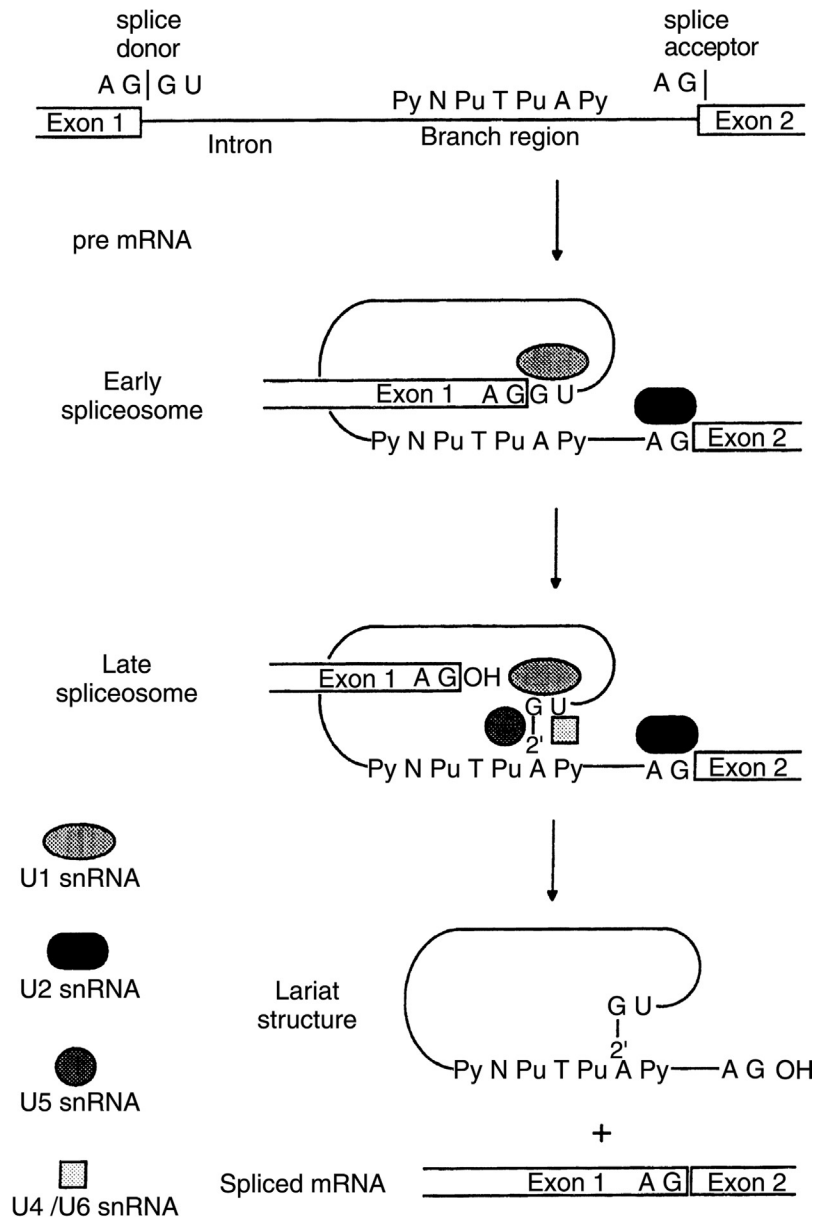


Figure 4.11 Splicing of mRNA.

(snRNPs), mediates the splicing of the large number of pre-mRNA transcripts, collectively referred to as heterogeneous nuclear RNA (hnRNA). Spliceosomes are multienzyme complexes that both catalyze the splicing reaction and stabilize the intermediates in the splicing process. The snRNPs composing the spliceosome consist of a set of five integral snRNA molecules (U1, U2, U4,

U5, and U6) tightly associated with a large number of proteins [34]. RNA molecules in the snRNPs are among the most highly evolutionarily conserved sequences among eukaryotes. An initial intermediate of the splicing reaction is formed when the 5' guanylate end of an intron (the splice donor) is joined to an adenylate residue near the 3' end of the intron (the branch point) through a 2'-5'

phosphodiester linkage (Fig. 4.11). After the completion of exon–exon fusion, the excised intron is released as a “lariat structure” by cleavage at the splice acceptor.

The genes encoding rRNA and tRNA also contain exons and introns but are spliced by different mechanisms than those required for mRNA splicing. Self-splicing of RNA without any protein factors is known to happen in prokaryotes, which suggests that introns have an extremely ancient evolutionary origin, predating not only the eukaryote/prokaryote divergence but also perhaps the origin of proteins as well.

4.4.6 3'-Untranslated Sequences and Transcriptional Termination

The 3' ends of primary transcripts are determined by transcriptional termination signals located downstream of the ends of each coding region. However, the 3' ends of mature mRNA molecules are created by cleavage of each primary precursor RNA and the addition of a several hundred nucleotide polyadenylate [poly(A)] tails (see Fig. 4.6). The cleavage site is marked by the sequence 5'-AAUAAA-3' located 15–20 nucleotides upstream of the poly(A) site and by additional GU-rich sequences 10–30 nucleotides downstream. Histone mRNAs, which do not have poly(A) tails, have stem-loop structures instead with cleavage of the primary transcript mediated by a distinct protein complex that includes the U7 snRNP [35].

Some complex transcriptional units contain several potential polyadenylation and/or transcription termination sites. It is often difficult to distinguish the latter from the former as the product available for analysis (mRNA) has lost the portion of the 3'-terminus originally transcribed by RNA polymerase. Alternative polyadenylation (or termination) sites can determine final protein structure if the longer precursor RNA contains an exon not found in the shorter precursor RNA. In a simple case, two proteins with different carboxyl termini are formed. But if alternative exon splice sites are made available in the longer precursor RNA, proteins with entirely different sequences can be produced.

The region from the translation termination codon to the poly(A) addition site may contain up to several hundred nucleotides of a 3'-UTR, which includes signals that affect mRNA processing and stability. Many mRNAs that are known to have a very short half-life contain AU-rich elements, 50- to 150-nucleotide sequences containing AUUUA motifs that regulate mRNA stability [36]. Other, less well-characterized sequences can have

similar effects. Removal or alteration of these sequences can prolong the half-life of mRNA, indicating that such elements represent a general regulatory feature of mRNAs whose level of expression can be rapidly altered.

4.5 TRANSLATION OF RNA INTO PROTEIN

4.5.1 Genetic Code

After intron sequences are spliced out of the primary RNA transcript and the 3'-terminus is generated [in most cases, by the addition of a poly(A) tail], the mature mRNA is transported from the nucleus to the cytoplasm, where it is translated into a polypeptide chain. In the cytoplasm, tRNA molecules provide a bridge between mRNA and free amino acids (see Fig. 4.3). Adjacent groups of three nucleotide sequences in the mRNA (codons) each bind to complementary three nucleotide sequences in tRNA (anticodons). Unlike most other nucleic acids, tRNA molecules have rigid tertiary structures. All tRNAs are L-shaped, with the anticodon located at one end and the amino acid binding site at the other end. Modified nucleotides, such as methylguanosine (mG) and pseudouridine (ψ), are common in tRNA and help determine the specific three-dimensional characteristics of tRNA molecules. Aminoacyl tRNA synthetases specifically recognize different tRNAs and attach each tRNA to the correct amino acid. The last base in each codon is followed by the first base in the next, and thus the first codon in an mRNA molecule determines the reading frame for all subsequent codons.

The relationship between codon and amino acid sequence is referred to as the genetic code (Fig. 4.12). Different tertiary structures of each tRNA are specifically recognized by the proper tRNA synthetase, ensuring the accuracy of the code. As the anticodon sequence itself does not determine tRNA tertiary structure, each amino acid may have several possible codons recognized by tRNAs with different anticodons but similar tertiary structures; that is, they are recognized by the same tRNA synthetase. For example, 5'-AAA-3' tRNA^{Phe} (the tRNA coding for phenylalanine with the anticodon 5'-AAA-3') has the same tertiary structure and is charged by the same tRNA synthetase as 5'-GAA-3' tRNA^{Phe}. Thus, both codons 5'-UUU-3' and 5'-UUC-3' code for phenylalanine using different tRNAs but the same tRNA synthetase. Additional redundancy in the genetic code arises because the third base in each codon–anticodon duplex (which is the first base from the 5' end of the anticodon)

UUU } Phe UUC } UUA } Leu UUG }	UCU } Ser UCC } UCA } UCG }	UAU } Tyr UAC } UAA - Stop UAG - Stop	UGU } Cys UGC } UGA - Stop UGG - Trp
CUU } Leu CUC } CUA } CUG }	CCU } Pro CCC } CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } Arg CGC } CGA } CGG }
AUU } Ile AUC } AUA } AUG - Met	ACU } Thr ACC } ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } Arg AGA } AGG }
GUU } Val GUC } GUA } GUG }	GCU } Ala GCC } GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } Gly GGC } GGA } GGG }

Figure 4.12 The genetic code.

can be flexible according to the rules of Watson–Crick base pairing. In particular, G:U or U:G base pairs are often found in the third position of a codon–anticodon duplex, and the guanine analog inosine, found only in tRNA, can pair or wobble with A, C, or U in the codon. Despite the redundancy of the genetic code, synonymous codons are not used with equal frequency, and the pattern of codon usage (codon bias) may vary tremendously among different species and between nuclear and mitochondrial mRNAs.

The AUG codon, which codes for methionine, nearly always begins the protein-coding portion of each mRNA molecule. Therefore, the vast majority of newly synthesized peptides begin with methionine. The tRNA^{Met} for the initiator AUG codon has a different tertiary structure from all other tRNAs, including the tRNA^{Met} that functions in elongation. Translation of most mRNAs generally begins with the first AUG from the 5' end, which is typically embedded within a Kozak consensus sequence (5'-CCA/GCCAUGG-3') and establishes the reading frame [37]. The UAA, UAG, and UGA codons are stop codons and have no cognate tRNAs. Thus, recognition of any one of these codons by the protein synthesis machinery terminates the protein-coding portion of every mRNA molecule.

Mutations that change a codon into a different codon and would therefore encode a different amino acid result in a protein with an amino acid substitution, and these are described as missense mutations. However, the UAA, UAG, and UGA codons do not code for an amino acid but instead serve as a signal to terminate protein synthesis. Mutations that produce one of these

codons in the middle of a normal reading frame cause truncation of the newly synthesized protein during protein synthesis and are referred to as nonsense mutations. Frequently transcripts with a premature termination codon are degraded by a process called nonsense-mediated decay, so that no protein product is synthesized from the mutant allele, resulting in haploinsufficiency for the gene product.

4.5.2 Protein Synthesis

The biochemistry of protein synthesis (Fig. 4.13) can be divided into the stages of initiation, elongation, and termination. All three processes occur on ribosomes, cytoplasmic particles of protein and rRNA that align the different substrates of each reaction. When inactive, ribosomes exist as separate pools of the two ribosome subunits, described by their size or sedimentation coefficient (S value). The small 40S ribosomal subunit contains 18S rRNA and approximately 33 different proteins, and the large 60S subunit contains 28S rRNA, 5.8S rRNA, 5S rRNA, and approximately 50 different proteins. Beyond these structural components, there is a wealth of additional factors, including both proteins and functional RNA molecules, that are required for ribosome biogenesis [38]. Translation begins with the formation of a preinitiation complex that contains the 40S ribosomal subunit, initiator tRNA^{Met}, GTP, and several protein initiation factors. An mRNA molecule initially binds to the preinitiation complex in conjunction with several initiation factors that interact with the 5' cap structure. The canonical model for identification of the AUG start codon involves scanning the mRNA in a 5'–3' direction until the consensus sequence 5'-CCA/GCCAUGG-3' is reached. However, internal ribosome entry sites (IRES) mediate ribosome recruitment and translational initiation for uncapped mRNA molecules and for translation when cap-dependent processes are inhibited [39]. Binding of the 60S ribosomal subunit and dissociation of several initiation factors generate a complex of proteins and subcellular particles poised to begin synthesis of the first peptide bond.

A ribosome contains room for two tRNAs and their respective amino acids (see Fig. 4.13). One tRNA at the peptidyl or P site is attached to the amino acid that has just been incorporated into a nascent peptide chain, and another tRNA at the aminoacyl or A site is attached to its cognate amino acid and is ready to participate in protein synthesis. During elongation, a peptide bond

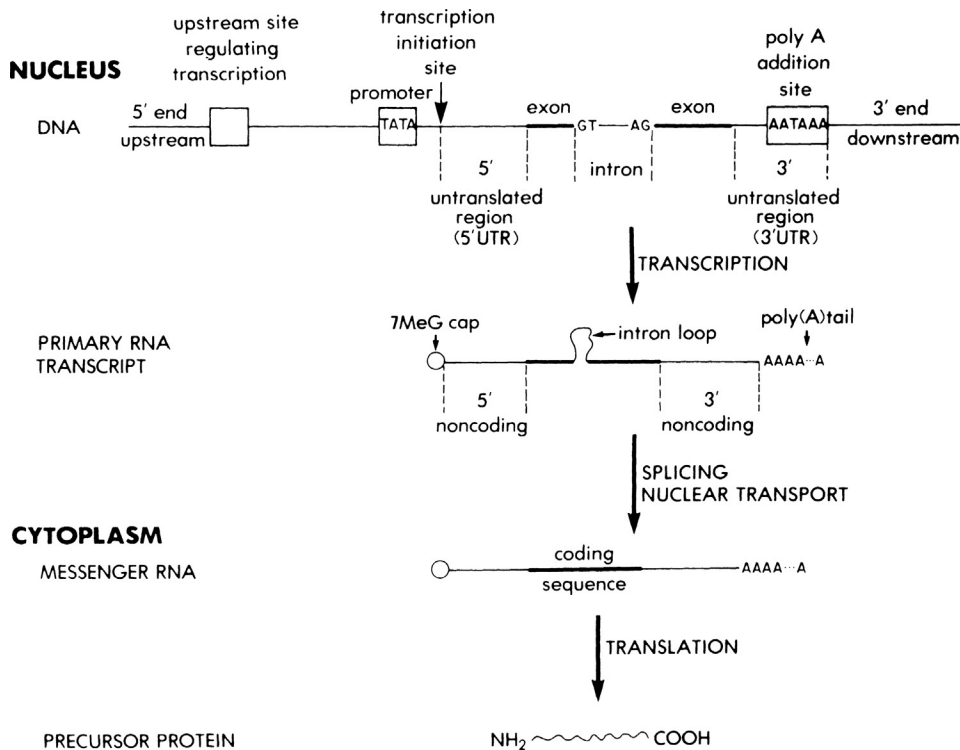


Figure 4.13 Translation of mRNA into protein.

is formed between the two adjacent amino acids, the ribosome moves to the next codon in the mRNA, the tRNA at the P site dissociates from the nascent peptide chain, and the tRNA at the A site is translocated to the P site. This series of reactions is dependent on elongation factor 1 (EF1), which binds to free charged tRNAs, and EF2, which facilitates translocation from the A site to the P site. A single mRNA can be simultaneously translated by several active ribosomes, forming a polysome that can contain as many as 50 ribosomes.

When a codon signifying termination of protein synthesis (UAA, UAG, or UGA) is reached, the completed polypeptide separates from the tRNA at the P site, and the ribosome dissociates. In bacteria and yeast, a unique group of suppressor tRNA mutations is caused by changes in an anticodon that then permit binding of a charged tRNA to a termination codon. Point mutations in an mRNA that would normally lead to premature termination codon (e.g., by changing a UUA codon into a UAA codon) are then partially suppressed by the mutant tRNA, allowing synthesis of a full-length protein with a missense change at the position of the mutant codon.

The principle of suppression of nonsense codons, which represent about 30% of disease-causing mutations in humans, can be mimicked by treatment with aminoglycoside antibiotics and other pharmaceutical compounds [40]. Used as a therapeutic approach, such treatment has the potential to affect therapy in a wide variety of genetic disorders.

4.5.3 Protein Localization

Gene products function in particular cellular compartments. For example, histones, tubulin, glycosyltransferases, peptide hormone receptors, and collagen are specifically localized to the nucleus, cytosol, Golgi apparatus, cell membrane, and extracellular space, respectively. Although many membranes contain pores large enough to accommodate a linear polypeptide chain, completely folded proteins are generally too large to fit through these pores. In addition, to the problem of translocating soluble proteins across membranes, proteins that remain attached to the membrane must be placed and oriented in specific ways. These problems—protein sorting, translocation, and membrane

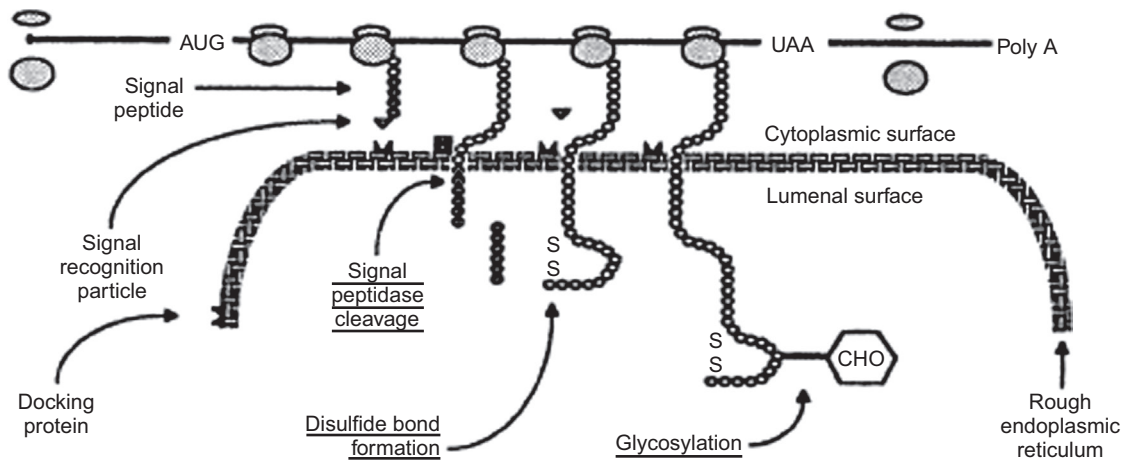


Figure 4.14 Translocation of newly synthesized proteins across the endoplasmic reticulum.

orientation—have been solved by complex biochemical mechanisms that depend in part on short peptide sequences in each protein. One of the most well-understood pathways is the initial sorting of gene products into those that will remain inside the cytosol or nucleus and those that pass across the endoplasmic reticulum (ER) membrane, and which are then available for secretion into the extracellular space. This initial sorting is determined early in the translation of proteins destined to cross the ER by the presence of a specialized hydrophobic signal sequence of 20–30 amino acids [41] usually located at the N-terminus (Fig. 4.14).

The signal sequence is first recognized when about 25 amino acids of the growing polypeptide have emerged from the ribosome and bind to a protein–RNA complex called the signal recognition particle (SRP). The SRP stops further translation until bound to a docking protein complex, the translocon, which is located on the surface of the ER and forms a hydrophilic membrane pore [42]. The signal peptide passes through the pore, translation recommences, and the growing polypeptide crosses the membrane cotranslationally. After the protein has passed into the lumen of the ER, signal peptidase, a protein on the luminal surface of the ER, cleaves the signal peptide to complete the initial phase of protein sorting.

Proteins extruded from the RER pass through the Golgi apparatus into secretory vesicles for transport to the cell surface. Proteins destined for the extracellular matrix are secreted from the cell when the vesicles fuse with the plasma membrane. Insertion and orientation

of proteins destined for the cell membrane, however, require additional sequences that function either to stop transfer across the membrane (stop transfer sequences) or to initiate transfer of an internal loop of the nascent polypeptide chain (start transfer sequences). Start transfer sequences are recognized by SRP-like N-terminal signal sequences but are not cleaved from the protein after translocation. The number, order, and orientation of start transfer and stop transfer sequences determines the conformation of complex integral membrane proteins that span the membrane multiple times. Some soluble proteins that contain the short peptide sequence Lys–Asp–Glu–Leu (KDEL) remain in the lumen of the RER, such as binding protein (BiP) and protein disulfide isomerase (PDI). Both BiP and PDI facilitate the folding of newly synthesized proteins in the RER. PDI catalyzes the rearrangement of Cys–Cys disulfide bonds; BiP is a so-called chaperone that binds temporarily to portions of other proteins normally not exposed to the surface and, in doing so, prevents partially folded proteins from misfolding and/or aggregating. Soluble proteins destined for specialized compartments inside the cell, such as lysosomes or peroxisomes, use a signal sequence to gain access to the ER lumen but require additional mechanisms for proper subcellular localization. Lysosomal sorting depends on amino acid sequences that specify posttranslational addition of a mannose 6-phosphate residue. Proteins containing this modification are selectively transferred from the Golgi apparatus to the lysosomal interior. Failure to modify proteins destined for the lysosome in

this way is responsible for the inherited disease mucopolipidosis II (I-cell disease).

There are more than 1000 mitochondrial proteins, the great majority of which are encoded in the nucleus and synthesized in the cytosol. Similar to the proteins destined for the ER, transport of proteins across the mitochondrial membrane also depends on a signal sequence and uses several translocator complexes [43,44]. The targeting sequence is usually located at the N-terminus and is cleaved on import, but sometimes is internal to the protein and is therefore not cleaved. Unlike RER proteins, for which translation and translocation are codependent, mitochondrial proteins are first translated completely, released into the cytosol, and then translocated into the mitochondrial membranes, intermembrane space or matrix. The potential problem of translocating a completely folded protein across a membrane pore is solved for mitochondria by complexes that include chaperone proteins of the Hsp70 family, which bind to proteins destined for the mitochondria and stabilize them in the unfolded state until after they have passed across the mitochondrial membrane, after which they assume their folded conformation.

4.5.4 Posttranslational Modification

Alterations to protein structure that occur after translation include the formation of disulfide bonds, hydroxylation, glycosylation, proteolytic cleavage, and phosphorylation. Phosphorylation of serine, tyrosine, and threonine residues is a common reversible modification that alters protein–protein interactions or controls enzymatic activity, mostly of intracellular proteins. The formation of disulfide bonds, hydroxylation, glycosylation, and proteolytic cleavage are generally not reversible and mostly involve extracellular proteins.

Intramolecular disulfide bond formation can begin cotranslationally as the growing polypeptide chain enters the lumen of the ER. Some proteins, such as immunoglobulin light chains, have a sequential pattern of intrachain disulfide bonds (e.g., between the first and second cysteines or third and fourth cysteines). Other proteins, such as proinsulin, have a more complicated pattern. Protein folding and establishment of the correct arrangement of disulfide bonds are critical steps in synthesizing a three-dimensional protein structure. Glycosylation of newly synthesized proteins may be O-linked

via serine, threonine, or hydroxylysine residues, or N-linked via asparagine residues. O-linked glycosylation is catalyzed by glycosyltransferases located on the luminal surface of the Golgi apparatus. N-linked glycosylation begins with transfer of a 14-residue oligosaccharide from a lipid molecule (dolichol) embedded in the RER membrane to the asparagine residue of a growing polypeptide chain. At some sites, the oligosaccharide is highly modified by removal of some carbohydrates and addition of other carbohydrates to form a complex glycoprotein modification. Other sites are less modified and contain the original high mannose composition of the dolichol intermediate. Many glycoprotein modifications help determine the specificity of extracellular protein–protein interactions, such as antigen–antibody binding or attachment of cells to the extracellular matrix.

Proteoglycans are a specialized class of extensively glycosylated proteins that contain a protein core with long disaccharide chains branching off at regular intervals and can contain as much as 95% carbohydrate by weight. Proteoglycans are extremely hydrophilic and form hydrated gels that provide structural integrity to the extracellular space. During growth and development, extracellular remodeling is accompanied by endocytosis and degradation of proteoglycans by lysosomal enzymes specific for different disaccharide chains; absence of these lysosomal enzymes produces mucopolysaccharidoses, such as Hunter syndrome or Hurler syndrome.

4.5.5 Expression of Housekeeping and Tissue-Specific Genes

Many proteins that operate in basic metabolic functions such as energy generation or nutrient transport are found in all cells, and the genes that encode these proteins are described as housekeeping genes. They are characteristically expressed at a relatively constant level in all cells. More specialized genes that are not housekeeping are used only at specific times and places during development or in one or a limited set of tissues. The sequence of the human genome has revealed that the most common genes in our genome are those encoding TFs and nucleic acid binding proteins, which together encode 26.8% of the proteins with known or putative function. Other highly represented genes encode receptors, transferases, signaling molecules, and transporters [45]. These and other housekeeping genes usually account for 90% or more of the transcripts expressed in any particular cell type.

Analysis of genome sequence data has revealed that segments of DNA with a relatively large proportion of 5'-CpG-3' dinucleotide pairs and a very high GC content are frequently located near the 5' ends of all housekeeping and a proportion of tissue-selective genes [46]. These CpG islands thus mark promoters and are present adjacent to about 70% of annotated genes, accounting for about half of the known CpG islands. Some of the remaining CpG islands have been associated with transcriptional activity, suggesting that these sequences mark the promoters of unknown transcripts and genes. On average, CpG islands are about 1000 bp long and most are less than 2000 bp in length. The CpG sequences within each island are generally nonmethylated and are thought to facilitate transcription either by preferential binding of TFs or by altering chromatin structure to a nucleosome-deficient and transcriptionally permissive state. Methylation of CpG islands is associated with reduced transcriptional activity, such as on the inactive X chromosome, although the methylation event appears to follow rather than initiate transcriptional silencing.

REFERENCES

- [1] Kauppi L, Jeffreys AJ, Keeney S. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet* 2004;5:413–24.
- [2] Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 1998;63:861–9.
- [3] Maeshima K, Imai R, Hikima T, Joti Y. Chromatin structure revealed by X-ray scattering analysis and computational modeling. *Methods* 2014;70:154–61.
- [4] Ou HD, Phan S, Deerinck TJ, Thor A, Ellisman MH, O'Shea CC. ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* 2017;357.
- [5] Cremer T, Cremer M. Chromosome territories. *Cold Spring Harb Perspect Biol* 2010;2:a003889.
- [6] Arthold S, Kurowski A, Wutz A. Mechanistic insights into chromosome-wide silencing in X inactivation. *Hum Genet* 2011;130:295–305.
- [7] Wang J, Jia ST, Jia S. New insights into the regulation of heterochromatin. *Trends Genet* 2016;32:284–94.
- [8] Mehta GD, Agarwal MP, Ghosh SK. Centromere identity: a challenge to be faced. *Mol Genet Genom* 2010;284:75–94.
- [9] Grandin N, Charbonneau M. Protection against chromosome degradation at the telomeres. *Biochimie* 2008;90:41–59.
- [10] Little PF. Structure and function of the human genome. *Genome Res* 2005;15:1759–66.
- [11] Makalowski W. The human genome structure and organization. *Acta Biochim Pol* 2001;48:587–98.
- [12] Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Sholders CC, Proudfoot NJ. The structure and evolution of the human beta-globin gene family. *Cell* 1980;21:653–68.
- [13] Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* 2007;39:S22–9.
- [14] Bagshaw ATM. Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol Evol* 2017;9:2428–43.
- [15] Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. *Annu Rev Genet* 2011;45:203–26.
- [16] Lupski JR. Genomic rearrangements and sporadic disease. *Nat Genet* 2007;39:S43–7.
- [17] Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet* 2009;10:551–64.
- [18] Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics* 2008;1:4.
- [19] Kim A, Dean A. Chromatin loop formation in the beta-globin locus and its role in globin gene transcription. *Mol Cell* 2012;34:1–5.
- [20] Gonzalez-Sandoval A, Gasser SM. On TADs and LADs: spatial control over gene expression. *Trends Genet* 2016;32:485–95.
- [21] Lupianez DG, Spielmann M, Mundlos S. Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet* 2016;32:225–37.
- [22] Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 2015;161:1012–25.
- [23] Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics* 2016;107:1–8.
- [24] Shendure JA, Porreca GJ, Church GM, Gardner AF, Hendrickson CL, Kieleczawa J, Slatko BE. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* 2011 (Chapter 7, Unit 7.1).

- [25] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* 2014;30:418–26.
- [26] Paule MR, White RJ. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res* 2000;28:1283–98.
- [27] Klug A. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Q Rev Biophys* 2010;43:1–21.
- [28] Scott MP, Tamkun JW, Hartzell 3rd GW. The structure and function of the homeodomain. *Biochim Biophys Acta* 1989;989:25–48.
- [29] Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* 2011;144:327–39.
- [30] Riethoven JJ. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol* 2010;674:33–42.
- [31] Thein SL, Menzel S, Lathrop M, Garner C. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Hum Mol Genet* 2009;18:R216–23.
- [32] Cowling VH. Regulation of mRNA cap methylation. *Biochem J* 2009;425:295–302.
- [33] Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat* 2007;28:150–8.
- [34] Valadkhan S, Jaladat Y. The spliceosomal proteome: at the heart of the largest cellular ribonucleoprotein machine. *Proteomics* 2010;10:4128–41.
- [35] Dominski Z, Marzluff WF. Formation of the 3' end of histone mRNA: getting closer to the end. *Gene* 2007;396:373–90.
- [36] Barreau C, Paillard L, Osborne HB. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res* 2005;33:7138–50.
- [37] Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 2005;361:13–37.
- [38] Strunk BS, Karbstein K. Powering through ribosome assembly. *RNA* 2009;15:2083–104.
- [39] Gilbert WV. Alternative ways to think about cellular internal ribosome entry. *J Biol Chem* 2010;285:29033–8.
- [40] Linde L, Kerem B. Introducing sense into nonsense in treatments of human genetic diseases. *Trends Genet* 2008;24:552–63.
- [41] High S. Protein translocation at the membrane of the endoplasmic reticulum. *Prog Biophys Mol Biol* 1995;63:233–50.
- [42] Swanton E, Bulleid NJ. Protein folding and translocation across the endoplasmic reticulum membrane. *Mol Membr Biol* 2003;20:99–104.
- [43] Endo T, Yamano K. Multiple pathways for mitochondrial protein traffic. *Biol Chem* 2009;390:723–30.
- [44] Mokranjac D, Neupert W. Thirty years of protein translocation into mitochondria: unexpectedly complex and still puzzling. *Biochim Biophys Acta* 2009;1793:33–41.
- [45] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 2003;13:2129–41.
- [46] Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;25:1010–22.

Epigenetics

*Rosanna Weksberg^{1,2}, Darci T. Butcher¹, Cheryl Cytrynbaum^{1,2},
Michelle T. Siu¹, Sanaa Choufani¹, Benjamin Tycko³*

¹Genetics and Genome Biology, Research Institute, The Hospital for Sick Children, Toronto, ON, Canada

²Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada

³Division of Genetics & Epigenetics, Hackensack Meridian Health Center for Discovery and Innovation, Nutley, NJ, United States

5.1 INTRODUCTION

Despite the tremendous advances in human genetics enabled by the original genome projects and brought to fruition with high-throughput genotyping and massively parallel DNA sequencing, many aspects of human biology still cannot be explained by genetics alone. Over the years a largely parallel line of research has implicated epigenetic dysregulation in human diseases. With advancing technologies, the two areas of genomics and epigenomics are starting to come together to yield fundamental insights. Epigenetics is defined as modifications of DNA and its associated proteins and ribonucleoproteins that do not involve alterations in the DNA sequence itself [1]. Normal human development requires the specification of a multitude of cell types and tissues that depend on transcriptional programs that are regulated by epigenetic mechanisms. These mechanisms enforce the acquisition and maintenance of the characteristic gene expression profiles of specific cell types in developing and adult tissues. Enzymes and other proteins that act as epigenetic regulators can be categorized based on their specific functions as “writers,” “erasers,” or “readers” of epigenetic modifications. Writers enzymatically deposit specific substrates (including acetyl, methyl, and phosphate groups) to modify DNA. Erasers catalyze the removal of these posttranslational modifications. Readers have specific domains that recognize individual epigenetic modifications, often in a unique

context, which can then recruit other epigenetic machinery to these sites and fine-tune transcriptional activity. Disruption of such control mechanisms is associated with a wide variety of human disorders with behavioral, endocrine, and/or neurologic manifestations, and quite strikingly with disorders of tissue growth, including cancer. While epigenetic factors affecting disease susceptibility and progression have been studied for decades, this topic has attracted much more attention in our current “postgenomic” era. Ongoing research is focused on understanding the functions of epigenetic regulators that modify or read epigenetic marks, characterizing *cis*- and *trans*-acting influences of the genetic background on epigenetic marks, delineating cell type/tissue-specific epigenetic marks in human health and disease, and studying the interactions between epigenetic marks and the environment, especially with respect to fetal programming and risks for common adult-onset disorders. Further, modulation of adverse epigenetic states by drug-based and nutritional therapies has become an important field of investigation. These efforts have been facilitated by large-scale “epigenome projects” that have defined epigenetic patterns across multiple human cell types and throughout different periods of development, producing whole-epigenome maps that are being integrated with parallel genomic data via publicly accessible platforms such as the Human Genome Browser at the University of California at Santa Cruz (UCSC) (see list of websites).

5.2 EPIGENETIC MECHANISMS: CHROMATIN, HISTONE MODIFICATIONS, DNA METHYLATION, AND LONG NONCODING RNAs

Site-specific epigenetic modifications, essential for controlling gene expression in normal growth and development, are established by multiple enzyme-catalyzed mechanisms that target DNA methylation at cytosine residues in CpG dinucleotides and covalent modifications of histone proteins, as well as by more recently appreciated mechanisms that involve untranslated RNAs and long-range chromatin architecture within the cell nucleus [2,3].

DNA in most eukaryotic cells is packaged with histone proteins to form nucleosomes, the beads in the well-known “beads on a string” structure of chromatin. In eukaryotic cells, most double-helical DNA is wrapped around an octamer core of four histone homodimers: H2A, H2B, H3, and H4. Through multiple levels of organization, these nucleosomes define chromatin conformations that can be relaxed or tightened to either facilitate or inhibit transcription in specific cells at critical times in development. Condensed states of chromatin (heterochromatin) inhibit transcription, while relaxed states (euchromatin) are permissive to transcription. For instance, nontranscribed telomeric and centromeric repeat regions are often silenced owing to their compact heterochromatin environment. Highly active genes, usually located within relaxed euchromatin, can be bound by transcription factors and expressed, often with a short nucleosome-free segment of DNA near the transcriptional start site. Further regulation is accomplished by assembling promoter–enhancer complexes via long-range chromatin looping, a process that is regulated by specific DNA sequences called insulators [3,4].

The core histones are subject to diverse posttranslational modifications that are closely associated with chromatin states. These include methylation and acetylation of lysine and arginine amino acid residues in the N-terminal histone tails that project from the nucleosome cores (Fig. 5.1). Depending on their pattern of modifications, these N-terminal histone tails are recognized by other chromatin proteins that signal reader proteins, which define chromatin states or bind regulatory proteins, activating or repressing

transcription (Fig. 5.2). The enzymes that catalyze histone modifications (epigenetic writers and erasers), and the chromatin-associated proteins that recognize the modifications (“epigenetic readers”) are therefore important regulators of gene expression. Mutations in some of these genes can lead to human diseases, which will be discussed in later sections of this chapter. Examples of epigenetic writers and erasers include histone acetyltransferases, histone deacetylases (HDACs), and histone methyltransferases. The modifications catalyzed by this large array of enzymes can be sequential and interdependent, mutually exclusive, or independent. Despite this complexity, data from large-scale projects such as ENCODE have revealed general rules that have substantiated the “histone code hypothesis,” in which particular combinations of histone modifications are reproducibly associated with specific types of regulatory elements such as gene promoters and enhancers, while other combinations of marks delineate broad regions of active, poised, or repressed chromatin [10]. The histone code can now be readily visualized for these specific marks or types of chromatin in various cell types as “tracks” in the Human Genome Browser at UCSC (<https://genome.ucsc.edu/>) and related websites such as Roadmap Epigenomics Consortium (<http://www.roadmapepigenomics.org/>) [11]. Ongoing efforts by the International Human Epigenetic Consortium (<http://ihc-epigenomes.org/>) and many individual laboratories are under way to fully catalog the different combinations of histone modification, DNA methylation, and gene expression that establish cell-type-specific chromatin states [12].

Methyl groups are added *de novo* to DNA at various times during development, for example, to regulate gene transcription both on the autosomes and in X inactivation, to establish parental imprints [13], to methylate centromeric DNA and other constitutive heterochromatin, and to defend the host against foreign DNA integration and expression [14]. DNA methylation involves the covalent transfer of a methyl group to a cytosine in DNA, most commonly occurring at the 5-carbon position on cytosine (5-methylcytosine or 5mC). In humans, DNA methylation occurs most often at dinucleotides of cytosine followed by guanine (CpG), which create a simple palindromic sequence, i.e., CpG is present on both strands 5′ to 3′. This process is catalyzed by DNA methyltransferase enzymes that establish and maintain these patterns through cell division. DNA methyltransferase

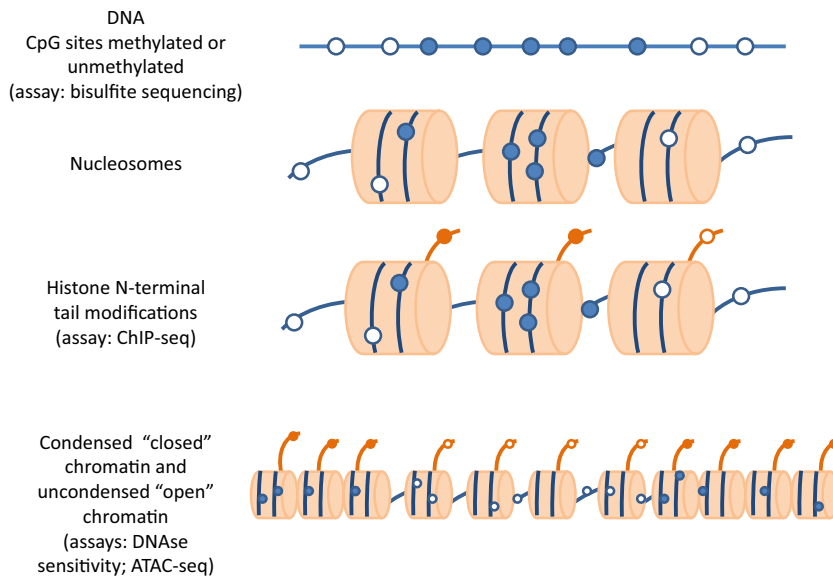


Figure 5.1 Epigenetic Organization of Chromatin: Layering of DNA Methylation and Histone Modification to Control Gene Expression. DNA of a gene promoter can be unmethylated (*white circles*), and in most cases the gene is expressed, or the promoter can be methylated (*blue circles*), and in most cases the gene is not expressed. DNA is not independent of its associated histone proteins. Histone modifications are established and maintained independent of or dependent on the DNA methylation state of the region. These protein modifications can activate (*open orange circles*) or repress (*filled orange circles*) gene transcription. Although not shown in this figure, but mentioned in the text, additional epigenetic processes, including microRNAs and long noncoding RNAs, also contribute to gene regulation. The DNA/histone protein nucleosome core is further compacted to form higher-order chromatin structures that also contribute to gene regulation. Methods for mapping the patterns of CpG methylation, histone modifications, and chromatin accessibility are indicated (ATAC-seq, assay for transposase-accessible chromatin).

1 (*DNMT1*), the major maintenance methyltransferase in mammalian tissues, has a high affinity for hemimethylated DNA (only one C in the duplex methylated) [15] and it therefore acts to propagate established methylation patterns. De novo methylation is carried out by two DNA methyltransferases, *DNMT3A* and *DNMT3B*, which can efficiently methylate CpG's that are not hemimethylated. Another member of the DNMT3 family, *DNMT3L*, has no catalytic activity but can bind to and activate *DNMT3A*; it is required to maintain allele-specific methylation in imprinted regions of the genome [16]. DNA demethylation occurs both passively, via replication of the cellular genome when DNMTs are not abundant, and by active processes that involve sequential modifications of the methylated base followed by thymine DNA glycosylase-dependent base excision repair [17,18].

In mammalian cells, high levels of DNA methylation (most CpG dinucleotides methylated) occur within repetitive elements and in some nonrepetitive sequences in intergenic and intragenic regions, while

CpG methylation is usually *excluded* from gene promoter regions that are especially CpG rich, referred to as CpG islands [19]. In 98% of the genome, CpG dinucleotides appear at a low frequency of about 1 per 80 nucleotides, but in the remaining 2% of the genome they are found at a much higher density, in CpG islands ranging in size from 200bp to several kilobases in length. Approximately 50%–60% of genes contain CpG islands, typically in their proximal promoters. Such CpG islands are almost always unmethylated in normal tissues, with the exception of imprinted genes, X-inactivated genes, retrotransposons, and a subset of genes with tissue-specific silencing [20,21]. In pathological situations such as cancers, DNA methylation can occur in CpG island-associated promoters and cause gene silencing, either by directly interacting with transcription factors or by recruiting methyl-binding proteins that then recruit histone-modifying enzymes to transform chromatin to a repressive state [22]; [23].

Although DNA methylation and histone modifications are regulated by different sets of enzymes, cross

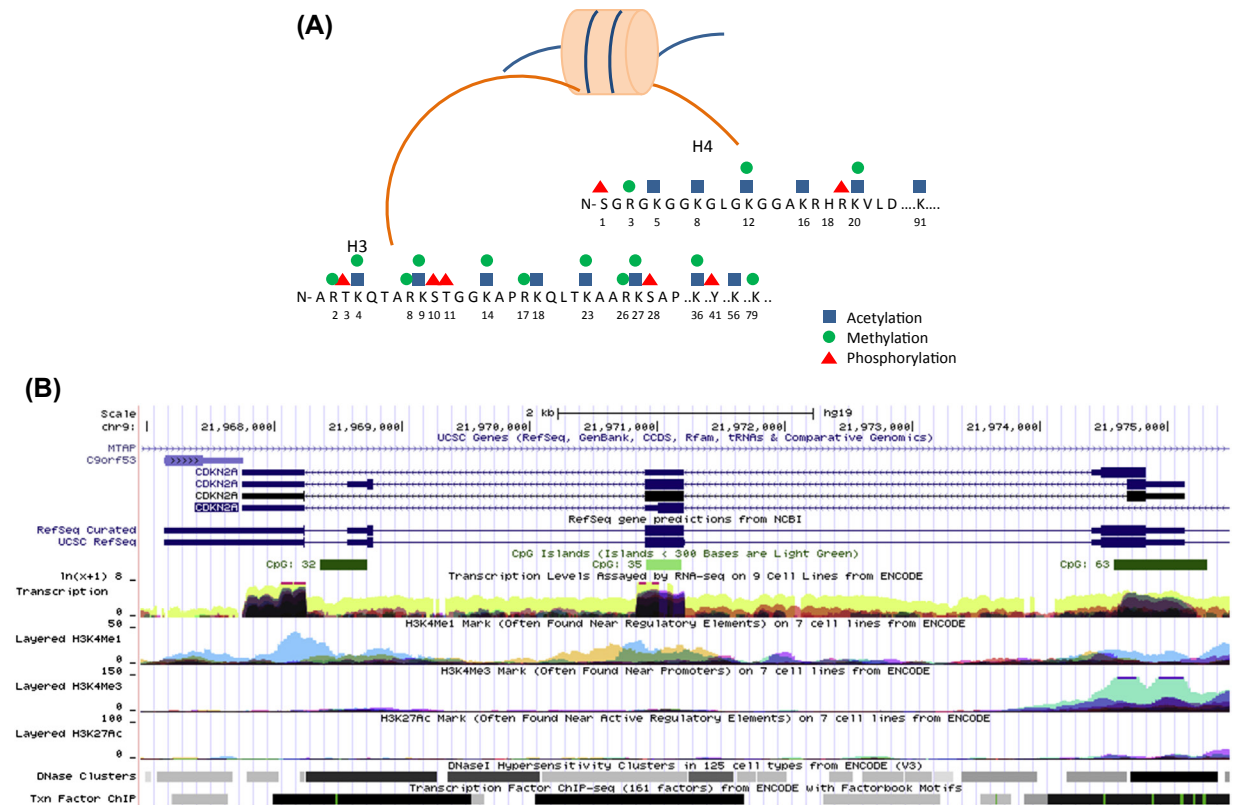


Figure 5.2 (A) Histone Modifications of Histone H3 and H4 N-Terminal Tails. Posttranslational modifications of N-terminal tails (these can also occur in the C-terminal domain but are not shown here) can occur in combination and are read by the appropriate protein to establish local and global open or closed chromatin states. (B) Snapshot From the University of California at Santa Cruz (UCSC) Genome Browser Demonstrating the Layers of Epigenetic Regulation in the Promoter of the Tumor-Suppressor Gene *CDKN2A*. This diagram is an example of epigenetic data available in UCSC genome browser. The description of each genomic feature is shown on the left. Two isoforms of the *CDKN2A* gene are shown in black and blue; here we focus on a shorter (black) isoform. Enrichment for the regulatory histone mark H3K4me1, the active histone H3K4me3, and the active histone H3K27Ac is shown by multiple colors for different cell lines. The peaks for each of these marks coincide with the transcription start site of *CDKN2A* overlapping a CpG island (green), as well as transcription factor binding sites (TxN factor ChIP) and DNase 1 clusters, which are indicators of open chromatin. All these marks—H3K4me3, transcription factor binding, and DNase1 clusters—indicate that *CDKN2A* is actively transcribed in these cell lines (shown in the transcription track). Information about other histone marks and DNA methylation levels is available from the UCSC genome browser under multiple tracks from the Regulation tab. This image was downloaded from the UCSC genome browser hg19: <http://genome.ucsc.edu/> [5,6]. The ENCODE regulation data are from Ref. [7–9].

talk between these modifications occurs through protein networks that write and read these patterns [24]. These two central types of epigenetic modifications are interdependent; histone marks are more labile and DNA methylation marks more stable [24–29], so DNA methylation can act to “lock in” epigenetic states. However, regulating metastable states of gene expression is so crucial in development and tissue homeostasis that other mechanisms, in addition to histone modifications and DNA methylation, come into play to establish and maintain epigenetic states.

Regulatory noncoding RNAs, including small interfering RNAs (siRNAs), microRNAs (miRNAs), and long noncoding RNAs (lncRNAs), play important roles in gene expression regulation at several levels, including transcription mRNA degradation, splicing, transport, and translation [30]. While in plants small RNAs are important for targeting specific loci for cytosine methylation [18], in mammals the main function of siRNAs and miRNAs is posttranscriptional regulation [31]. However, lncRNAs, defined as RNA species >200 nucleotides in length (and often much larger), are increasingly

recognized as playing a key role in modulating the epigenetic states of some highly regulated mammalian genes. lncRNAs are transcribed from various genomic locations in relation to their target protein-coding genes. They may be antisense, intronic, intergenic, or promoter or enhancer associated and can regulate transcription both *in cis* and *in trans* by a number of different mechanisms [2,32]. Specifically, certain lncRNAs have been shown to establish specialized nuclear compartments devoid of RNA polymerase II, in which the chromatin-associated polycomb repressive complex-2 (PRC2) catalyzes the formation of a repressive constellation of histone marks [2,30,32,33]. A particularly well-studied example is the *XIST* lncRNA that is essential for initiating the silencing *in cis* of one X-chromosome in female cells [2,34], discussed in greater detail in Section 5.4. lncRNAs are also found in imprinted gene clusters, where monoallelically expressed lncRNAs mediate the repression of one allele via the propagation of repressive chromatin marks such as DNA methylation and histone (H3K27 and H3K9) methylation on that allele *in cis* [35]. In contrast to gene silencing mediated *in cis* by *XIST* and imprinted lncRNAs, the nonimprinted lncRNA *HOTAIR* expressed from the *HOXC* gene recruits PRC2 and the histone demethylase LSD1, resulting in acquisition *in trans* of silencing histone marks at genes within the *HOXD* gene cluster located on another chromosome [36]. It has been estimated that 20% of lncRNAs expressed in human cells are bound by polycomb group proteins, suggesting a shared biochemical mechanism for their role in epigenetic silencing [37,38]. The roles of lncRNAs are diverse; some of them, exemplified by the *H19* untranslated RNA, which is abundant in fetal tissues, are host transcripts that give rise to specific miRNAs, while others appear to function as “molecular sponges” to sequester miRNAs in the cytoplasm [2].

In addition to the “fifth base,” 5mC, mammalian DNA contains a related modified base, 5-hydroxymethylcytosine (5hmC). This “sixth base” has taken center stage with the demonstration that proteins in the Ten–Eleven–Translocation (TET) family of oxygenases catalyze the conversion of 5mC to 5hmC, and that the gene for one of these enzymes, *TET2*, is sometimes mutated in human cancers, notably in cases of acute myeloid leukemia and myelodysplastic syndrome (MDS). The formation of 5hmC can, in principle, lead to demethylation of DNA, either by a passive process (failure of remethylation of these sites in S phase) or by an active mechanism

(direct demethylation of 5hmC or base excision of a further modified product such as carboxy-mC), so it has been suggested that TET family oxidases probably contribute to the dynamics of DNA methylation [18]. 5hmC has been found in many cell types and tissues, with particularly high levels in the brain and in embryonic stem (ES) cells. As discussed in the next section, on epigenetic reprogramming, TET1 has been shown to be important for self-renewal and maintenance of ES cells.

5.3 EPIGENETIC REPROGRAMMING

Global epigenetic reprogramming occurs first in cell populations destined to become germ cells. Epigenetic marks (including DNA methylation) are erased, following which the egg and sperm acquire their highly specialized and divergent epigenetic marks [39]. The second wave of epigenetic reprogramming occurs after fertilization [40]. A small number of genomic regions are protected from this zygotic wave of epigenetic reprogramming; these regions retain an epigenetic “memory” of their parent of origin and thus are crucially involved in the non-Mendelian phenomenon of genomic imprinting [41].

Genomic imprinting (also known as parental imprinting) has been well studied at the whole-genome level and for specific imprinted genes. In most areas of the diploid genome, epigenetic marks on the paternal and maternal alleles are equalized during preimplantation development, with strong allelic asymmetries persisting mainly at imprinted loci [42] and at certain other loci with haplotype-dependent allele-specific methylation. This equalization of the two alleles at most autosomal loci is essential for classical Mendelian transmission of human genetic disorders and involves early postzygotic reductions in DNA methylation. The reduction seems to result from both active demethylation, via mechanisms that are still being investigated experimentally, and passive demethylation, in which CpG methylation is diluted through early rounds of DNA replication in the presence of low-maintenance methylase activity.

Genomic levels of 5hmC change during development, probably as a function of the activity of TET family enzymes, which are particularly highly expressed in ES cells [43–45]. In the mouse genome, 5hmC is widely distributed throughout nonrepetitive regions, whereas satellite repeats (which are located in heterochromatin) are highly enriched for 5mC but substantially less so

for 5hmC [46]. Distinct 5hmC and 5mC patterns are observed at CpG islands overlapping gene promoters: while 5mC is depleted from these regions, 5hmC is well represented. Interestingly, the presence of 5hmC and depletion of 5mC at CpG island promoters is associated with increased transcription in ES cells [46]. Consistent with these observations, active histone marks H3K4me3 are enriched in promoters with high 5hmC, suggesting that enriched 5hmC at CpG island promoters is positively correlated with active transcription. Decline of TET oxidase levels during differentiation of ES cells is accompanied by reduced promoter 5hmC and increased 5mC levels, correlating with silencing of certain key developmental regulator genes [46,47]. Thus, one current hypothesis is that hydroxymethylation and the TET proteins could play a role in erasing methylation marks from promoters of pluripotency-related genes during differentiation [46,48]. Hydroxymethylation may also play a role in epigenetic reprogramming of primordial germ cells (paternal genome) and early embryos [49,50].

5.4 EPIGENETIC REGULATION OF X INACTIVATION

Inactivation of one of the two X chromosomes in female cells was first described by Mary Lyon in 1961 [51] and since then remains the prototypical example of chromosomal epigenetic silencing. The process of X inactivation is regulated in *cis* by a small control region known as the X-inactivation center (XIC). X-chromosome inactivation (XCI) has evolved in placental mammals to achieve dosage compensation of X-linked genes between females, who have two X chromosomes, and males, with only one X and one Y chromosome. It is hypothesized that the X and Y chromosomes evolved from a pair of autosomes, coinciding in time with the evolution of the placenta and driven by acquisition of the Sex-Determining Region Y (SRY) gene. In the course of mammalian evolution, the SRY-carrying Y chromosome has been significantly reduced in size and lost most of its active genes, retaining, in addition to SRY, a few other genes playing a role in male reproduction, whereas the X chromosome has acquired additional genetic material through translocations from autosomes. As a result, homology between the X and the Y chromosomes is now limited to two small pseudoautosomal regions (PAR1 and PAR2) (reviewed in

Ref. [52,53]). It is estimated that the human Y chromosome contains ~45 expressed genes, whereas the X chromosome has ~1300 [54]. The resulting major bias in copy number for X-linked genes between males and females is, for the most part, transcriptionally compensated via XCI.

The XIC is critical for the mechanism of XCI. X-chromosome deletion and translocation mapping in mouse and human defined a critical region for the XIC covering ~1 Mb and mapping to chromosome band Xq13 in humans [55,74]. A major breakthrough in understanding the mechanism of XCI followed the discovery of an lncRNA within this region, the X-inactive-specific transcript *XIST*, which is expressed from the XIC at high levels only on the X chromosome destined for inactivation, and which quickly spreads to coat the entire inactive X (Xi) in female cells [56–58]. The majority of the work on mechanisms of X inactivation was performed using mouse preimplantation embryos and ES cells. Before turning to the details of this mechanism, it should be noted that although mouse models have been very valuable in unraveling XCI, there are a number of substantive differences in the XCI process between mice and humans. In mice, there are two waves of XCI; the first is imprinted inactivation of the paternal X, which is initiated shortly after fertilization. This pattern of paternal X-chromosome silencing is maintained in extraembryonic tissues, but is erased and reestablished in a random manner in the inner cell mass (ICM) of blastocysts, which gives rise to the embryo proper [59,60].

In human embryos, XCI is initiated at the blastocyst stage and occurs at random in both the ICM and the trophoblast [61]. Random XCI is a multistep process, which can be divided into three steps: initiation, spreading, and maintenance. Initiation involves counting the number of X chromosomes per cell so that one X remains active per diploid number of autosomes. In other words, XY males and XO females keep their single X chromosome active, whereas XX, XXX, and XXXX females and XXY males inactivate all but one X, upregulating *XIST* RNA on all the X chromosomes destined to become inactive. The spreading of inactivation is achieved through sequential acquisition of epigenetic marks starting with *XIST* RNA, followed by PRC2 recruitment, a shift to late replication timing, enrichment of histone macro H2A, and silencing of chromatin marks, such as histone H3 and H4 hypoacetylation, H3 lysine 27 methylation, and finally

DNA methylation of CpG-rich promoters [62,63]. Once established early in embryonic development, the inactive state of an X chromosome is maintained through somatic cell divisions utilizing epigenetic modifications of DNA and histones. As DNA methylation is sufficient to lock in the inactive state, *XIST* expression is no longer required [64].

Regulation of *XIST/Xist* expression itself is a complex process involving multiple *cis*- and *trans*-acting factors [2]. In mice, the pluripotency transcription factors Oct4 and Nanog are negative regulators of *Xist* expression, such that a decrease in their expression early in post-zygotic development coincides with cell differentiation, upregulation of *Xist*, and onset of XCI [65–67]. Several subregions within the mouse Xic are involved in *Xist* regulation: *Tsix*, an antisense transcript of *Xist*, is a negative regulator of its expression and it protects the active X (Xa) from inactivation [68]; the noncoding RNA *Jpx* [69], X-pairing region [59,70], and protein-coding *Rnf12* [71,72] are positive regulators of *Xist* expression. There are a number of differences in the organization of the mouse and human XIC/Xic and there is a little sequence conservation between mouse and human [73,74]. Based on comparative analysis of Xic in several mammalian species, it seems that Xic is an evolutionarily labile locus and the orthologs across mammalian species act via multiple diverse strategies [61]. For example, in mice, *Jpx* is located 9 kb upstream of *Xist*, whereas in human it is separated by 90 kb [73,74]. Furthermore, human *TSIX* shows little conservation with mouse and is not transcribed through the entire *XIST*, as in mouse [75]. Also, in human fetal cells *XIST* and *TSIX* are coexpressed from the Xi [76], suggesting that *TSIX*-mediated downregulation of *XIST* might be less functional in humans.

More research is required to understand the regulation of XCI in human embryonic development. Despite the elegant picture that has emerged, some steps of the XCI process are still not completely understood. Many intriguing questions remain, such as how the cell counts the number of X chromosomes and decides how many to inactivate, how X chromosomes communicate with each other to retain one Xa and avoid a lethal state with two active or inactive X chromosomes, how *XIST* RNA recruits repressive chromatin markings, and how spreading of the inactivated state occurs along the Xi chromosome. These are active areas of current research [2,60].

5.4.1 Special Aspects of X Inactivation Relevant to Human Genetic Diseases

Not all X-linked genes are subject to X inactivation; some genes are robustly expressed from both the Xa and the Xi. It is estimated that 3% and 15% of genes escape XCI in mouse and human, respectively [77]. Again, there are fundamental differences between human and mouse genes that escape X inactivation. In mouse, genes escaping X inactivation are randomly distributed along X, whereas in humans they tend to cluster together. Only six of these genes overlap between mouse and human [74]. Genes that escape XCI are frequently expressed at lower levels from Xi than from Xa [78]. Further, based on DNA methylation analysis of X-linked promoters it is estimated that 12% of the X-linked genes show variable inactivation status among different somatic tissues [79]. Most, but not all, genes that escape XCI have a Y-linked homologue either within or outside of a PAR; some are functional, whereas others are pseudogenes. On the human X chromosome, the locations of genes that escape XCI are seemingly nonrandom, as the majority are clustered within regions of X that have the highest degree of homology to the Y chromosome [78].

Genes that escape XCI are important potential contributors to phenotypes of X-chromosome aneuploidy, both in the relatively common situation of X chromosome deficiency (females with Turner syndrome, karyotype 45,X; often abbreviated XO) and in supernumerary X-chromosome syndromes. Less than 1% of XO conceptuses survive to birth [80], with surviving individuals having a female phenotype and manifesting Turner syndrome, characterized by short stature, ovarian dysfunction, and a variety of somatic abnormalities such as webbed neck, high arched palate, increased carrying angle of elbows, aortic coarctation, renal malformations, and cognitive problems with visual-spatial perception and social interactions [81]. These problems are presumed to be due to haploinsufficiency of the few genes that normally escape X inactivation (i.e., are present in two active copies in normal XX females). Most of these genes are in the PARs: one of the genes, *SHOX*, has been implicated in the short stature phenotype of females with Turner syndrome, and short stature is observed in individuals with subchromosomal deletions encompassing this region on either the X or the Y chromosome [82]. Individuals carrying supernumerary X chromosomes,

XXX and XXY, have increased mortality rates, possibly resulting from overexpression of X-linked genes that escape XCI [74].

The human X chromosome is enriched for genes expressed in the brain [83,84], and many X-linked conditions present clinically as syndromic or nonsyndromic intellectual disability. Sex chromosome dimorphism makes the inheritance of X-chromosome conditions more complex than patterns observed for autosomal dominant or recessive inheritance. Most X-linked disorders affect males, with carrier females being either unaffected or mildly affected, depending on whether there is skewed or random XCI in critical tissues. Normally, XCI is random, resulting in cellular mosaicism, with two approximately equal populations of cells that express either the paternal or the maternal X chromosome. However, there are cases of significant skewing of X inactivation. Rare mutations within XIC may result in a failure to inactivate the X that carries these mutations, with preferential inactivation of the other X chromosome. Usually skewed XCI selection favors cells with a normal Xa, selectively silencing the mutation-bearing X chromosome [85]. However, there are rare examples of preferential activation of a mutant X chromosome in female carriers, resulting in more severe disease phenotype, such as rare cases of Duchenne muscular dystrophy in females [86], adrenoleukodystrophy caused by *ABCD1* mutation [87], or Xp11.22-23 duplication in females with intellectual disability, speech delay, and autism [88,89]. These situations can result from stochastic factors, presence of a genetic mutation or autosomal translocation on another X, or faster proliferation of cells carrying the mutation on the Xa [85]. There are at least two known exceptional situations in which female carriers of X-linked mutations are more severely affected than males: mutations in the ephrin-B1 gene [90] and in the epilepsy in females with mental retardation (EFMR) locus [91].

Rett syndrome is another example of females affected by an X-linked disorder and is discussed in a later section. Rett syndrome is a severe neurodevelopmental disorder (NDD) caused by heterozygous mutations of the X-linked *MECP2* gene. *MECP2* mutations are extremely rare in XY males, either because of early embryonic lethality or because sporadic *MECP2* mutations almost exclusively occur on the X chromosome transmitted from fathers to daughters, possibly secondary to deleterious effects of *MECP2* mutation on the oocyte [92].

In classical Rett syndrome, X inactivation is generally not skewed in the brain tissue of affected girls [93], but skewed XCI has been reported in females with very mild symptoms of Rett syndrome [94,95].

For all these reasons, when counseling families with suspected or known X-linked conditions, knowledge of the inactivation status of the gene, as well as the presence of XCI skewing, is an important factor. An X-inactivation skewing test based on DNA methylation analysis of a polymorphic CAG repeat within exon 1 of the *AR* gene [96] is routinely used in molecular diagnostic laboratories to address this question. It should be kept in mind that usually this test is performed on clinically accessible tissues, such as blood or buccal swabs, and the degree of skewing might vary among different tissues [85]. Therefore, the clinical relevance of skewed X inactivation may be difficult to interpret.

5.5 GENOMIC IMPRINTING

Genomic or parental imprinting was discovered by embryologists in the late 1970s, but the phenomenon of parent of origin-specific inheritance was observed 3 millennia ago, when mule breeders found that mating a female horse to a male donkey gives rise to a “mule,” whereas the reciprocal cross yields a phenotypically distinct equine, called a “hinny” [97]. Thus, in violation of classical Mendelian principles, two animals with the same diploid genome can be phenotypically distinct depending only on the parental origin of each haploid chromosome complement. This phenomenon can now be explained by epigenetics. The first embryological evidence of genomic imprinting came from experiments done in the mid-1980s in which attempts to reconstitute a viable mouse embryo entirely from either the maternal germline (gynogenetic conceptus derived from the fusion of two female pronuclei) or the paternal germline (androgenetic conceptus from two male pronuclei) were uniformly unsuccessful [98,99]. From these results the investigators immediately postulated that the maternally and paternally transmitted genomes (really epigenomes) are not functionally equivalent, and subsequent results analyzing mouse transgenes suggested that different DNA methylation on maternal versus paternal alleles might be an important mechanism accounting for this nonequivalence.

Based on extensive subsequent research, imprinting has been shown to affect a relatively small but

important subset of mammalian genes. The first endogenous imprinted genes, including the maternally expressed *Igf2r* gene, the paternally expressed *Igf2* gene, and the maternally expressed untranslated *H19* RNA, were discovered and studied in the early 1990s, first in mice and quickly thereafter in humans. Imprinting tends to be well conserved between the two species, but there are some exceptions in which specific genes, e.g., *Igf2r*, are functionally imprinted in mice but not in humans. Recently updated catalogs of imprinted genes in human and mouse (<http://www.geneimprint.com> and <http://www.otago.ac.nz/IGC>) document >100 imprinted transcripts in humans, which involve only around 1% of the total genome (Fig. 5.3). Many

of these imprinted genes play vital roles in embryonic growth and/or behavior and sometimes exhibit tissue- or developmental stage-specific monoallelic expression patterns [100]. This parent-specific expression of imprinted genes imposes a functionally haploid state at imprinted loci in normal tissues and hence deleterious effects from heterozygous mutations or hemizygous DNA deletions at these loci. This situation underlies a variety of human disorders that are inherited in a non-Mendelian fashion, that is, with phenotypic effects seen only after transmission of the mutant allele from one specific parent [101,102].

While most autosomal genes are expressed roughly equally from the two parental alleles, imprinted genes

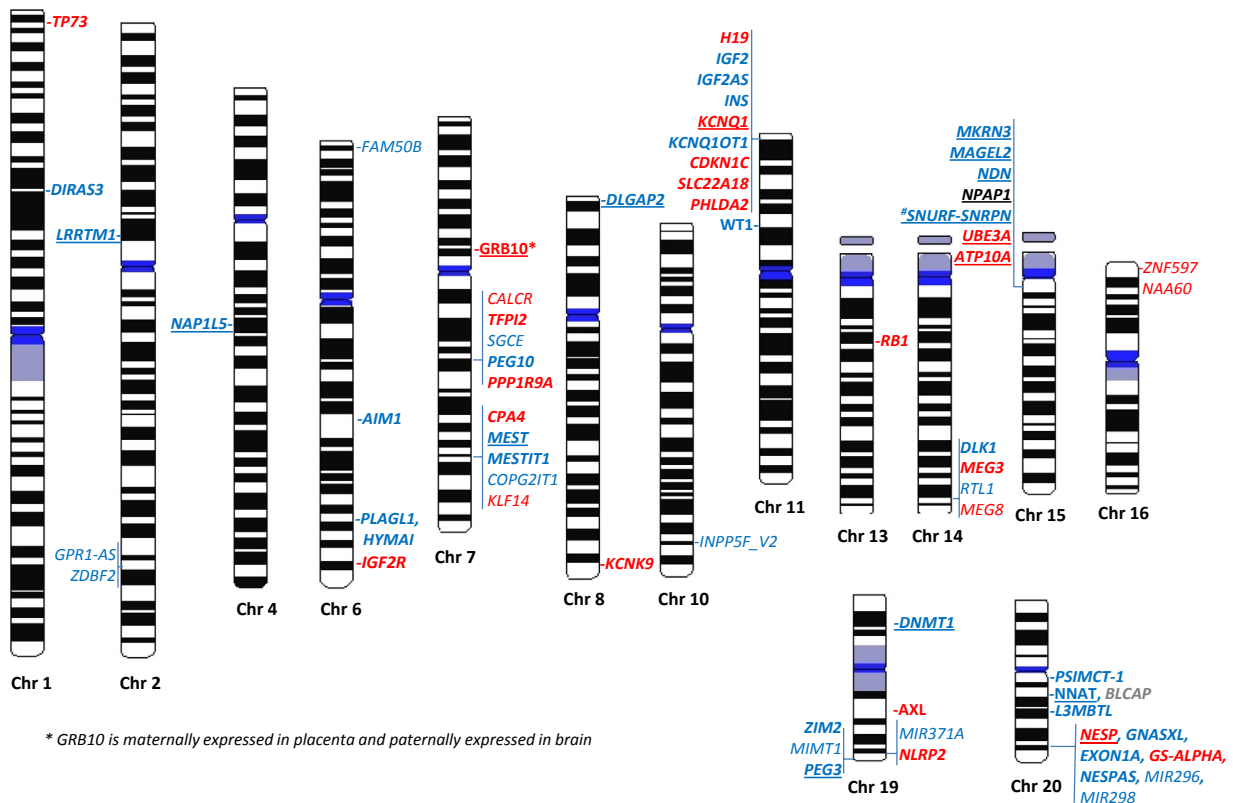


Figure 5.3 Idiograms of Human Imprinted Genes. Idiograms were generated using <http://www.dna-rainbow.org/ideograms/>. An idiogram of each human chromosome known to have an imprinted gene, based on the imprinted gene catalog last updated in January 2016 (<http://igc.otago.ac.nz>) and GeneImprint portal (<http://www.geneimprint.com>), is shown. Imprinted genes are listed on each idiogram if they were designated as imprinted in both of the aforementioned human imprinted gene catalogs. Blue genes are paternally expressed, red genes are maternally expressed, black genes have unknown parent-of-origin expression, gray genes have parental expression that is isoform dependent. Bold genes are implicated in growth, underlined genes play roles in neurodevelopment. Genes in *italic* have no reported function in growth or neurodevelopment.

are expressed preferentially or completely from only one allele, paternal or maternal, depending on the specific imprinted gene under consideration. The complex molecular mechanisms involved produce differential epigenetic marks on the two parental chromosomes: typically, allele-specific DNA methylation and/or allele-specific histone modifications. As noted previously, these epigenetic marks, also known as imprinted differentially methylated regions (DMRs), are established in the male and female gametes. As of this writing, there are 38 germline DMRs in humans, the majority originating in the oocyte [103]. Interestingly, most imprinted genes reside in 100-kb- to several megabase-sized gene clusters, which define subchromosomal imprinted domains. Imprinted domains generally have their own *cis*-acting imprinting control region, referred to as either a germline DMR or an imprinting center (IC). These ICs have acquired a crucial DNA methylation mark in the paternal or maternal germline, and they preserve this monoallelic mark in one or more tissues of the offspring, ultimately resulting in monoallelic or preferential allelic expression of one or more clustered imprinted genes [104]. Such domain-wide effects are mediated by several long-range *cis*-acting mechanisms, often including lncRNAs [2].

Geneticists have long debated the *raison d'être* for this surprising non-Mendelian phenomenon. The paternal–maternal intergenomic conflict hypothesis was proposed by evolutionary biologists to explain the observed patterns of imprinting for growth-related imprinted genes, and it currently appears to be a plausible biological rationale for imprinting [100]. Key to this hypothesis is the observation that imprinted genes that are expressed from paternal alleles tend to drive increased growth of the fetus and placenta and/or promote neonatal activity including suckling, thereby placing a greater demand on maternal resources, while, in contrast, imprinted genes expressed from maternal alleles tend to inhibit growth and downregulate demands on the mother [105]. Behavior of offspring can also affect demands on maternal resources, and evidence from human syndromes and mouse models has indicated a role for a subset of imprinted genes in neurodevelopment, cognition, and behavior, with possible links to common human psychiatric disorders [106].

5.5.1 Androgenetic and Gynogenetic Tumors: Hydatidiform Moles and Ovarian Teratomas

Complete hydatidiform moles (CHMs) are trophoblast tumors arising from an oocyte fertilized by one or rarely two sperm during early cell divisions; they retain a diploid set of paternal chromosomes and lose the maternal chromosome complement [107]. This scenario typically leads to paternal uniparental disomy (UPD) for all chromosomes; thus almost all CHMs are androgenetic tumors, and the dysregulated growth of these neoplasms is consistent with the physiological roles of paternally expressed imprinted genes in promoting trophoblast proliferation [105]. Nearly all hydatidiform moles have lost the maternal chromosomes; but there are rare cases of biparental moles that retain the maternal genome. Such cases usually occur in families and/or repeatedly in successive pregnancies, suggesting a genetic predisposition. In these tumors there is nonetheless markedly reduced expression of paternally imprinted/maternally expressed genes, indicating that the androgenetic gene expression state in these variant cases has resulted from failure of maternal imprinting, and that this state is essential for tumor formation [108]. Germline mutations in the *NLRP7* and *KHDC3L* oocyte-expressed genes have been identified in some women with this syndrome of familial biparental hydatidiform moles [109–111]. The broad loss of maternally methylated DMRs suggests that *NLRP7* normally functions to establish oocyte-specific methylation at imprinted DMRs [112].

Mature cystic teratomas constitute one of the most common types of benign ovarian tumors. They originate from a parthenogenetically activated oocyte after first meiosis and carry two maternal genomes and no paternal genome. This situation results in the formation of a cyst containing disorganized but histologically mature tissues from each of the three germ cell layers, with a predominance of ectodermal tissues [113–115]. It has been suggested that imprinting dysregulation is a major factor in the development of these tumors [115], and genome-wide disruption of normal methylation profiles at ICs has been observed in mature ovarian cystic teratomas [116].

5.5.2 Genomic Imprinting and Human Developmental Disorders

Imprinted genes typically function in growth regulation and neurodevelopment. Genetic and/or epigenetic

aberrations in these genes result in major abnormalities of intrauterine growth or postnatal cognition and behavior (Fig. 5.3). We will first discuss imprinting disorders that feature overgrowth or growth restriction as one of their major clinical characteristics.

Beckwith–Wiedemann syndrome (BWS): This disorder is an etiologically heterogeneous congenital overgrowth disorder with an incidence of ~1/13,700 live births [102,117,118]. However, this is likely to be an underestimation, as milder and overlapping phenotypes may not be ascertained. Clinically, diagnostic criteria include macrosomia (somatic overgrowth), macroglossia (large tongue), abdominal wall defects (omphalocele, often requiring surgical repair), ear creases and pits, kidney malformations, neonatal hypoglycemia, visceromegaly, and somatic hemihyperplasia. Certain tissues and organs can also become disproportionately large (tongues, kidneys, liver). There is also an increased incidence of embryonal tumors (7.5%). Most common are Wilms tumor and hepatoblastoma; but a variety of other tumor types are seen, including neuroblastoma, rhabdomyosarcoma, and adrenocortical carcinoma.

BWS is caused by epigenomic and/or genomic alterations that impact the regulation of imprinted genes on chromosome band 11p15.5 [119]. The 11p15.5 region

is organized into two distinct imprinted domains separated by a nonimprinted region (Fig. 5.4). Each domain has its own *cis*-acting imprinting control region: the IC1 domain in the telomeric region and the IC2 domain in the centromeric region. The IC1 domain contains the insulin-like growth factor (*IGF2*) and *H19* genes and the IC2 domain contains the *CDKN1C* (negative regulator of cell proliferation), *KCNQ1*, and *KCNQ1OT1* genes. The genes in these regions undergo parent-of-origin imprinting such that, typically, IC1 is methylated on the paternally derived chromosome, resulting in *IGF2* expression and silencing of *H19*. On the maternally derived chromosome IC2 is methylated, resulting in silencing of *KCNQ1OT1* and expression of *KCNQ1* and *CDKN1C* [104]. The molecular alterations detected in individuals with BWS include loss of methylation at IC2 (50%–60%), which silences *CDKN1C* plus several nearby maternally expressed genes; paternal UPD of 11p15 (20%–25%), which leads to biallelic *IGF2* expression; gain of methylation on the maternal allele of IC1 (5%–7%), which silences *H19* and activates expression of *IGF2*; and loss-of-function mutations of the maternally derived *CDKN1C* gene (3%–8%). Although cytogenetically detectable abnormalities involving chromosome 11p15 are present in fewer than

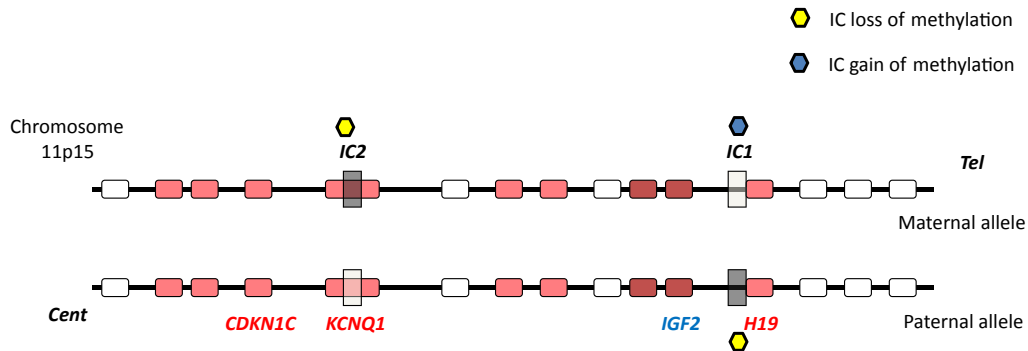


Figure 5.4 Schematic Representation of Imprinted Gene Clusters on Human Chromosome 11p15. Imprinted genes are indicated as filled boxes and nonimprinted genes as empty boxes. Paternally expressed genes are indicated in blue and maternally expressed genes in red. Hollow vertical rectangles show the location on normally unmethylated imprinting centers (IC) and filled vertical rectangles indicate that the IC is normally methylated. In the telomeric domain are two imprinted genes, *H19* and insulin-like growth factor 2 (*IGF2*). *IGF2* is a paternally expressed fetal growth factor and *H19* is a maternally expressed noncoding RNA. IC1 is usually methylated on the paternal chromosome and unmethylated on the maternal chromosome. The centromeric domain contains several imprinted genes, including *KCNQ1*, *KCNQ1OT1* (long noncoding RNA within the *KCNQ1* gene, not shown in this figure), and *CDKN1C*. IC2 at the promoter for *KCNQ1OT1* regulates the expression of *KCNQ1OT1*, which is a paternally expressed noncoding transcript that further regulates *in cis* the expression of the maternally expressed imprinted genes in the centromeric domain. *Tel*, telomere; *Cent*, centromere.

1% of affected individuals, genomic alterations such as microdeletions/microduplications may occur in association with methylation alterations. Most commonly these are microdeletions associated with gain of methylation at IC1. These epimutations occur as epigenetic programming errors early in postzygotic development, often resulting in tissue mosaicism for the cells carrying the epimutation.

Research studies have shown that the methylation alterations in BWS are not always restricted to the imprinted domain on chromosome 11. Multilocus imprinting defect (MLID), with mostly loss of maternal methylation at multiple imprinted loci, has been reported in children with BWS [120]. These MLIDs are frequently found to be mosaic, suggesting that the mechanism is failure of postfertilization maintenance of DNA methylation imprints [121]. Other causes of MLID include rare maternal-effect homozygous mutations in *NLRP2* identified in one family with BWS/MLID [122].

There are distinct phenotype–epigenotype correlations that have been well described in BWS [123]. Notably, Wilms tumors are more common in children with BWS who have either a gain of methylation in the IC1 imprinted domain or paternal UPD of 11p15. Studies report rare cases of Wilms tumor in children with BWS carrying epigenetic lesions in the IC2 imprinted domain [124]. Omphalocele is primarily associated with epigenetic lesions in the IC2 imprinted domain, specifically, loss of methylation at IC2 and mutation on the maternally derived *CDKN1C* gene [125].

Russell–Silver syndrome (RSS): RSS and BWS are two clinically opposite growth disorders that demonstrate opposite molecular alterations in the chromosome 11p15 imprinted domain. RSS is characterized by intra-uterine growth restriction, postnatal growth deficiency with normal head circumference, and characteristic facial features [126]. RSS is clinically heterogeneous, most cases are sporadic; occasionally there are familial cases. Both genetic and epigenetic alterations have been described in RSS, including loss of methylation of the chromosome 11p15 IC, IC1 (containing the *IGF2* and *H19* genes) in about 50% of cases, and maternal UPD for chromosome 7, a chromosome with known imprinted loci, in about 10% cases [127]. There have also been some case reports of gain of methylation at IC2, resulting in increased expression of *CDKN1C* in patients with RSS [128].

Prader–Willi and Angelman syndromes (PWS and AS): These two disorders are discussed together because they both map to the imprinted gene cluster on chromosome band 15q11–q13 (Fig. 5.5). These are two distinct neurogenetic disorders, both occurring at a frequency of 1 in 15,000–25,000 live births [129]. PWS is characterized by hypotonia and feeding difficulties early in life, with failure to thrive in early infancy, followed by a shift to excessive eating that can lead to morbid obesity in childhood. Individuals with PWS also exhibit developmental delay, mild to moderate intellectual disability, a distinctive behavioral phenotype including temper tantrums and obsessive–compulsive features, short stature, hypogonadism, characteristic facial features, scoliosis, and non-insulin-dependent diabetes mellitus. Consensus diagnostic criteria for PWS were first developed by Holm et al. [130], and further revised based on molecular diagnostics [131]. In contrast, AS is characterized by microcephaly, severe intellectual disability, severe speech impairment, gait ataxia, seizures, and a unique behavioral profile including frequent laughter, smiling, and excitability. Consensus diagnostic criteria were developed by Williams et al. [132].

The PWS and AS-associated locus is a large (~2.5 Mb) imprinted region that contains several paternally expressed (that is, expressed in an imprinted fashion from only the paternal allele) genes including *MKRN3*, *MAGEL2*, *NDN*, *C15ORF2*, *SNURF/SNRPN*, and a cluster of C/D small nucleolar RNAs (snoRNAs), plus a maternally expressed imprinted gene, *UBE3A*, and the *ATP10C* gene, which exhibits polymorphic maternal allele expression [129,133]. The expression of genes within the 15q11–q13 imprinted domain is regulated by an IC containing two critical control elements located at the 5' end of *SNURF/SNRPN* (Fig. 5.5). The differentially methylated IC and the promoter regions of *MKRN3* and *NDN* are methylated only on the maternal allele.

Accumulated molecular data have shown that AS results from functional loss of the maternally expressed *UBE3A* gene. Conversely, PWS arises from functional loss of paternally expressed genes in the 15q11–q13 region; no single candidate gene has been identified to date; however, atypical microdeletions suggest the important role of the snoRNA *SNORD116*. The most frequent cause of both syndromes (~70%) is a de novo ~5- to 7-Mb deletion, typically visible by standard cytogenetics. This recurrent deletion occurs as a result

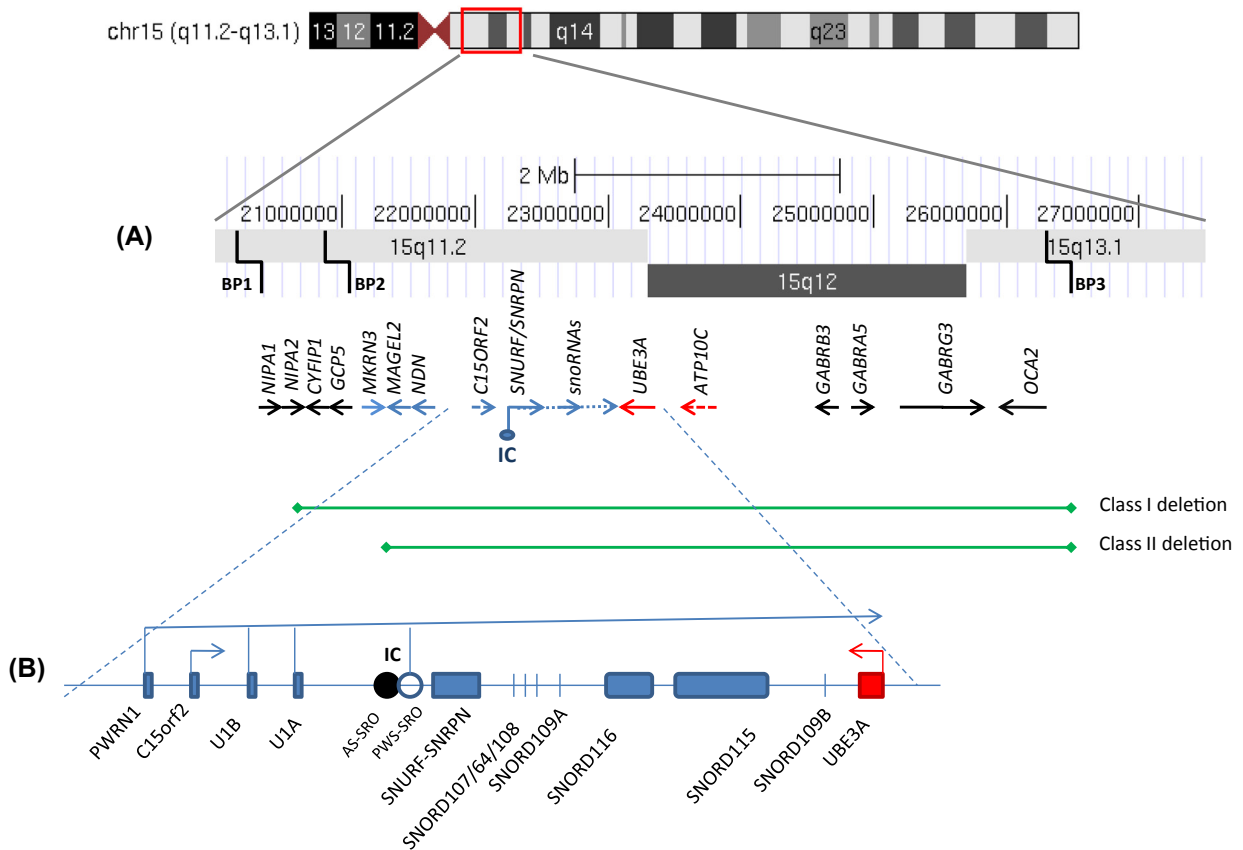


Figure 5.5 Schematic Maps of the Imprinted Domains on Chromosome Bands 15q11–q13. (A) Schematic representation of the 2-Mb domain on chromosome 15q11–q13 that carries the imprinted genes involved in Prader–Willi (PWS) and Angelman (AS) syndromes. Arrows represent the genes (note that sizes of genes are not to scale). Colors of the arrows represent pattern of expression: blue, paternal; red, maternal; black, biallelic. Dashed arrows show unconfirmed monoallelic patterns of expression. The names of the genes are shown above the respective arrows. This imprinted domain is under the control of imprinting centers (ICs) located upstream of *SNURF/SNRPN*. BP1, BP2, and BP3 are recurrent breakpoints. Green lines indicate regions of typical deletions (classes I and II) associated with AS (maternal deletions) and PWS (paternal deletions) and maternal duplications associated with autism spectrum disorder. (B) Zoom into the *SNURF/SNRPN–UBE3A* region. Circles are IC critical elements comprising two regulatory regions, the PWS smallest region of overlap (SRO), located around the *SNRPN* promoter (white circle), and the AS SRO, located 35 kb upstream (black circle). Boxes are genes, colors of boxes represent pattern of expression: blue, paternal; red, maternal. Arrows denote the direction of expression. *SNURF/SNRPN* is a multiexonic gene, expressed in multiple isoforms, with the first three coding exons encoding SNURF, a protein of unknown function, and *SNRPN* encodes SmN, a spliceosomal protein involved in mRNA splicing. *PWRN1*, *U1A*, and *U1B* are alternative transcription start sites of *SNURF/SNRPN*. Small nucleolar RNAs (snoRNAs) are encoded within introns of *SNURF/SNRPN*, with individual genes for *SNORD107*, *-64*, *-108*, *-109A*, and *-109B*, while *SNORD116* and *-115* are multicopy gene clusters. The function of the snoRNAs is not completely understood; they are possibly involved in modulating methylation or alternative splicing, because the snoRNAs in the PWS region lack the usual rRNA complementarity and each snoRNA gene may have multiple targets. Some of the splice variants of *SNURF/SNRPN* span *UBE3A* (*UBE3A-as*), which possibly regulates imprinted expression of *UBE3A*.

of low-copy-number repeats in the 15q11–q13 region, which increases susceptibility to nonhomologous recombination. These interstitial deletions involve the entire imprinting domain and several nonimprinted genes (class I and II deletions) [129,134]. In PWS, deletions occur on the paternal chromosome, whereas in AS deletions occur on the maternal chromosome. In addition to deletions of the 15q11–q13 region, other mechanisms that result in PWS and AS include UPD 15q11–q13 and imprinting defects. Specifically, 25%–30% of PWS cases result from maternal UPD for this chromosomal region, and conversely, 2%–5% of AS results from paternal UPD [129]. Imprinting defects have been reported in 1%–3% of PWS and 2%–4% AS patients. This type of defect results from an acquisition of maternal-type imprint (gain of DNA methylation) on the paternal chromosome in PWS and, conversely, loss of the maternal imprint (loss of DNA methylation) on the maternal chromosome in AS. Imprinting defects typically result from microdeletions in the 15q11–q13 region. The location of microdeletions is distinct in AS versus PWS cases; the smallest region of overlap (SRO) of the microdeletions in PWS is 4.3 kb in size and is located within the *SNURF/SNRPN* exon 1/promoter [135], while the AS SRO is 880 bp and is located more centromerically; ~35 kb upstream of exon 1 of *SNURF/SNRPN* [136]. For both syndromes, a microdeletion can be de novo or inherited from an unaffected carrier parent, through the male germline in PWS and through the female germline in AS [129]. Rarely, IC methylation defects occur in individuals with AS or PWS; these can result from failure of imprint erasure/acquisition or maintenance. No mutations in single genes have been demonstrated to date in PWS, but in 10% of AS cases mutations of the maternal copy of *UBE3A* have been documented [137,138]. This gene encodes an E3 ubiquitin ligase, a protein that functions in protein degradation. About 80% of mutations occur de novo, and about 20% are inherited from unaffected mothers [139]. By a mechanism that is not yet fully understood, maternal expression of *UBE3A* is observed only in the brain and not in other tissues, accounting for the neurobehavioral AS phenotype of deficient expression [140,141].

Last, in contrast to PWS, where most cases can be explained by genetic and/or epigenetic alterations at the 15q11–q13 imprinted cluster, 10%–15% of suspected AS cases have no identifiable molecular alteration [129].

5.5.3 Diagnostic Testing and Recurrence Risk in Imprinting Disorders

Identifying specific molecular defects in imprinting disorders provides important information for patient management and for estimating recurrence risk. Molecular diagnosis, specifically testing for abnormal DNA methylation in the relevant imprinted domains, is usually the first line of investigation for many imprinting disorders, including PWS/AS (domain in chromosome band 15q11–q13) and BWS/RSS (domains in chromosome band 11p15.5). The majority of molecular alterations within imprinting domains (UPD, IC epimutations, and microdeletions) can be diagnosed by assaying DNA methylation in the respective IC. Techniques such as Southern blot or PCR-based assays for DNA methylation [33,142] are useful, in addition to the more recently developed method, methylation-sensitive multiplex ligation-dependent probe amplification (MS-MLPA) [143]. The benefit of using MS-MLPA is that it can assess both methylation levels and copy number across several sites within the imprinted cluster. There are advantages and limitations of DNA methylation analysis for the various imprinting disorders. For PWS and AS, methylation analysis is sufficient to establish a clinical diagnosis, but cannot usually elucidate the underlying genetic mechanism, i.e., it cannot distinguish between methylation abnormalities due to a deletion, UPD, or IC mutation. In some cases, MS-MLPA can identify microdeletions within the IC in PWS and AS; however, typically further testing is required to identify the specific mechanism, which is critical to provide accurate recurrence risk information. This is particularly important as the recurrence risk can range from less than 1% for de novo deletions or UPD to 50% for some IC defects. For BWS, MS-MLPA can detect methylation abnormalities and strongly suggest 11p15 UPD when simultaneous gain of DNA methylation at H19 DMR and loss of DNA methylation at *KCNQ1* intronic DMR is identified. Additional testing for chromosome 11p15 UPD using either a PCR-based dosage assay or microsatellite genotyping should be undertaken if this molecular change is suspected, especially given the high frequency of somatic mosaicism. For RSS, DNA methylation is usually performed for the distal (IC1) IC on chromosome 11p15.5 and the imprinted genomic regions at 7p13 and 7q32 [126]. For AS and BWS, if no loss of DNA methylation at the respective ICs is detected, sequencing of *UBE3A* or *CDKN1C*, respectively, should be performed.

If no molecular defects are identified by methylation and mutation screening, comparative genome hybridization arrays to identify small atypical deletions or duplications could be pursued. It should be kept in mind that identification of small deletions or duplications is dependent on the resolution of microarrays, which can vary significantly among different diagnostic laboratories. Although chromosome translocations, inversions, and duplication infrequently cause imprinting disorders, their presence is associated with a significant risk of recurrence. A chromosome abnormality associated with an imprinting disorder may or may not have an associated methylation defect. For BWS, whether or not a methylation defect is present, high-resolution banding of the critical chromosomal region(s) should be considered for individuals who do not have a positive molecular diagnosis.

Individuals with imprinting disorders and UPD have not, as of this writing, been reported to transmit the molecular alteration or imprinting disorder to the next generation. In fact, theoretically this is very unlikely. Recurrence risk is usually low (<1%) in individuals with IC epimutations and imprinting disorders. This low risk implies that, in most cases, the IC defect can be rectified by the normal germline reprogramming mechanism; however, there are a small percentage of cases with IC defects that are heritable. These can be recognized from a positive family history or an associated genomic alteration. When inherited genetic alterations are present, such as IC

microdeletions in AS and PWS, *UBE3A* mutation in AS, or *CDKN1C* mutation in BWS, the associated recurrence risk rises to 50%. Such genetic alterations segregate in a Mendelian fashion, but the penetrance of the imprinting disorder depends on which parent transmits the mutation. For example, a parent carrying a mutation in *CDKN1C* (BWS) or *UBE3A* (AS) has a 50% chance of transmitting the mutation, but the imprinting disorder is expressed only if the mother transmits the mutation, since the paternally transmitted gene, whether it carries a mutation or not, is normally silenced in the male germline.

5.6 GENETIC DISORDERS CAUSED BY MUTATIONS IN EPIGENES

Within the last decade, an increasing number of Mendelian disorders have been recognized to be caused by mutations in genes that are important for maintaining normal epigenetic regulation, or “epigenes” [144]; [145–147]. Among these disorders, neurologic function, in particular intellectual disability, and growth dysregulation appear to be common features. Loss-of-function mutations in epigenes can disrupt normal establishment, maintenance, or reading of epigenetic marks, thereby resulting in altered chromatin structure and gene expression. Here we will provide examples of disorders caused by pathogenic mutations in four categories of epigenes: writers, erasers, readers, and chromatin remodelers. For a more comprehensive list see Table 5.1.

TABLE 5.1 Genetic Disorders Caused by Mutations in Epigenes*

Disorder	Gene	Function	Locus	OMIM
Sotos syndrome	<i>NSD1</i>	Histone methyltransferase (writer)	5q35.3	117550
Weaver syndrome	<i>EZH2</i>	Histone methyltransferase (writer)	7q36.1	277590
Cohen–Gibson syndrome	<i>EED</i>	Histone methyltransferase (writer)	11q14.2	617561
Wiedemann–Steiner syndrome	<i>KMT2A</i>	Histone methyltransferase (writer)	11q23.3	605130
Tatton–Brown–Rahman syndrome	<i>DNMT3A</i>	DNA methyltransferase (writer)	2p23.3	615879
Rubinstein–Taybi syndrome	<i>CREBBP</i>	Histone acetyltransferase (writer)	16p13.3	180849
	<i>EP300</i>	Histone acetyltransferase (writer)	22q13.2	613684
Hereditary sensory neuropathy type 1E	<i>DNMT1</i>	DNA methyltransferase (writer)	19p13.2	614116
Cerebellar ataxia, deafness, and narcolepsy, autosomal dominant	<i>DNMT1</i>	DNA methyltransferase (writer)	19p13.2	604121
Kleefstra syndrome	<i>EHMT1</i>	DNA methyltransferase (writer)	9q34.3	610253
Kabuki syndrome	<i>KMT2D</i>	Histone methyltransferase (writer)	12q13.12	147920
	<i>KDM6A</i>	Histone demethylase (eraser)	Xp11.3	300867

Continued

TABLE 5.1 Genetic Disorders Caused by Mutations in Epigenes^a—cont'd

Disorder	Gene	Function	Locus	OMIM
X-linked syndromic mental retardation; Claes–Jensen type	<i>KDM5C</i>	Histone demethylase (eraser)	Xp11.22	300534
Brachydactyly–mental retardation syndrome	<i>HDAC4</i>	Histone deacetylase (eraser)	2q37.3	600430
Cornelia de Lange syndrome	<i>HDAC8</i>	Histone deacetylase (eraser)	Xq13.1	300882
	<i>NIPBL</i>	Cohesin complex (reader)	5p13.2	122470
	<i>RAD21</i>	Cohesin complex (reader)	8q24.11	614701
	<i>SMC1A</i>	Cohesin complex (reader)	Xp11.22	300590
	<i>SMC3</i>	Cohesin complex (reader)	10q25.2	610759
Rett syndrome	<i>MECP2</i>	Methyl-binding domain (reader)	Xq28	312750
CHARGE syndrome	<i>CHD7</i>	Chromatin remodeling (remodeler)	8q12.2	214800
Coffin–Siris syndrome	<i>ARID1B</i>	Chromatin remodeling (remodeler)	6q25.3	135900
	<i>ARID1A</i>	Chromatin remodeling (remodeler)	1p36.11	614607
	<i>SMARCA4</i>	Chromatin remodeling (remodeler)	19p13.2	135900
	<i>SMARCB1</i>	Chromatin remodeling (remodeler)	22q11.23	135900
	<i>SMARCE1</i>	Chromatin remodeling (remodeler)	17q21.2	135900
Nicolaides–Baraitser syndrome	<i>SMARCA2</i>	Chromatin remodeling (remodeler)	9p24.3	601358
α -Thalassemia mental retardation syndrome, X-linked	<i>ATRX</i>	Chromatin remodeling (remodeler)	Xq21.1	301040
Floating–Harbor syndrome	<i>SRCAP</i>	Chromatin remodeling (remodeler)	16p11.2	136140
Epileptic encephalopathy	<i>CHD2</i>	Chromatin remodeling (remodeler)	15q26.1	615369
Autism spectrum disorder susceptibility	<i>CHD8</i>	Chromatin remodeling (remodeler)	14q11.2	615032

^aThere are over 55 known epigenes—readers, writer, erasers, and chromatin remodelers—that are associated with Mendelian disorders. This table includes examples for each type of epigene. For a more detailed list see Kleefstra et al. *Neuropharmacology* 2014;80:83–94 and Björnsson. *Genome Res* 2015;25:1473–81.

5.6.1 Human Disorders due to Mutations in Writers of Epigenetic Marks

Writers of epigenetic marks place specific modifications on either the DNA or its associated histone proteins, which constitutes one of the mechanisms that control gene transcription. Hemizygous, typically loss-of-function, mutations in a variety of epigenes encoding writers, including histone methyltransferases, histone acetyltransferases, and DNA methyltransferases, have been associated with well-known Mendelian disorders.

Sotos syndrome (*NSD1*): Sotos syndrome is caused by hemizygous mutations (including deletions) of the *NSD1* gene, which encodes a histone methyltransferase. Sotos syndrome is an overgrowth condition associated with macrocephaly, facial dysmorphism, advanced bone age, and learning difficulties or intellectual disability [148]. *NSD1*, which has a catalytic lysine methyltransferase Su(var)3-9, Enhancer-of-zeste and Trithorax (SET) domain and four zinc-binding plant homeodomain

(PHD domains), functions primarily to mono- and dimethylate H3K36 [149]. The role of H3K36 methylation is not completely understood; in model organisms it has been found within gene bodies of expressed genes and is associated with suppression of intragenic transcriptional initiation [150]. Chromatin immunoprecipitation (ChIP)–ChIP experiments using promoter microarrays have shown that *NSD1* binds to promoters of genes, playing a role in regulating various cellular processes, such as cell growth/cancer, keratin biology, and bone morphogenesis [151]. In addition, four of the *NSD1* PHD domains normally bind histone H3 methylated at K4 and K9; this binding is disrupted by most of the point mutations found in Sotos syndrome [152].

Weaver syndrome (*EZH2*): Weaver syndrome is characterized by pre- and postnatal overgrowth, accelerated osseous maturation, characteristic craniofacial appearance, and developmental delay. Mutations in *EZH2* (*Enhancer of Zeste, Drosophila, homolog 2*) are

the primary cause of this syndrome [153,154]. *EZH2* encodes the catalytic component of PRC2, which regulates chromatin structure and gene expression through trimethylation of lysine 27 of histone H3. The core complex of PRC2 is made up of three components: *EZH2*, *EED*, and *SUZ12*. It has been established that mutations in all three components of the PRC2 complex cause Weaver or a Weaver-like syndrome [155,156]. Further studies are needed to determine if the phenotypes associated with mutations in these different proteins represent a single clinical entity—Weaver syndrome—or diverse syndromes with overlapping clinical features.

Kabuki syndrome (*KMT2D*): Kabuki syndrome is characterized by typical facial features, skeletal and dermatoglyphic anomalies, postnatal growth deficiency, and mild to moderate intellectual disability [157]. Mutations in *KMT2D* constitute the primary cause of Kabuki syndrome and are identified in about 80% of individuals. A second gene, *KDM6A*, accounts for about 6% of cases (see next section on erasers). *KMT2D* belongs to the SET1 family of histone H3K4 methyltransferases. It has a catalytic SET domain, five PHD domains, and an HMG-I binding motif [158]. *KMT2D* is a part of a multiprotein complex (which includes *KDM6A*) that catalyzes mono-, di-, and trimethylation of H3K4 [159] and regulates the expression of a wide range of downstream genes. H3K4 trimethylation is associated with active transcription [150], and the reduction of *KMT2D* in human HeLa cells results in downregulation of a number of genes involved in cell adhesion, cytoskeleton organization, transcriptional regulation, and development [159]. Interestingly, in a mouse model, *Kmt2d* has been shown to be crucial for the epigenetic reprogramming that takes place before fertilization in oocytes, such that reduced trimethylation of H3K4 due to deficiency of *Kmt2d* results in anovulation [160].

5.6.2 Human Disorders Due to Mutations in Erasers of Epigenetic Marks

Erasers of epigenetic marks remove specific modifications from either the DNA or its associated histone proteins and constitute another mechanism that controls gene transcription. Hemizygous mutations in several epigenetic erasers, including histone demethylases and HDACs, have been associated with well-known Mendelian disorders.

Kabuki syndrome (*KDM6A*): Kabuki syndrome, as described in the writer section, can be caused by

loss-of-function mutations in both a writer, *KMT2D* (80%), and an eraser, *KDM6A* (6%). *KDM6A* encodes a histone H3 lysine 27 (H3K27) demethylase that is important for general chromatin remodeling, creating a closed chromatin mark. *KDM6A* interacts with *KMT2D* in a conserved SET1-like complex that targets H3K4, and they are both trithorax complex-like proteins. Therefore, the overall effect of mutations in either gene is predicted to be the same: regulating chromatin state and transcriptional activity at a specific set of target genes [161].

X-linked mental retardation (*KDM5C*): Mutations in the X-linked gene *KDM5C*, encoding a histone demethylase, cause a spectrum of phenotypes, ranging from syndromic to nonsyndromic intellectual disability. The clinical features in males with *KDM5C* mutations include mild to severe intellectual disability, epilepsy, short stature, aggressive behaviors, and microcephaly [162–167]. *KDM5C* escapes X inactivation and has a functional Y-linked homologue, *KDM5D*. Therefore, female heterozygous mutation carriers are usually unaffected; some demonstrate mild intellectual disability or learning difficulties [84]. *KDM5C* has several conserved functional domains, including the Bright/ARID domain responsible for DNA binding, the catalytic JmjC domain, and two PHD domains, responsible for histone binding [168,169]. *KDM5C* can bind the repressive histone mark H3K9me3 and removes the active mark H3K4me3/2, thus establishing a repressive chromatin state [170,171]. *KDM5C* point mutations identified in humans can suppress demethylase activity and/or H3K9me3 binding in vitro, depending on the location of the mutation [171]. ChIP in cell lines showed that *KDM5C* colocalizes with REST, a transcriptional repressor acting via neuron-restrictive silencing elements in the promoters of target genes such as *BDNF* and *SCN2A*, suggesting that downregulation of *KDM5C* activity impairs REST-mediated neuronal gene regulation [172].

5.6.3 Human Disorders Due to Abnormal Readers of Epigenetic Marks

Epigenetic regulators can affect gene expression by encoding proteins that read epigenetic marks in a temporal and spatial-specific manner.

Rett syndrome (*MECP2*): Heterozygous mutations of the X-linked gene *MECP2* cause Rett syndrome, a condition that occurs almost exclusively in girls. Rett

syndrome is characterized by developmental arrest between 5 and 18 months of age, followed by regression of acquired skills, loss of speech, stereotypical movements, microcephaly, seizures, and severe intellectual disability [173]. The function of MECP2 has been studied extensively but the mechanism by which its deficiency results in the phenotypes of Rett syndrome remains incompletely understood. Initially, MECP2 was identified as a protein capable of binding methylated DNA [174]. It was found to have abundant binding sites distributed throughout the genome and was demonstrated to function in repression of transcription [175–177]. The best established mechanism by which MECP2 downregulates gene expression is through recruitment of HDACs, which transform specific regions of chromatin into a repressive state by removing acetyl groups from histones H3 and H4 [175,177,178]. There is growing evidence that the role of MECP2 in transcription regulation is more complex; for example, in mice it was shown to bind to the transcriptional activator CREB to activate transcription of a large number of genes in the hypothalamus [179]. Further, MECP2 deficiency is associated with dysregulation of specific genes, such as *BDNF*, which has been shown to have MECP2 binding sites. Reduction of *Bdnf* in a mouse model mimics some features of the *Mecp2*-null mouse phenotype [180] and *Bdnf* overexpression in these mice can partially rescue the phenotype by improving locomotor function, extending life span [181], and rescuing synaptic dysfunction [182]. These data suggest that *Bdnf* is indeed an important and clinically relevant *Mecp2* transcriptional target. Findings suggest that *Mecp2* is almost as abundant as histone H1 in mouse neurons but not in glia [183], so *Mecp2* function in neurons might affect genome-wide chromatin remodeling in addition to regulating the expression of specific genes. Moreover, targeted deletion of *Mecp2* in mice results in increased expression of repetitive elements in neurons [183], prompting investigators to suggest a role for this protein in limiting overall transcriptional noise in neurons and the capacity of neurons to respond to environmental signals [184].

5.6.4 Human Disorders Due to Mutations in Chromatin Remodelers

Chromatin remodeling is an important epigenetic mechanism critical for nucleosome mobility and

transcriptional control. Chromatin remodelers often function as part of large macromolecular protein complexes that can reorganize nucleosome structure in an ATP-dependent fashion [185]. Many such genes have been implicated in Mendelian disorders.

CHARGE syndrome (*CHD7*) and other disorders due to mutations in chromodomain helicase enzymes: CHARGE syndrome is characterized by coloboma, heart defects, atresia of the choanae, retardation of growth and development, genital hypoplasia, and ear abnormalities, including deafness and vestibular disorders [186,187]. Nonsense or missense mutations and deletions that result in haploinsufficiency of the *CHD7* gene cause the majority of CHARGE syndrome cases [187,188]. *CHD7* is an ATP-dependent chromodomain helicase chromatin-remodeling protein involved in the formation of several large protein complexes that regulate the movement of nucleosomes along DNA, thereby affecting the activity of numerous signaling pathways during embryonic development. The *CHD7* gene is expressed in ES cells, and its expression becomes restricted to specific tissues, including the brain, eye, heart, and ear, during differentiation. *CHD7*-containing protein complexes bind to DNA at specific sites, the majority of which constitute regulatory elements such as gene promoters or enhancers [189]. The epigenetic effects of *CHD7* on chromatin and gene regulation appear to vary both temporally and spatially, depending largely upon the function of the protein complex with which it interacts. Other genes in the Chromodomain Helicase DNA-binding protein family, including *CHD2* and *CHD8*, have been associated with epileptic encephalopathy and susceptibility to autism, respectively.

α -Thalassemia/mental retardation syndrome, X-linked (ATR-X) (*ATRX*): Mutations in an X-linked gene, *ATRX*, cause ATR-X, which is characterized by severe intellectual disability, facial dysmorphism, urogenital anomalies, and α -thalassemia. As the *ATRX* gene normally undergoes X inactivation, affected individuals are almost exclusively males, while females usually are unaffected, due to preferential X inactivation of the chromosome with the *ATRX* mutation [190]. The *ATRX* protein is involved in epigenetic regulation through two functional domains: an ATP/helicase domain and an *ATRX*-Dnmt3-Dnmt3L (ADD) domain that shares homology with de novo methyltransferases. The ATP/helicase domain is proposed to be involved in nucleosome repositioning and making DNA more accessible for

protein binding [191], while the ADD domain has been shown to bind histone H3 tails with the silencing mark H3K9me3 but not the active mark H3K4me3/2 [192]. In terms of genomic targets, ATRX has been shown to localize to the nucleus in heterochromatin, telomeric/subtelomeric chromosomal regions, ribosomal DNA (rDNA) and promyelocytic leukemia bodies [193–195]. Furthermore, peripheral blood cells of ATR-X patients exhibit changes in DNA methylation of rDNA, subtelomeric repeats, and Y-chromosome-specific satellites [193]. By ChIP-sequencing, it was established that in erythroid cells ATRX binds to CpG-rich tandem repeat sequences clustered at subtelomeric regions, thereby affecting the expression of associated genes, including α -globin, which accounts for the α -thalassemia phenotype of ATR-X syndrome [196].

5.7 METHODS FOR STUDYING EPIGENETIC MARKS

5.7.1 Mapping DNA Methylation

One of the first methods to score DNA methylation at specific loci was Southern blotting of genomic DNA digested with methylation-sensitive restriction enzymes [197]. Certain restriction enzymes (e.g., *HpaII*, *SmaI*, *NotI*) that contain a CpG as part of their recognition sequence do not cut at methylated sites. Therefore, failure to cleave by a methyl-sensitive restriction enzyme is evidence of DNA methylation at that site. Restriction enzymes can also be used in combination with microarray platforms to evaluate genome-wide DNA methylation patterns, including promoter methylation and allele-specific methylation [198,199]. Methylation-specific PCR, based on predigestion of genomic DNAs with such restriction enzymes, provides a semiquantitative measurement of DNA methylation levels. While such methods have generally been supplanted by approaches based on bisulfite conversion of DNA (see later), assays with methylation-sensitive restriction enzymes can still be useful for independently validating the results from bisulfite-based methods.

Chemical conversion of DNA via sodium bisulfite is the gold standard for DNA methylation analysis, allowing for quantitative analysis of CpG methylation. Sodium bisulfite deaminates nonmethylated dC's to dU residues. During subsequent PCR amplification, the dU's are paired with T's and amplified as A/T base pairs; however, if a C is methylated, the DNA sequence

does not change, and a C will be paired with a G [200]. Pyrosequencing after bisulfite conversion can determine average DNA methylation levels at individual CpG sites across a short genomic region. A more informative approach involves amplification of bisulfite PCR products followed by cloning and Sanger sequencing or deep massively parallel sequencing. This approach reveals DNA methylation levels of a larger number of individual CpG sites and can be useful for detection of heterogeneous and allele-specific patterns of methylation [201].

By combining sodium bisulfite conversion and microarrays or massively parallel bisulfite sequencing, net and allele-specific DNA methylation patterns are being mapped genome-wide in an ever-increasing number of cell types from individuals with a variety of diseases as well as controls (e.g., [116,202]). These methods to characterize the methylation status of all or many of the ~28 million genomic CpG sites can be broadly classified into two categories: microarray and next-generation sequencing (NGS) based. The Illumina Methylation BeadChip assays (450K and EPIC arrays) are the most commonly used methods for epigenome-wide association studies (EWAS) and related types of studies [203,204]. The BeadChips are limited by a fixed number of probes, but are widely used because of their low cost, low DNA input requirement, and significantly reduced sample processing time. NGS-based approaches such as whole-genome bisulfite sequencing are generally regarded as the gold standard genome-wide methods because they provide the broadest coverage at single-base resolution. So far, this definitive approach has been less widely used, since the read depth required, data storage requirements, and computational processing time for the resulting terabyte-scale data have remained cost prohibitive for many researchers [205]. Two alternative NGS approaches, reduced representation bisulfite sequencing (RRBS) [206] and methyl-capture sequencing (MC-Seq), aim to enrich for specific regulatory genomic regions known to be targets for epigenetic dysregulation (i.e., CpG islands, CpG shores, and CpG shelves, close to transcriptional start sites and promoters) [207]. RRBS uses a restriction enzyme that recognizes CpG's to enrich for CpG-rich regions, while MC-Seq uses target-specific bait sequences. Going forward, efforts are under way to adopt single-molecule sequencing technologies for the direct detection of multiple types of DNA methylation in unamplified DNA, with a view to analyzing various sample types, including low-abundance specimens [208].

5.7.2 Mapping Histone Modifications and Chromatin Structure

ChIP followed by microarray-based assays or massively parallel sequencing is the primary method to determine the interactions of histone proteins, or chromatin-associated proteins such as transcription factors or insulator binding proteins, with DNA [209]. Thus, ChIP-Seq has become the standard approach for mapping the genomic distributions of histone modifications and DNA binding proteins. Cells or tissues are briefly fixed with formaldehyde to crosslink the chromatin-associated proteins to the DNA. The fixed chromatin is then fragmented, usually by sonication, to generate DNA fragments of appropriate size for analysis. Next, immunoprecipitation of chromatin with an antibody specific for a specific histone modification or protein is used to pull down associated genomic regions. The DNA is then freed from crosslinks and identified and analyzed by NGS to map specific types of epigenetic states genome-wide in various tissues and disease states. Large multi-investigator projects, such as ENCODE, the NIH Roadmap Epigenetics Mapping Consortium, and the International Human Epigenome Consortium, have mapped both DNA methylation and a large number of histone modifications in many human cell types and tissues, including cancer cell lines and pluripotent cells, with the data being made publicly available [210] (see websites).

The study of the three-dimensional structure of chromatin is also an area of active research. Chromatin conformation capture (3C) uses ligation of DNA–DNA interactions followed by PCR to identify interacting genomic regions [211]. As genome-wide techniques have improved, the number of 3C-derivative techniques has increased dramatically, supporting the creation of genome-wide interaction maps using massively parallel sequencing [212]. These data have also been made publicly accessible as genome browser tracks (see websites) [213]. Additional epigenetic features of interest can also be targeted for analysis; these include DNase hypersensitive and transposase-accessible regions (two assays for open chromatin), transcription factor binding sites, multifunctional epigenetic domains, and FANTOM5 enhancers [214,215], as annotated by ENCODE, FANTOM, IHEC, and the RoadMap Epigenomics Consortium (see websites).

5.7.3 Epigenome-Wide Association Studies

An emerging approach in epigenomics is performing EWAS. EWAS seek to use case–control or cohort designs to detect changes in DNA methylation in various disease states. These include disease pathology (i.e., disease progression, not genetic susceptibility); environmental factors, such as dietary influences, including over- or undernutrition; exposure to environmental toxins; and effects of substance abuse, including common situations such as alcohol consumption and cigarette smoking [216,217]. Issues of experimental design and caveats for interpreting EWAS data have been discussed in several papers, including by our groups [218–220], but the number of studies completed as of this writing is smaller than for genome-wide association studies (GWAS), and the criteria for calling true-positive “hits” have yet to be standardized. Among the phenotypes that have been investigated are body mass index and type 2 diabetes (T2D) [221–223], cardiovascular phenotypes [224–226], Alzheimer disease [227–229], autoimmune and inflammatory diseases [230,231], and neuropsychiatric disorders [232,233].

5.8 CANCER EPIGENETICS

There is extensive literature on epigenetics and cancer. In this chapter we will restrict our discussion to general principles, with selected illustrations. Changes in DNA methylation were the first epigenetic alterations identified in cancer [234], and subsequent work over 3 decades has shown that both hyper- and hypomethylation are important and pervasive pathogenic mechanisms in early as well as late stages of human tumorigenesis [235,236]. Not surprisingly, histone modifications and RNA expression, including miRNA and lncRNA, are also altered in cancer [237–239]. Genetic and epigenetic international mapping projects have made it increasingly apparent that proteins regulating epigenetic marks, including writers, erasers, and readers, as well as chromatin remodeling complexes, are dysregulated in cancer [11,235]. In fact, over 50% of cancers have mutations in enzymes that regulate the epigenome [235]. Aberrant epigenetic profiles, such as DNA methylation, histone modifications, lncRNA, and miRNA signatures, are being used as biomarkers to diagnose specific tumors and to predict recurrence risk, and to test the efficacy of therapies in certain types of cancer, including colorectal cancer and acute myeloid leukemia [240,241].

5.8.1 DNA Hypermethylation in Cancer

CpG hypermethylation in gene promoters is the best-characterized epigenetic abnormality in human malignancies. A common paradigm in cancer epigenetics is hypermethylation of the CpG-rich promoter regions of tumor-suppressor genes, resulting in epigenetic silencing of these genes [22]. Indeed, for some of the most important tumor suppressors, such as the *CDKN2A* gene encoding the p16 cell cycle inhibitor, promoter hypermethylation can be the most common mechanism underlying their functional loss during tumor formation, with the corresponding genetic pathways for loss of function (deletion/mutation) being utilized less commonly [242]. Hypermethylated promoter DNA is associated with virtually every type of human tumor, with each type of tumor having its own signature of methylated genes, such as the methylation of *GSTP1* in prostate cancer, the von Hippel–Landau syndrome gene *VHL* in renal cancer, the mismatch repair gene *MLH1* in colon and endometrial cancers, and sometimes *BRCA1* in breast cancer [243–248]. In some of these examples, the same tumor-suppressor gene is mutated or methylated as alternative pathways in the same tumor type: loss-of-function mutations in *MLH1* and *VHL* are found in the germlines of patients with hereditary colon and renal cancer, respectively, and these same genes are hypermethylated and silenced in sporadic tumors of the same histologic type [248,249].

While the gain of DNA methylation is often discussed as a late event in tumor progression, CpG hypermethylation in specific sequences often occurs early in cancer formation, sometimes preceding tumorigenesis. Examples of early epigenetic aberrations can also be cited in other adult malignancies: in cigarette smokers, *CDKN2A* promoter methylation occurs in dysplastic bronchial epithelial cells prior to the formation of overt lung cancers [250], and promoter hypermethylation of tumor-suppressor genes is already detectable in the premalignant lesion Barrett's esophagus [251,252]. One of the best substantiated examples of a very early epigenetic lesion predisposing to subsequent tumor formation is gain of methylation of the *H19* DMR on the maternal allele, which leads to loss of imprinting of *IGF2* expression and can often be detected in nonneoplastic kidney cells in both BWS-associated and sporadic cases of the pediatric kidney cancer Wilms tumor [236].

DNA hypermethylation has attracted much attention as a biomarker for cancer detection and classification. To be clinically applicable, an ideal tumor biomarker must be specific for cancer, and readily detectable in clinical specimens obtained through minimally invasive procedures. DNA hypermethylation seems to fulfill these requirements and has been considered to be a promising biomarker. Examining the methylation of a subset of genes (*GSTP1*, *APC*, *RASSF1*, and *MDR1*) distinguished primary prostate cancer from benign prostate tissues with sensitivities and specificities of greater than 90% [253,254]. DNA methylation alterations can be detected and used as biomarkers in fecal samples for colorectal cancer, urine for bladder cancer screening, and sputum to predict the occurrence of lung cancer [254–256]. For colorectal cancer, DNA hypermethylation at a set of specific genes has been defined as a biomarker for tumor stage as well as for recurrence risk estimation [257]. Further investigation is under way to support the adoption of this biomarker into clinical practice.

5.8.2 DNA Hypomethylation in Cancer

Global DNA hypomethylation in cancer cells was in fact identified prior to promoter hypermethylation [258], with studies indicating that genome-wide 5mC is reduced an average of 10% in a number of different tumor types [234,259]. The net decrease in the genomic methyl-C content in cancer cells often exceeds the localized increases in DNA methylation [260]. There is evidence that hypomethylation of DNA can result in genomic instability, leading to mutations, deletions, amplifications, inversions, and translocations [261–263]. Regions of hypomethylation in many cancers correspond to lamina-associated domains (LADs) and large organized chromatin K9 modifications (LOCKs), which are partially methylated in normal tissues [235]. Hypomethylation of LADs and LOCKs is a common feature of solid tumors, including colorectal, breast, and pancreatic cancer [264,265]. In some tumor types, including cervical carcinoma and meningioma, increasing hypomethylation has been correlated with progression from benign to malignant disease [266,267]. Hypomethylation can also lead to upregulation of miRNAs, e.g., miR-21 in chronic lymphocytic leukemia, which can lead to a downstream cascade of aberrant gene expression [238,260,268].

5.8.3 Abnormalities of Histones and Histone Modifications in Cancer

Epigenetic alterations in cancer are not restricted to DNA methylation. Genome-wide mapping of histone marks has demonstrated global changes in histone modifications in most cancer types. For example, acetylated H4 lysine 16 and H4 lysine 20 trimethylation are commonly identified alterations in many tumors correlating with repression of gene expression [269]. HDACs have been found to be overexpressed in a number of cancer types, and in some cancers there can be dysregulation of histone acetyltransferases due to translocations resulting in deleterious gene fusion products [270–272]. Aberrant histone methylation of H3K9 and H3K27 also results in gene silencing in many cancers [26,273,274]. *EZH2*, a histone methyltransferase of H3K27, is frequently overexpressed in breast and prostate tumors, in addition to other tumors [275,276]. Loss-of-function mutations in *EZH2* have also been seen in myeloid malignancies and T cell leukemias, demonstrating the importance of H3K27 methylation [277,278]. As is true for aberrant DNA methylation, most abnormalities in histone modifications in cancer are not yet explained by a single genetic lesion. However, with high-coverage sequencing technologies an increasing number of chromatin-modifying enzymes, such as CREB, JARID1C, *EZH2*, and the SWI/SNF family proteins hSNF5/INI1 and PBRM1, are now being found mutated in specific types of human cancers [235,277,279–282]. Mutations have been found not only in histone-modifying proteins but also in histones themselves. A recurrent mutation, p.K36M of histone H3 variants, which disrupts H3K36 methylation, has been identified in a number of tumor types, including glioblastomas and head and neck tumors [283,284]. In diffuse intrinsic pontine glioma, a fatal pediatric brain tumor, this mutation has been shown to be a driver of oncogenesis [285].

5.8.4 Aberrant miRNA and lncRNA Expression in Cancer

Comparisons of tumor tissues and corresponding normal tissues have revealed global changes in miRNA expression during tumorigenesis [286]. Upregulation of *miR-21*, which targets *PTEN*, occurs in glioblastomas [287], and in chronic lymphocytic leukemia *miR-15* and *-16*, which target the antiapoptotic gene *BCL2*, are downregulated [288]. Similarly, *let-7*, which targets the oncogene *RAS* is downregulated in lung cancer [288].

These alterations in miRNA expression may occur through a number of mechanisms, including chromosomal abnormalities, transcription factor binding, and epigenetic alterations [289]. Silencing of miRNA expression has been shown to occur by aberrant hypermethylation in a number of cancers [290,291] and, like other epigenetic factors discussed earlier, the role of miRNA dysregulation in cancer has been validated genetically by findings of DNA deletions encompassing miRNA genes, e.g., on chromosome 13 in chronic lymphocytic leukemias [292]. The possibility of introducing miRNA as a therapy using a variety of methods is being explored in vitro and also in in vivo models of many different cancers [293]. In an early study restoration of *let-7* in a mouse model of lung cancer resulted in decreased tumor growth [294].

lncRNAs have also been shown to be dysregulated in cancer [295,296]. Overexpression of *HOTAIR* in early-stage breast cancer is predictive of progression to metastatic disease as well as overall survival [297]. Further studies have shown that aberrant expression of *HOTAIR* is associated with cancer progression in a number of tumor types [298]. *MALAT1* overexpression predicts metastatic progression in early-stage non-small-cell lung cancer (NSCLC) [299]. Decreased expression of *Malat1* in a mouse model reduces metastasis of lung carcinomas [300]. Overexpression of *MALAT1* has also been reported in other cancer types, including breast cancer [301]. These studies suggest that lncRNAs may provide targets for therapeutics that could be applicable to a wide range of tumor types [296].

5.8.5 Therapies Targeting Epigenetic Modifications

A number of classes of medications that target epigenetic pathways that are disrupted in cancer have been approved by the US Food and Drug Administration (FDA), either as a treatment for cancer or for clinical trials [240,302]. The DNA methylation inhibitors 5-azacytidine (azacytidine) and 5-aza-2'-deoxycytidine (decitabine), are nucleoside analogues that get incorporated into the genomes of growing tumor cells and inhibit DNA methyltransferase enzymes, leading to progressive loss of DNA methylation with each S phase of the cell cycle. These medications have been approved for use in the treatment of MDS and have shown some promise for treating acute lymphoblastic leukemia and other hematological malignancies. [303–305]. HDAC

inhibitors (HDACi), which have been approved by the FDA (including vorinostat, suberoylanilide hydroxamic acid), are being used with good results in treating patients with cutaneous T cell lymphoma in the United States [306]. There have been a number of clinical trials in solid tumors with HDACi with little success, due to adverse events [307]. This problem can be addressed by using an HDACi in combination with other chemotherapeutics, specifically for NSCLC and ovarian cancer [302,308].

There are several recurrent cancer-associated somatic mutations, for example, activating mutations in *IDH1/2* or *EZH2*, which result in the alteration of histone marks and DNA methylation [240]. For both of these genes, drugs that have been designed to inhibit the aberrant protein are in clinical trials [240]. *EZH2* is activated by mutations in lymphomas; the use of an inhibitor has been shown to induce apoptosis or differentiation in cell lines, and this drug is now in clinical trials [309].

Beyond treatment for cancers, such medications are being used and developed to treat a wider spectrum of diseases. Resveratrol, a natural compound in red wine, which inhibits sirtuins (a family of HDACs), is being evaluated as a treatment for T2D and metabolic syndrome [310]. Valproic acid (VPA), also an HDAC inhibitor, is used to treat seizures and mood disorders [311,312], and VPA in combination with other medications has been shown to inhibit cancers in vitro [313]. The wide array of epigenetic alterations identified in human disease could present valuable targets for repurposing approved medications and developing novel agents. For example, small molecules that inhibit bromodomain and extraterminal proteins, which are important epigenetic readers of acetylated lysine residues in histones, are showing promising preclinical results in diseases ranging from cancers to organ fibrosis and osteoporosis [314–319].

5.9 ENVIRONMENTAL INFLUENCES ON THE EPIGENOME

The concept of environmental influence on the genome encompassing transgenerational epigenetic inheritance is of great importance to our current understanding of the underlying molecular mechanisms of health and disease. These environmental influences can be transgenerational, i.e., they can affect not only the individuals exposed but also future generations. The effects of

environment on our epigenome have been documented for two tragic historical events, the Dutch Hunger Winter (1944) and the Great Chinese Famine (1958–61). Studies of children born following these periods reported them to be small for gestational age at birth and to have an increased risk for schizophrenia, effects that were found to last two generations following the famine [320,321]. Children who were conceived during the Dutch famine demonstrated long-term epigenetic effects, in that changes in DNA methylation were observed six decades later in the *IGF2* gene, which encodes a growth hormone critical for normal embryonic/fetal development [322]. These data constitute the first concrete evidence that in humans the maternal diet early in pregnancy can directly affect epigenetic programming early in utero, impacting growth, metabolism, and neurodevelopment throughout life. Further, the impact is carried forward to future generations. The “developmental origins of health and disease” (DOHAD) hypothesis, pioneered by David Barker [323], has predicted, among other things, that maternal stress during pregnancy (dietary inadequacy, toxic exposures, and perhaps psychological stress) might lead to persistent epigenetic changes in the fetus, which could play a role in modulating the subsequent onset of adult cardiovascular, metabolic, and psychiatric diseases. Other maternal gestational environmental factors, such as gestational diabetes, autoimmune diseases, and infections, have also been associated with adverse pregnancy outcomes, including an increased risk for neuropsychiatric disorders [324–331]. The transgenerational effects of these exposures are not yet documented.

For many of the diseases and health outcomes discussed here, the underlying etiology has not been fully elucidated. It is likely, for example, in the case of phenotypically heterogeneous diseases such as schizophrenia and autism spectrum disorder (ASD), that there are multifactorial etiologic factors to consider. These encompass genetics and environment (often referred to as “G×E” effects), where epigenetics can be used as a powerful tool to reflect their interactions. We are only beginning to understand which environmental signals can alter epigenetic marks, most commonly DNA methylation, of specific genes/regions or the genome. Data are emerging from the exploration of other stable epigenetic marks such as miRNAs, which appear to be a promising epigenetic and molecular mark for studying the origins of brain disorders, given their stability [332–335], their abundance in the brain,

and their role in regulating neuronal plasticity and development [336,337]. We will discuss these data further in the context of exogenous, endogenous, or social environmental exposures.

5.9.1 Exogenous Exposures

Epigenetic marks can be altered by a broad range of exogenous exposures. Theoretically, it is possible to detect and quantify specific epigenetic alterations for many types of exposures. Whether the observed epigenetic dysregulation is directly linked to adverse health outcomes resulting from these exposures has yet to be confirmed, especially given that the tissues assessed in such studies are not necessarily the primary tissue impacted by the exposure. For example, environmental exposure to compounds such as cadmium [338] and arsenic [339] may predispose to epigenetic instability, aging, and cancer. Lifestyle factors such as diet, cigarette smoking, and alcohol have also been shown to alter DNA methylation in the individual exposed, as well as in the offspring, if exposure occurs during pregnancy [340,341]. Gestational exposure to cigarette smoking and alcohol has been associated with growth dysregulation and an increase in the rate of certain NDDs (e.g., fetal alcohol spectrum disorder, attention deficit hyperactivity disorder) and cancer. DNA methylation and histone methylation alterations have been associated with smoking or alcohol exposure in both human and animal models, across various tissues [240,342–344], demonstrating the utility of epigenetic marks as potential biomarkers of environmental exposures and/or disease.

Certain medications can also have direct or indirect effects on DNA methylation and histone acetylation, as demonstrated in both animal and human studies [324,340,345,346]. The commonly prescribed antiepileptic drug VPA, if taken during pregnancy, can lead to adverse birth outcomes, including teratogenic effects and impaired postnatal cognitive development [347,348]. VPA inhibits HDACs and has been shown to alter DNA methylation and DNA methyltransferase activity in various tissues in both the mother and the exposed embryo [347,349]. Other drugs with epigenetic targets have been used widely in the treatment of certain cancers and neurodegenerative diseases [350–352], but their therapeutic potential may extend beyond these diseases. Further, the detection and quantitation of epigenetic alterations may be useful as a predictive biomarker

to identify individuals who may be more amenable to specific interventions and also as an endpoint for evaluating therapeutic efficacy.

Alterations to the one-carbon metabolism (OCM) cycle can have a significant impact on the cell's ability to methylate DNA, and thus can disrupt normal epigenetic regulation [353]. Both folate, which is obtained in part through our diet, and the enzyme methyltetrahydrofolate reductase (MTHFR) are components of OCM, and are important for DNA synthesis and methylation. Mice lacking the enzyme *Mthfr* have been shown to have decreased global DNA methylation. In humans, MTHFR enzyme activity depends on an individual's genotype for the functional polymorphism MTHFR 1298A→C, which correlates positively with the level of global DNA methylation [354]. Further, humans on a folate-depleted diet demonstrate decreased global DNA methylation [355]. Conversely, in a study of adult males on hemodialysis, adding an exogenous source of folate led to an increase in both global and locus-specific DNA methylation, including *H19*, *IGF2*, and *SYDL1* [356]. Maternal folate and/or related B vitamin status has been shown to influence the cognitive outcome of offspring, although the evidence for this remains conflicting and the direct mediating role of DNA methylation remains unclear [357,358]. Last, dietary supplementation with folate and B vitamins has been clearly shown to modify tumor incidence in mouse models [359,360]. Interest in investigating the impact and therapeutic potential of folate and related vitamins in brain disorders in which epigenetic dysregulation has been implicated in the molecular etiology (e.g., ASD, neurodegenerative diseases) has increased in recent years [361–363], but is still in the early stages of research.

5.9.2 Impact of Endogenous Gestational Environment and Assisted Reproductive Technologies on Epigenetic Programming

Infertility and assisted reproductive technologies (ART) may have an impact on epigenetic reprogramming. In mice and humans, oocytes retrieved following hormonal induction or embryos studied after in vitro culture have shown DNA methylation and/or expression anomalies in several imprinted genes [364–367]. Studies of human oocytes harvested after medical hormonal induction showed loss of methylation at the maternal *MEST/PEG1* DMR on chromosome band 7q33 [368] and gain of methylation at the maternal

H19 DMR (IC1) on chromosome band 11p15.5 [367]. Increasing attention has been focused on reports of increased rates of epigenetic errors in humans following infertility/ART, particularly given that ART are administered during a sensitive developmental period when epigenetic reprogramming is occurring. The use of ART, as of this writing, accounts for ~6% of live births in North America, with continually increasing rates [369]; [370]. In particular, two rare epigenetic disorders, BWS and AS, exhibited an increased incidence in retrospective studies (odds ratios of 6–17 and 6–12, respectively) in children born following infertility/ART [371–376]. The data are especially compelling in that the increased incidence is attributable to an increase in specific epigenetic errors at chromosome regions 11p15 (BWS) and 15q11–q13 (AS), with both locations being affected by abnormal imprinting on the maternal (oocyte-derived) alleles. Furthermore, in ART-conceived AS and BWS patients, loss of maternal methylation at their respective DMRs occurs 8 and 1.9 times more often, respectively, than in individuals born from spontaneous conceptions [377–379]. Such evidence supports the hypothesis that ART-conceived children have an increased rate of epigenetic errors over that in the general population.

In humans, it is still unclear whether maternal loss of methylation observed in children post-ART is the result of the procedure itself or of an underlying infertility with oocyte abnormalities in the couple seeking ART interventions, or both. Idiopathic male infertility is also associated with aberrant methylation at both maternal and paternal alleles, suggesting that male germ cells represent another potential source for methylation defects in children conceived via ART [380,381]. Ovarian stimulation (part of subfertility/infertility treatment) has been linked to perturbed genomic imprinting at both maternally and paternally expressed genes. These data demonstrate that superovulation has dual effects during oogenesis: disruption of imprint acquisition in growing oocytes and disruption of a maternal-effect gene product subsequently required for imprint maintenance during preimplantation development [382]. The use of ART itself is associated with an increased risk for several adverse birth outcomes, including preterm labor, multiple births, and low birth weight, that, independent of ART, also confer an increased risk for adverse health outcomes, including NDDs such as ASD [383–385].

5.9.3 Social Environmental Exposures and Their Impact on Epigenetic Outcomes

In both mice and humans, maternal behavior in the neonatal period may correlate with epigenetic programming of the offspring's adult behavior [386,387]. Studies in mice have indicated that mothers showing strong nurturing behavior toward their pups, by frequently licking and grooming their offspring, produce alterations in the patterns of DNA methylation, for example, in the promoter of the *glucocorticoid receptor* gene (*GR*), in the hippocampus of their pups [388]. This area of research has been steadily growing and it will be important to follow this work over the next several years. The link between altered epigenetic marks and postnatal health outcomes is still unclear. The literature on this topic has not been well replicated, and further, there is not yet a clear causative link between the genes and biological pathways affected by altered epigenetic mechanisms and the health outcomes discussed.

As an extension of these observations on gestational exposures, early life adversity and early childhood environment are well known to be associated with long-term health outcomes, like neuropsychiatric problems and cancer [389–391]. One of the mechanisms hypothesized to be mediating these outcomes involves the perturbation of epigenetic marks such as DNA methylation. It has been suggested that specific biological pathways are particularly affected by epigenetic changes as a result of these environmental conditions. In both mouse and human models, it has been shown that early life environment, including the quality of maternal care, can influence hypothalamic–pituitary–adrenal (HPA) axis function [392,393]. Alterations in DNA methylation of the aforementioned *GR* gene, a critical component regulating the HPA axis and stress response, have been well studied in the context of the early life social environment [394]. Although results can be conflicting, this is largely attributable to the variation in experimental designs, models, tissue types, and gene regions examined. There is now growing evidence to corroborate the role of the methylation status of several important players in the HPA axis (e.g., *FKBP5* in humans, *Pomc* in mice) [395,396]. Interestingly, the involvement of the methylation status of *GR* and other components of the HPA axis has also been observed in the social environmental effects on epigenetic mechanisms

in adulthood. In particular, such effects have been reported in association with posttraumatic stress disorder and paralleled by the study of fear conditioning in animal studies [397,398].

5.10 INTERACTIONS BETWEEN THE GENOME AND THE EPIGENOME

With the exponentially increasing volume of human genetic data from copy number analyses, GWAS, and whole-exome/genome sequencing, it is crucial to consider interactions between the genome and the epigenome. These interactions can occur *in cis*, in which the local DNA sequence and haplotype can affect the pattern of epigenetic marks on a given allele, or *in trans*, in which specific chromosomal aneuploidies or point mutations can affect epigenetic patterns at various sites distributed across all the chromosomes.

Regarding *cis* interactions between the genome and the epigenome, Kerkel et al. [199] used predigestion of genomic DNA by methylation-sensitive restriction enzymes, followed by probe synthesis and hybridization to single-nucleotide polymorphism (SNP) arrays, to examine allele-specific DNA methylation (ASM) in several human tissues [199]. Their study was designed to detect new examples of imprinted genes, but instead they found numerous examples of previously unsuspected ASM at loci outside of imprinted regions. Most of these examples of nonimprinted ASM showed a strong correlation of CpG methylation patterns with local SNP genotypes, indicating *cis* regulation of this epigenetic phenomenon. That paper was quickly followed by other reports examining various types of human cells and tissues for ASM or similar phenomena of methylation quantitative trait loci (mQTL) and allele-specific transcription factor binding (ASTF). All these studies confirmed that for most genes and intergenic regions that show strong ASM, mQTL, or ASTF, the allelic asymmetry is dictated not by the parent of origin but rather by local SNPs, i.e., by the haplotype in which the epigenetic pattern is embedded (reviewed in [218]). Thus, while ASM due to parental imprinting is a potent mechanism for regulating functional gene dosage, it affects fewer genes than this more newly recognized phenomenon of haplotype-dependent ASM.

In parallel with this work, many laboratories have mapped the related phenomena of haplotype-dependent allele-specific RNA expression (ASE) and expression quantitative trait loci, which strongly affect up to

10% of human genes (reviewed in [218]; see Chapter [Multifactorial Inheritance and Complex Diseases](#)). Fine mapping of regions of haplotype-dependent ASM and ASTF under GWAS peaks is now being used to hone in on genetic variants (SNPs and indels) that have bona fide functional effects, as revealed by their ability to confer the observed physical asymmetry (ASM, ASTF) between the two alleles in heterozygotes [218]. These “post-GWAS” studies are seeking to complete the central task begun by GWAS, namely, to identify the specific functional genetic variants that confer susceptibility to common cardiovascular, metabolic, inflammatory, neuropsychiatric, and neoplastic human diseases.

Haplotype-dependent ASM, ASTF, and ASE, in contrast to imprinting, do not violate any Mendelian principles. While the allele-specific epigenetic patterns are not actually passed through the germline, in each generation these patterns are reestablished and maintained in the fetal and adult tissues under the strong *cis*-acting influence of the local DNA sequence. Therefore, for counseling purposes, each locus with haplotype-dependent epigenetic asymmetry can be thought of as inherited with the DNA sequence as a Mendelian trait.

Accumulating data have indicated that there are *trans*-acting effects of chromosomal aneuploidies on epigenetic patterns in human tissues. The chromosomal aneuploidy that causes Down syndrome (trisomy 21) has been shown to produce specific and highly recurrent changes in DNA methylation in sets of genes distributed across most of the other chromosomes. These epigenetic changes are readily detected in blood leukocytes and brain cells of individuals with this syndrome [399]. Studies are ongoing to test for this phenomenon in analogous situations, including developmental disorders with large subchromosomal DNA gains and losses, such as Williams and dup(7) syndromes [400], and cancer cells with recurrent simple chromosomal aneuploidies. In addition, there is a growing literature showing that in genetic syndromes where mutations occur in genes involved in epigenetic regulation (see [Section 5.6](#)), there are distinct downstream patterns of epigenetic marks (e.g., DNA methylation, histone marks), often referred to as “epigenetic signatures” [401–408]. These signatures have been found in various tissue types, including blood, brain, and buccal epithelium; the cross-tissue relevance of these signatures and their reflection in primary tissues of interest to each disorder are still being explored.

5.11 THE FUTURE OF EPIGENOMICS

As we have highlighted throughout this chapter, the role of epigenetic marks in translating the primary genomic sequence has now moved to the forefront of human genetics, with clear implications for our understanding of human development and disease. While a good part of what we know about epigenetics in disease came initially from cancer research, this topic has now been richly extended to nearly every common but genetically complex human disease. Large-scale epigenome projects have generated comprehensive maps of cell-type-specific epigenetic patterns, EWAS are beginning to reveal environmental effects on epigenomes, and fine mapping of allele-specific epigenetic marks is taking center stage as a post-GWAS approach for identifying functional genetic variants.

In addition, epigenomics holds tremendous potential for identifying new molecular biomarkers of disease. Moreover, given their metastable and potentially transgenerational inheritance, epigenetic markers could function in a predictive capacity, allowing for the earlier detection of diseases and health outcomes. At the same time, NGS has revealed mutations in genes that code for epigenetic readers and writers associated not only with human cancers but also with a growing number of syndromic and nonsyndromic developmental disorders. These advances from basic research are starting to translate to drug discovery and will likely also provide useful molecular endpoints for monitoring individual therapeutic responses. For all these reasons, the field of epigenetics will likely have a crucial role in the future of precision medicine.

REFERENCES

- [1] Berger SL, Kouzarides T, Shiekhatter R, Shilatifard A. An operational definition of epigenetics. *Gene Dev* 2009;23:781–783.
- [2] Kung J.T., Colognori D., Lee J.T. Long noncoding RNAs: past, present, and future. *Genetics* 2013;193:651–669.
- [3] Ong C.T., Corces V.G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014;15:234–246.
- [4] Espinoza C.A., Ren B. Mapping higher order structure of chromatin domains. *Nat Genet* 2011;43:615–616.
- [5] Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006.
- [6] Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer L.R., Pohl A., Raney B.J., Wang T., Hinrichs A.S., Zweig A.S., Fujita P.A., Learned K., Rhead B., Smith K.E., Kuhn R.M., Karolchik D., Haussler D., Kent W.J. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* 2010;38:D620–D625.
- [7] Gerstein M.B., Kundaje A., Hariharan M., Landt S.G., Yan K.K., Cheng C., Mu X.J., Khurana E., Rozowsky J., Alexander R., Min R., Alves P., Abyzov A., Addleman N., Bhardwaj N., Boyle A.P., Cayting P., Charos A., Chen D.Z., Cheng Y., Clarke D., Eastman C., Euskirchen G., Fietze S., Fu Y., Gertz J., Grubert F., Harmanci A., Jain P., Kasowski M., Lacroute P., Leng J.J., Lian J., Monahan H., O'Geen H., Ouyang Z., Partridge E.C., Patocsil D., Pauli F., Raha D., Ramirez L., Reddy T.E., Reed B., Shi M., Slifer T., Wang J., Wu L., Yang X., Yip K.Y., Zilberman-Schapira G., Batzoglou S., Sidow A., Farnham P.J., Myers R.M., Weissman S.M., Snyder M. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;489:91–100.
- [8] Rosenbloom K.R., Sloan C.A., Malladi V.S., Dreszer T.R., Learned K., Kirkup V.M., Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res* 2013;41:D56–63.
- [9] Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012;22:1798–812.
- [10] Turner BM. Histone acetylation and an epigenetic code. *Bioessays* 2000;22:836–45.
- [11] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shores N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthal KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang

- W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30.
- [12] Stunnenberg HG, International Human Epigenome C, Hirst M. The international human epigenome consortium: a blueprint for Scientific collaboration and discovery. *Cell* 2016;167:1897.
- [13] Tilghman S.M. The sins of the fathers and mothers: genomic imprinting in mammalian development. *Cell* 1999;96:185–193.
- [14] Yoder J.A., Walsh C.P., Bestor T.H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 1997;13:335–340.
- [15] Pradhan S, Bacolla A., Wells R.D., Roberts R.J. Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *J Biol Chem* 1999;274:33002–33010.
- [16] Bourc'his D., Xu G.L., Lin C.S., Bollman B., Bestor T.H. Dnmt3L and the establishment of maternal genomic imprints. *Science* 2001;294:2536–2539.
- [17] Law J.A., Jacobsen S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 2010;11:204–220.
- [18] Wu X., Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet* 2017;18:517–534.
- [19] Wolffe A.P., Matzke M.A. Epigenetics: regulation through repression. *Science* 1999;286:481–486.
- [20] Costello J.F., Plass C. Methylation matters. *J Med Genet* 2001;38:285–303.
- [21] Straussman R., Nejman D., Roberts D., Steinfeld I., Blum B., Benvenisty N., Simon I., Yakhini Z., Cedar H. Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol* 2009;16:564–571.
- [22] Jones P.A., Baylin S.B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 2002;3:415–428.
- [23] Portela A., Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol* 2010;28:1057–1068.
- [24] Cedar H., Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* 2009;10:295–304.
- [25] Epsztejn-Litman S., Feldman N., Abu-Remaileh M., Shufaro Y., Gerson A., Ueda J., Deplus R., Fuks F., Shinkai Y., Cedar H., Bergman Y. De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. *Nat Struct Mol Biol* 2008;15:1176–83.
- [26] Kondo Y. Epigenetic cross-talk between DNA methylation and histone modifications in human cancers. *Yonsei Med J* 2009;50:455–463.
- [27] Lehnertz B., Ueda Y., Derijck A.A., Braunschweig U., Perez-Burgos L., Kubicek S., Chen T., Li E., Jenuwein T., Peters A.H. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* 2003;13:1192–1200.
- [28] Ooi S.K., Qiu C., Bernstein E., Li K., Jia D., Yang Z., Erdjument-Bromage H., Tempst P., Lin S.P., Allis C.D., Cheng X., Bestor T.H. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* 2007;448:714–717.
- [29] Weber M., Hellmann I., Stadler M.B., Ramos L., Paabo S., Rebhan M., Schubeler D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 2007;39:457–466.
- [30] Kaikkonen M.U., Lam M.T., Glass C.K. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc Res* 2011;90:430–440.
- [31] Kim D.H., Saetrom P., Snove Jr. O., Rossi J.J. MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proc Natl Acad Sci U S A* 2008;105:16230–16235.
- [32] Ponting C.P., Oliver P.L., Reik W. Evolution and functions of long noncoding RNAs. *Cell* 2009;136:629–641.
- [33] Ditttrich B., Robinson W.P., Knoblauch H., Buiting K., Schmidt K., Gillissen-Kaesbach G., Horsthemke B. Molecular diagnosis of the Prader-Willi and Angelman syndromes by detection of parent-of-origin specific DNA methylation in 15q11-13. *Hum Genet* 1992;90:313–315.
- [34] Zhao J., Sun B.K., Erwin J.A., Song J.J., Lee J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 2008;322:750–6.
- [35] Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith A.C., Feil R., Fraser P. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 2008;322:1717–1720.
- [36] Tsai M.C., Manor O., Wan Y., Mosammamaparast N., Wang J.K., Lan F., Shi Y., Segal E., Chang H.Y. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010;329:689–693.
- [37] Davidovich C., Cech T.R. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA* 2015;21:2007–2022.
- [38] Khalil A.M., Guttman M., Huarte M., Garber M., Raj A., Rivea Morales D., Thomas K., Presser A., Bernstein B.E., van Oudenaarden A., Regev A., Lander E.S., Rinn J.L. Many human large intergenic noncoding

- RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* 2009;106:11667–11672.
- [39] Mackay D.J.G., Temple I.K. Human imprinting disorders: principles, practice, problems and progress. *Eur J Med Genet* 2017;60:618–626.
- [40] Arand J., Wossidlo M., Lepikhov K., Peat J.R., Reik W., Walter J. Selective impairment of methylation maintenance is the major cause of DNA methylation reprogramming in the early embryo. *Epigenet Chromatin* 2015;8:1.
- [41] Barlow D.P., Bartolomei M.S. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol* 2014;6.
- [42] Feng S., Jacobsen S.E., Reik W. Epigenetic reprogramming in plant and animal development. *Science* 2010;330:622–627.
- [43] Ito S., D'Alessio A.C., Taranova O.V., Hong K., Sowers L.C., Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* 2010;466:1129–1133.
- [44] Kriaucionis S., Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 2009;324:929–930.
- [45] Tahiliani M., Koh K.P., Shen Y., Pastor W.A., Bandukwala H., Brudno Y., Agarwal S., Iyer L.M., Liu D.R., Aravind L., Rao A. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 2009;324:930–5.
- [46] Ficz G., Branco MR., Seisenberger S., Santos F., Krueger F., Hore TA., Marques CJ., Andrews S., Reik W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 2011.
- [47] Koh K.P., Yabuuchi A., Rao S., Huang Y., Cunniff K., Nardone J., Laiho A., Tahiliani M., Sommer C.A., Mostoslavsky G., Lahesmaa R., Orkin S.H., Rodig S.J., Daley G.Q., Rao A. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* 2011;8:200–213.
- [48] Bhutani N., Brady J.J., Damian M., Sacco A., Corbel S.Y., Blau H.M. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* 2010;463:1042–1047.
- [49] Iqbal K., Jin S.G., Pfeifer G.P., Szabo P.E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci U S A* 2011;108:3642–3647.
- [50] Wossidlo M., Nakamura T., Lepikhov K., Marques C.J., Zakhartchenko V., Boiani M., Arand J., Nakano T., Reik W., Walter J. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* 2011;2:241.
- [51] Lyon M.F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 1961;190:372–373.
- [52] Graves J.A. Review: Sex chromosome evolution and the expression of sex-specific genes in the placenta. *Placenta* 2010;31(Suppl.):S27–S32.
- [53] Vicoso B., Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* 2006;7:645–653.
- [54] Graves J.A. Sex chromosome specialization and degeneration in mammals. *Cell* 2006;124:901–914.
- [55] Heard E., Avner P. Role play in X-inactivation. *Hum Mol Genet* 1994;(3 Spec No):1481–1485.
- [56] Brown CJ., Ballabio A., Rupert J.L., Lafreniere RG., Grompe M., Tonlorenzi R., Willard H.F. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* 1991;349:38–44.
- [57] Brown C.J., Hendrich B.D., Rupert J.L., Lafreniere R.G., Xing Y., Lawrence J., Willard H.F. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 1992;71:527–542.
- [58] Clemson C.M., Chow J.C., Brown C.J., Lawrence J.B. Stabilization and localization of Xist RNA are controlled by separate mechanisms and are not sufficient for X inactivation. *J Cell Biol* 1998;142:13–23.
- [59] Augui S., Filion G.J., Huart S., Nora E., Guggiari M., Maresca M., Stewart A.F., Heard E. Sensing X chromosome pairs before X inactivation via a novel X-pairing region of the Xic. *Science* 2007;318:1632–1636.
- [60] Augui S., Nora E.P., Heard E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat Rev Genet* 2011;12:429–442.
- [61] Okamoto I., Patrat C., Thepot D., Peynot N., Fauque P., Daniel N., Diabangouaya P., Wolf J.P., Renard J.P., Duranthon V., Heard E. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* 2011;472:370–374.
- [62] Morey C., Avner P. Genetics and epigenetics of the X chromosome. *Ann N Y Acad Sci* 2010;1214:E18–E33.
- [63] Okamoto I., Heard E. Lessons from comparative analysis of X-chromosome inactivation in mammals. *Chromosome Res* 2009;17:659–669.
- [64] Brown C.J., Willard H.F. The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature* 1994;368:154–156.
- [65] Donohoe M.E., Silva S.S., Pinter S.F., Xu N., Lee J.T. The pluripotency factor Oct4 interacts with Ctfcl and also controls X-chromosome pairing and counting. *Nature* 2009;460:128–132.

- [66] Guo G., Yang J., Nichols J., Hall J.S., Eyres I., Mansfield W., Smith A. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* 2009;136:1063–9.
- [67] Navarro P., Oldfield A., Legoupi J., Festuccia N., Dubois A., Attia M., Schoorlemmer J., Rougeulle C., Chambers I., Avner P. Molecular coupling of Tsix regulation and pluripotency. *Nature* 2010;468:457–460.
- [68] Lee J.T., Davidow L.S., Warshawsky D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* 1999;21:400–404.
- [69] Tian D., Sun S., Lee J.T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* 2010;143:390–403.
- [70] Xu N., Tsai C.L., Lee J.T. Transient homologous chromosome pairing marks the onset of X inactivation. *Science* 2006;311:1149–1152.
- [71] Jonkers I., Barakat T.S., Achame E.M., Monkhorst K., Kenter A., Rentmeester E., Grosveld F., Grootegoed J.A., Gribnau J. RNF12 is an X-encoded dose-dependent activator of X chromosome inactivation. *Cell* 2009;139:999–1011.
- [72] Shin J., Bossenz M., Chung Y., Ma H., Byron M., Taniguchi-Ishigaki N., Zhu X., Jiao B., Hall L.L., Green M.R., Jones S.N., Hermans-Borgmeyer I., Lawrence J.B., Bach I. Maternal Rnf12/RLIM is required for imprinted X-chromosome inactivation in mice. *Nature* 2010;467:977–981.
- [73] Chureau C., Prissette M., Bourdet A., Barbe V., Cattolico L., Jones L., Eggen A., Avner P., Duret L. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res* 2002;12:894–908.
- [74] Yang C., Chapman A.G., Kelsey A.D., Minks J., Cotton A.M., Brown C.J. X-chromosome inactivation: molecular mechanisms from the human perspective. *Hum Genet* 2011.
- [75] Migeon B.R., Chowdhury A.K., Dunston J.A., McIntosh I. Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. *Am J Hum Genet* 2001;69:951–60.
- [76] Migeon BR, Lee CH, Chowdhury AK, Carpenter H. Species differences in TSIX/Tsix reveal the roles of these genes in X-chromosome inactivation. *Am J Hum Genet* 2002;71:286–293.
- [77] Berletch J.B., Yang F., Xu J., Carrel L., Disteche C.M. Genes that escape from X inactivation. *Hum Genet* 2011;130:237–245.
- [78] Carrel L., Willard H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 2005;434:400–404.
- [79] Cotton A.M., Lam L., Affleck J.G., Wilson I.M., Penaherrera M.S., McFadden D.E., Kobor M.S., Lam W.L., Robinson W.P., Brown C.J. Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum Genet* 2011;130(2):187–201.
- [80] Nussbaum R.L., McInnes R.R., Willard H.F. Thompson & Thompson genetics in medicine. 6th ed. 2001.
- [81] Ross J., Zinn A., McCauley E. Neurodevelopmental and psychosocial aspects of Turner syndrome. *Ment Retard Dev Disabil Res Rev* 2000;6:135–141.
- [82] Rao E., Weiss B., Fukami M., Rump A., Niesler B., Mertz A., Muroya K., Binder G., Kirsch S., Winkelmann M., Nordsiek G., Heinrich U., Breuning M.H., Ranke M.B., Rosenthal A., Ogata T., Rappold G.A. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nat Genet* 1997;16:54–63.
- [83] Chiurazzi P., Schwartz C.E., Gecz J., Neri G. XLMR genes: update 2007. *Eur J Hum Genet* 2008;16:422–434.
- [84] Tarpey P.S., Smith R., Pleasance E., Whibley A., Edkins S., Hardy C., O'Meara S., Latimer C., Dicks E., Menzies A., Stephens P., Blow M., Greenman C., Xue Y., Tyler-Smith C., Thompson D., Gray K., Andrews J., Barthorpe S., Buck G., Cole J., Dunmore R., Jones D., Maddison M., Mironenko T., Turner R., Turrell K., Varian J., West S., Widaa S., Wray P., Teague J., Butler A., Jenkinson A., Jia M., Richardson D., Shepherd R., Wooster R., Tejada M.I., Martinez F., Carvill G., Goliath R., de Brouwer A.P., van Bokhoven H., Van Esch H., Chelly J., Raynaud M., Ropers HH, Abidi FE, Srivastava AK, Cox J, Luo Y, Mallya U, Moon J, Parnau J, Mohammed S, Tolmie JL, Shoubridge C, Corbett M, Gardner A, Haan E, Rujirabanjerd S, Shaw M, Vandeleur L, Fullston T, Easton DF, Boyle J, Partington M, Hackett A, Field M, Skinner C, Stevenson RE, Bobrow M, Turner G, Schwartz CE, Gecz J, Raymond FL, Futreal PA, Stratton MR. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* 2009;41:535–43.
- [85] Orstavik K.H. X chromosome inactivation in clinical practice. *Hum Genet* 2009;126:363–373.
- [86] Burn J., Povey S., Boyd Y., Munro E.A., West L., Harper K., Thomas D. Duchenne muscular dystrophy in one of monozygotic twin girls. *J Med Genet* 1986;23:494–500.
- [87] Maier E.M., Kammerer S., Muntau A.C., Wichers M., Braun A., Roscher A.A. Symptoms in carriers of adrenoleukodystrophy relate to skewed X inactivation. *Ann Neurol* 2002;52:683–688.
- [88] Chung B.H., Drmic I., Marshall C.R., Grafodatskaya D., Carter M., Fernandez B.A., Weksberg R., Roberts W., Scherer S.W. Phenotypic spectrum associated with duplication of Xp11.22-p11.23 includes autism spectrum disorder. *Eur J Med Genet* 2011;54(5):e516–520.

- [89] Giorda R., Bonaglia M.C., Beri S., Fichera M., Novara F., Magini P., Urquhart J., Sharkey F.H., Zucca C., Grasso R., Marelli S., Castiglia L., Di Benedetto D., Musumeci S.A., Vitello G.A., Failla P., Reitano S., Avola E., Bisulli F., Tinuper P., Mastrangelo M., Fiocchi I., Spaccini L., Torniero C., Fontana E., Lynch S.A., Clayton-Smith J., Black G., Jonveaux P., Leheup B., Seri M., Romano C., dalla Bernardina B., Zuffardi O. Complex segmental duplications mediate a recurrent dup(X)(p11.22-p11.23) associated with mental retardation, speech delay, and EEG anomalies in males and females. *Am J Hum Genet* 2009;85:394–400.
- [90] Twigg SR, Kan R, Babbs C, Bochukova EG, Robertson SP, Wall SA, Morriss-Kay GM, Wilkie AO. Mutations of ephrin-B1 (EFNB1), a marker of tissue boundary formation, cause craniofrontonasal syndrome. *Proc Natl Acad Sci U S A* 2004;101:8652–8657.
- [91] Ryan S.G., Chance P.F., Zou C.H., Spinner N.B., Golden J.A., Smietana S. Epilepsy and mental retardation limited to females: an X-linked dominant disorder with male sparing. *Nat Genet* 1997;17:92–95.
- [92] Trappe R., Laccone F., Cobilanschi J., Meins M., Huppke P., Hanefeld F., Engel W. MECP2 mutations in sporadic cases of Rett syndrome are almost exclusively of paternal origin. *Am J Hum Genet* 2001;68:1093–1101.
- [93] Van den Veyver I.B. Skewed X inactivation in X-linked disorders. *Semin Reprod Med* 2001;19:183–191.
- [94] Huppke P., Maier E.M., Warnke A., Brendel C., Laccone F., Gartner J. Very mild cases of Rett syndrome with skewed X inactivation. *J Med Genet* 2006;43:814–816.
- [95] Wan M., Lee S.S., Zhang X., Houwink-Manville I., Song H.R., Amir R.E., Budden S., Naidu S., Pereira J.L., Lo I.F., Zoghbi H.Y., Schanen N.C., Francke U. Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. *Am J Hum Genet* 1999;65:1520–1529.
- [96] Allen R.C., Zoghbi H.Y., Moseley A.B., Rosenblatt H.M., Belmont J.W. Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *Am J Hum Genet* 1992;51:1229–1239.
- [97] Hunter P. The silence of genes. Is genomic imprinting the software of evolution or just a battleground for gender conflict? *EMBO Rep* 2007;8:441–443.
- [98] McGrath J., Solter D. Nuclear and cytoplasmic transfer in mammalian embryos. *Dev Biol* 1986;4:37–55.
- [99] Surani M.A., Barton S.C., Norris M.L. Nuclear transplantation in the mouse: heritable differences between parental genomes after activation of the embryonic genome. *Cell* 1986;45:127–136.
- [100] Tycko B., Morison I.M. Physiological functions of imprinted genes. *J Cell Physiol* 2002;192:245–58.
- [101] Murphy SK, Jirtle RL. Imprinting evolution and the price of silence. *Bioessays* 2003;25:577–588.
- [102] Weksberg R. Imprinted genes and human disease. *Am J Med Genet C Semin Med Genet* 2010;154C:317–320.
- [103] Monk D. Genomic imprinting in the human placenta. *Am J Obstet Gynecol* 2015;213:S152–S162.
- [104] Reik W., Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* 2001;2:21–32.
- [105] Tycko B. Imprinted genes in placental growth and obstetric disorders. *Cytogenet Genome Res* 2006;113:271–278.
- [106] Peters J. The role of genomic imprinting in biology and disease: an expanding view. *Nat Rev Genet* 2014;15:517–530.
- [107] Roberts D.J., Mutter G.L. Advances in the molecular biology of gestational trophoblastic disease. *J Reprod Med* 1994;39:201–208.
- [108] Van den Veyver I.B., Al-Hussaini T.K. Biparental hydatidiform moles: a maternal effect mutation affecting imprinting in the offspring. *Hum Reprod Update* 2006;12:233–242.
- [109] Monk D., Sanchez-Delgado M., Fisher R. NLRPs, the subcortical maternal complex and genomic imprinting. *Reproduction* 2017;154:R161–R170.
- [110] Murdoch S., Djuric U., Mazhar B., Seoud M., Khan R., Quick R., Bagga R., Kircheisen R., Ao A., Ratti B., Hanash S., Rouleau G.A., Slim R. Mutations in NALP7 cause recurrent hydatidiform moles and reproductive wastage in humans. *Nat Genet* 2006;38:300–302.
- [111] Parry D.A., Logan C.V., Hayward B.E., Shires M., Landolsi H., Diggle C., Carr I., Rittore C., Touitou I., Philibert L., Fisher R.A., Fallahian M., Huntriss J.D., Picton H.M., Malik S., Taylor G.R., Johnson C.A., Bonthron D.T., Sheridan E.G. Mutations causing familial biparental hydatidiform mole implicate c6orf221 as a possible regulator of genomic imprinting in the human oocyte. *Am J Hum Genet* 2011;89:451–458.
- [112] Sanchez-Delgado M., Martin-Trujillo A., Tayama C., Vidal E., Esteller M., Iglesias-Platas I., Deo N., Barney O., Maclean K., Hata K., Nakabayashi K., Fisher R., Monk D. Absence of maternal methylation in biparental hydatidiform moles from women with NLRP7 maternal-effect mutations reveals widespread placenta-specific imprinting. *PLoS Genet* 2015;11:e1005644.
- [113] de Grouchy J. Human parthenogenesis: a fascinating single event. *Biomedicine* 1980;32:51–53.
- [114] Mutter G.L. Teratoma genetics and stem cells: a review. *Obstet Gynecol Surv* 1987;42:661–670.
- [115] Mutter G.L. Role of imprinting in abnormal human development. *Mutat Res* 1997;396:141–147.
- [116] Choufani S., Shapiro J.S., Susiarjo M., Butcher D.T., Grafodatskaya D., Lou Y., Ferreira J.C., Pinto D.,

- Scherer S.W., Shaffer L.G., Coullin P., Caniggia I., Beyene J., Slim R., Bartolomei M.S., Weksberg R. A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. *Genome Res* 2011;21:465–476.
- [117] Pettenati M.J., Haines J.L., Higgins R.R., Wappner R.S., Palmer C.G., Weaver D.D. Wiedemann-Beckwith syndrome: presentation of clinical and cytogenetic data on 22 new cases and review of the literature. *Hum Genet* 1986;74:143–154.
- [118] Weksberg R., Shuman C., Smith A.C. Beckwith-Wiedemann syndrome. *Am J Med Genet C Semin Med Genet* 2005;137C:12–23.
- [119] Weksberg R., Smith A.C., Squire J., Sadowski P. Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Hum Mol Genet* 2003;12(Spec No 1):R61–R68.
- [120] Bens S., Kolarova J., Beygo J., Buiting K., Caliebe A., Eggermann T., Gillissen-Kaesbach G., Prawitt D., Thiele-Schmitz S., Begemann M., Enklaar T., Gutwein J., Haake A., Paul U., Richter J., Soellner L., Vater I., Monk D., Horsthemke B., Ammerpohl O., Siebert R. Phenotypic spectrum and extent of DNA methylation defects associated with multilocus imprinting disturbances. *Epigenomics* 2016;8:801–816.
- [121] Eggermann T., Elbracht M., Schroder C., Reutter H., Soellner L., Spengler S., Begemann M. Congenital imprinting disorders: a novel mechanism linking seemingly unrelated disorders. *J Pediatr* 2013;163:1202–1207.
- [122] Meyer E., Lim D., Pasha S., Tee L.J., Rahman F., Yates JR, Woods CG, Reik W., Maher ER. Germline mutation in NLRP2 (NALP2) in a familial imprinting disorder (Beckwith-Wiedemann Syndrome). *PLoS Genet* 2009;5:e1000423.
- [123] Choufani S., Shuman C., Weksberg R. Molecular findings in Beckwith-Wiedemann syndrome. *Am J Med Genet C Semin Med Genet* 2013;163C:131–140.
- [124] Brzezinski J., Shuman C., Choufani S., Ray P., Stavropoulos D.J., Basran R., Steele L., Parkinson N., Grant R., Thorner P., Lorenzo A., Weksberg R. Wilms tumour in Beckwith-Wiedemann Syndrome and loss of methylation at imprinting centre 2: revisiting tumour surveillance guidelines. *Eur J Hum Genet* 2017;25:1031–1039.
- [125] Blik J., Gicquel C., Maas S., Gaston V., Le Bouc Y., Mannens M. Epigenotyping as a tool for the prediction of tumor risk and tumor type in patients with Beckwith-Wiedemann syndrome (BWS). *J Pediatr* 2004;145:796–799.
- [126] Eggermann T. Russell-Silver syndrome. *Am J Med Genet C Semin Med Genet* 2010;154C:355–364.
- [127] Gucev Z.S., Saranac L., Jancevska A., Tasic V. The degree of H19 hypomethylation in children with Silver-Russel syndrome (SRS) is not associated with the severity of phenotype and the clinical severity score (CSS). *Prilozi* 2013;34:79–83.
- [128] Cytrynbaum C., Chong K., Hannig V., Choufani S., Shuman C., Steele L., Morgan T., Scherer S.W., Stavropoulos D.J., Basran R.K., Weksberg R. Genomic imbalance in the centromeric 11p15 imprinting center in three families: further evidence of a role for IC2 as a cause of Russell-Silver syndrome. *Am J Med Genet* 2016;170:2731–2739.
- [129] Buiting K. Prader-Willi syndrome and Angelman syndrome. *Am J Med Genet C Semin Med Genet* 2010;154C:365–376.
- [130] Holm V.A., Cassidy S.B., Butler M.G., Hanchett J.M., Greenswag L.R., Whitman B.Y., Greenberg F. Prader-Willi syndrome: consensus diagnostic criteria. *Pediatrics* 1993;91:398–402.
- [131] Gunay-Aygun M., Schwartz S., Heeger S., O'Riordan MA, Cassidy SB. The changing purpose of Prader-Willi syndrome clinical diagnostic criteria and proposed revised criteria. *Pediatrics* 2001;108:E92.
- [132] Williams C.A., Beaudet A.L., Clayton-Smith J., Knoll J.H., Kyllerman M., Laan L.A., Magenis R.E., Moncla A., Schinzel A.A., Summers J.A., Wagstaff J. Angelman syndrome 2005: updated consensus for diagnostic criteria. *Am J Med Genet* 2006;140:413–418.
- [133] Hogart A., Patzel K.A., LaSalle J.M. Gender influences monoallelic expression of ATP10A in human brain. *Hum Genet* 2008;124:235–242.
- [134] Nicholls R.D., Knepper J.L. Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes. *Annu Rev Genomics Hum Genet* 2001;2:153–175.
- [135] Ohta T., Gray T.A., Rogan P.K., Buiting K., Gabriel J.M., Saitoh S., Muralidhar B., Bilienska B., Krajewska-Walasek M., Driscoll D.J., Horsthemke B., Butler M.G., Nicholls R.D. Imprinting-mutation mechanisms in Prader-Willi syndrome. *Am J Hum Genet* 1999;64:397–413.
- [136] Buiting K., Lich C., Cottrell S., Barnicoat A., Horsthemke B. A 5-kb imprinting center deletion in a family with Angelman syndrome reduces the shortest region of deletion overlap to 880 bp. *Hum Genet* 1999;105:665–666.
- [137] Kishino T., Lalande M., Wagstaff J. UBE3A/E6-AP mutations cause Angelman syndrome. *Nat Genet* 1997;15:70–73.
- [138] Matsuura T., Sutcliffe J.S., Fang P., Galjaard R.J., Jiang Y.H., Benton C.S., Rommens J.M., Beaudet A.L. De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nat Genet* 1997;15:74–77.

- [139] Clayton-Smith J., Laan L. Angelman syndrome: a review of the clinical and genetic aspects. *J Med Genet* 2003;40:87–95.
- [140] Rougeulle C., Glatt H., Lalande M. The Angelman syndrome candidate gene, UBE3A/E6-AP, is imprinted in brain. *Nat Genet* 1997;17:14–5.
- [141] Vu TH, Hoffman AR. Imprinting of the Angelman syndrome gene, UBE3A, is restricted to brain. *Nat Genet* 1997;17:12–13.
- [142] Velinov M., Jenkins E.C. PCR-based strategies for the diagnosis of Prader-Willi/Angelman syndromes. *Methods Mol Biol* 2003;217:209–216.
- [143] Bittel D.C., Kibiryeva N., Butler M.G. Methylation-specific multiplex ligation-dependent probe amplification analysis of subjects with chromosome 15 abnormalities. *Genet Test* 2007;11:467–475.
- [144] Bjornsson H.T. The Mendelian disorders of the epigenetic machinery. *Genome Res* 2015;25:1473–1481.
- [145] Fahrner J.A., Bjornsson H.T. Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states. *Annu Rev Genomics Hum Genet* 2014;15:269–293.
- [146] Kleefstra T., Schenck A., Kramer J.M., van Bokhoven H. The genetics of cognitive epigenetics. *Neuropharmacology* 2014;80:83–94.
- [147] Tatton-Brown K., Loveday C., Yost S., Clarke M., Ramsay E., Zachariou A., Elliott A., Wylie H., Ardisson A., Rittinger O., Stewart F., Temple I.K., Cole T., Mahamallie S., Seal S., Ruark E., Rahman N. Mutations in epigenetic regulation genes are a major cause of overgrowth with intellectual disability. *Am J Hum Genet* 2017;100:725–736.
- [148] Baujat G., Cormier-Daire V. Sotos syndrome. *Orphanet J Rare Dis* 2007;2:36.
- [149] Li Y., Trojer P., Xu C.F., Cheung P., Kuo A., Drury 3rd W.J., Qiao Q., Neubert T.A., Xu R.M., Gozani O., Reinberg D. The target of the NSD family of histone lysine methyltransferases depends on the nature of the substrate. *J Biol Chem* 2009;284:34283–34295.
- [150] Bannister A.J., Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* 2011;21:381–395.
- [151] Lucio-Eterovic A.K., Singh M.M., Gardner J.E., Veerapan C.S., Rice J.C., Carpenter P.B. Role for the nuclear receptor-binding SET domain protein 1 (NSD1) methyltransferase in coordinating lysine 36 methylation at histone 3 with RNA polymerase II function. *Proc Natl Acad Sci U S A* 2010;107:16952–7.
- [152] Pasillas MP, Shah M, Kamps MP. NSD1 PHD domains bind methylated H3K4 and H3K9 using interactions disrupted by point mutations in human sotos syndrome. *Hum Mutat* 2011;32:292–298.
- [153] Gibson W.T., Hood R.L., Zhan S.H., Bulman D.E., Fejes A.P., Moore R., Mungall A.J., Eyedoux P., Babul-Hirji R., An J., Marra M.A., Consortium F.C., Chitayat D., Boycott K.M., Weaver D.D., Jones S.J. Mutations in EZH2 cause Weaver syndrome. *Am J Hum Genet* 2012;90:110–118.
- [154] Tatton-Brown K., Hanks S., Ruark E., Zachariou A., Duarte Sdel V., Ramsay E., Snape K., Murray A., Perdeaux E.R., Seal S., Loveday C., Banka S., Clericuzio C., Flinter F., Magee A., McConnell V., Patton M., Raith W., Rankin J., Splitt M., Strenger V., Taylor C., Wheeler P., Temple K.I., Cole T., Childhood Overgrowth C., Douglas J., Rahman N. Germline mutations in the oncogene EZH2 cause Weaver syndrome and increased human height. *Oncotarget* 2011;2:1127–1133.
- [155] Cohen A.S., Tuysuz B., Shen Y., Bhalla S.K., Jones S.J., Gibson W.T. A novel mutation in EED associated with overgrowth. *J Hum Genet* 2015;60:339–342.
- [156] Imagawa E., Higashimoto K., Sakai Y., Numakura C., Okamoto N., Matsunaga S., Ryo A., Sato Y., Sanefuji M., Ihara K., Takada Y., Nishimura G., Saito H., Mizuguchi T., Miyatake S., Nakashima M., Miyake N., Soejima H., Matsumoto N. Mutations in genes encoding polycomb repressive complex 2 subunits cause Weaver syndrome. *Hum Mutat* 2017;38:637–648.
- [157] Ng S.B., Bigham A.W., Buckingham K.J., Hannibal M.C., McMillin M.J., Gildersleeve H.I., Beck A.E., Tabor H.K., Cooper G.M., Mefford H.C., Lee C., Turner E.H., Smith J.D., Rieder M.J., Yoshiura K., Matsumoto N., Ohta T., Niikawa N., Nickerson D.A., Bamshad M.J., Shendure J. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010;42:790–793.
- [158] Prasad R, Zhadanov AB, Sedkov Y, Bullrich F, Druck T, Rallapalli R, Yano T, Alder H, Croce CM, Huebner K, Mazo A, Canaani E. Structure and expression pattern of human ALR, a novel gene with strong homology to ALL-1 involved in acute leukemia and to *Drosophila* trithorax. *Oncogene* 1997;15:549–560.
- [159] Issaeva I., Zonis Y., Rozovskaia T., Orlovsky K., Croce C.M., Nakamura T., Mazo A., Eisenbach L., Canaani E. Knockdown of ALR (MLL2) reveals ALR target genes and leads to alterations in cell adhesion and growth. *Mol Cell Biol* 2007;27:1889–1903.
- [160] Andreu-Vieyra C.V., Chen R., Agno J.E., Glaser S., Anastassiadis K., Stewart A.F., Matzuk M.M. MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol* 2010;8.
- [161] Lederer D., Grisart B., Digilio M.C., Benoit V., Crespin M., Ghariani S.C., Maystadt I., Dallapiccola B., Verellen-Dumoulin C. Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *Am J Hum Genet* 2012;90:119–124.

- [162] Abidi F, Holloway L., Moore C.A., Weaver D.D., Simensen R.J., Stevenson R.E., Rogers R.C., Schwartz C.E. Novel human pathological mutations. Gene symbol: JARID1C. Disease: mental retardation, X-linked. *Hum Genet* 2009;125:345.
- [163] Abidi F.E., Holloway L., Moore C.A., Weaver D.D., Simensen R.J., Stevenson R.E., Rogers R.C., Schwartz C.E. Mutations in JARID1C are associated with X-linked mental retardation, short stature and hyperreflexia. *J Med Genet* 2008;45:787–793.
- [164] Jensen L.R., Amende M., Gurok U., Moser B., Gimmel V., Tzschach A., Janecke A.R., Tariverdian G., Chelly J., Fryns J.P., Van Esch H., Kleefstra T., Hamel B., Moraine C., Gecz J., Turner G., Reinhardt R., Kalscheuer V.M., Ropers H.H., Lenzner S. Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. *Am J Hum Genet* 2005;76:227–36.
- [165] Rujirabanjerd S, Nelson J, Tarpey PS, Hackett A, Edkins S, Raymond FL, Schwartz CE, Turner G, Iwase S, Shi Y, Futreal PA, Stratton MR, Gecz J. Identification and characterization of two novel JARID1C mutations: suggestion of an emerging genotype-phenotype correlation. *Eur J Hum Genet* 2009;18(3):330–5.
- [166] Santos C., Rodriguez-Revena L., Madrigal I., Badenas C., Pineda M., Mila M. A novel mutation in JARID1C gene associated with mental retardation. *Eur J Hum Genet* 2006;14:583–586.
- [167] Tzschach A., Lenzner S., Moser B., Reinhardt R., Chelly J., Fryns J.P., Kleefstra T., Raynaud M., Turner G., Ropers H.H., Kuss A., Jensen L.R. Novel JARID1C/SMCX mutations in patients with X-linked mental retardation. *Hum Mutat* 2006;27:389.
- [168] Baker L.A., Allis C.D., Wang G.G. PHD fingers in human diseases: disorders arising from misinterpreting epigenetic marks. *Mutat Res* 2008;647:3–12.
- [169] Huang F, Chandrasekharan M.B., Chen Y.C., Bhaskara S., Hiebert S.W., Sun Z.W. The JmjN domain of Jhd2 is important for its protein stability, and the plant homeodomain (PHD) finger mediates its chromatin association independent of H3K4 methylation. *J Biol Chem* 2010;285:24548–24561.
- [170] Christensen J., Agger K., Cloos P.A., Pasini D., Rose S., Sennels L., Rappsilber J., Hansen K.H., Salcini A.E., Helin K. RBP2 belongs to a family of demethylases, specific for tri- and dimethylated lysine 4 on histone 3. *Cell* 2007;128:1063–1076.
- [171] Iwase S., Lan F., Bayliss P., de la Torre-Ubieta L., Huarte M., Qi H.H., Whetstone J.R., Bonni A., Roberts T.M., Shi Y. The X-linked mental retardation gene SMCX/JARID1C defines a family of histone H3 lysine 4 demethylases. *Cell* 2007;128:1077–1088.
- [172] Tahiliani M., Mei P., Fang R., Leonor T., Rutenberg M., Shimizu F., Li J., Rao A., Shi Y. The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. *Nature* 2007;447:601–605.
- [173] Amir R.E., Van den Veyver I.B., Wan M., Tran C.Q., Francke U., Zoghbi H.Y. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* 1999;23:185–8.
- [174] Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, Bird A. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* 1992;69:905–914.
- [175] Jones P.L., Veenstra G.J., Wade P.A., Vermaak D., Kass S.U., Landsberger N., Strouboulis J., Wolffe A.P. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* 1998;19:187–191.
- [176] Nan X., Campoy F.J., Bird A. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* 1997;88:471–481.
- [177] Nan X., Ng H.H., Johnson C.A., Laherty C.D., Turner B.M., Eisenman R.N., Bird A. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* 1998;393:386–389.
- [178] Kokura K., Kaul S.C., Wadhwa R., Nomura T., Khan M.M., Shinagawa T., Yasukawa T., Colmenares C., Ishii S. The Ski protein family is required for MeCP2-mediated transcriptional repression. *J Biol Chem* 2001;276:34115–34121.
- [179] Chahrour M., Jung S.Y., Shaw C., Zhou X., Wong S.T., Qin J., Zoghbi H.Y. MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science* 2008;320:1224–1229.
- [180] Chahrour M., Zoghbi H.Y. The story of Rett syndrome: from clinic to neurobiology. *Neuron* 2007;56:422–437.
- [181] Chang Q., Khare G., Dani V., Nelson S., Jaenisch R. The disease progression of Mecp2 mutant mice is affected by the level of BDNF expression. *Neuron* 2006;49:341–348.
- [182] Kline D.D., Ogier M., Kunze D.L., Katz D.M. Exogenous brain-derived neurotrophic factor rescues synaptic dysfunction in Mecp2-null mice. *J Neurosci* 2010;30:5303–5310.
- [183] Skene P.J., Illingworth R.S., Webb S., Kerr A.R., James K.D., Turner D.J., Andrews R, Bird AP. Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol Cell* 2010;37:457–68.
- [184] Ramocki MB, Zoghbi H.Y. Failure of neuronal homeostasis results in common neuropsychiatric phenotypes. *Nature* 2008;455:912–918.

- [185] Lopez A.J., Wood M.A. Role of nucleosome remodeling in neurodevelopmental and intellectual disability disorders. *Front Behav Neurosci* 2015;9:100.
- [186] Lalani S.R., Safiullah A.M., Fernbach S.D., Harutyunyan K.G., Thaller C., Peterson L.E., McPherson J.D., Gibbs R.A., White L.D., Hefner M., Davenport S.L., Graham J.M., Bacino C.A., Glass N.L., Towbin J.A., Craigen W.J., Neish S.R., Lin A.E., Belmont J.W. Spectrum of CHD7 mutations in 110 individuals with CHARGE syndrome and genotype-phenotype correlation. *Am J Hum Genet* 2006;78:303–314.
- [187] Visser L.E., van Ravenswaaij C.M., Admiraal R., Hurst J.A., de Vries B.B., Janssen I.M., van der Vliet W.A., Huys E.H., de Jong P.J., Hamel B.C., Schoenmakers E.F., Brunner H.G., Veltman J.A., van Kessel A.G. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat Genet* 2004;36:955–957.
- [188] Zentner G.E., Layman W.S., Martin D.M., Scacheri P.C. Molecular and phenotypic aspects of CHD7 mutation in CHARGE syndrome. *Am J Med Genet* 2010;152A:674–686.
- [189] Schnetz M.P., Bartels C.F., Shastri K., Balasubramanian D., Zentner G.E., Balaji R., Zhang X., Song L., Wang Z., Laframboise T., Crawford G.E., Scacheri P.C. Genomic distribution of CHD7 on chromatin tracks H3K4 methylation patterns. *Genome Res* 2009;19:590–601.
- [190] Gibbons R.J., Wada T., Fisher C.A., Malik N., Mitson M.J., Steensma D.P., Fryer A., Goudie D.R., Krantz I.D., Traeger-Synodinos J. Mutations in the chromatin-associated protein ATRX. *Hum Mutat* 2008;29:796–802.
- [191] Hargreaves D.C., Crabtree G.R. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell Res* 2011;21:396–420.
- [192] Dhayalan A., Tamas R., Bock I., Tattermusch A., Dimitrova E., Kudithipudi S., Ragozin S., Jeltsch A. The ATRX-ADD domain binds to H3 tail peptides and reads the combined methylation state of K4 and K9. *Hum Mol Genet* 2011;20:2195–203.
- [193] Gibbons R.J., McDowell T.L., Raman S., O'Rourke D.M., Garrick D., Ayyub H., Higgs D.R. Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nat Genet* 2000;24:368–371.
- [194] McDowell T.L., Gibbons R.J., Sutherland H., O'Rourke D.M., Bickmore W.A., Pombo A., Turley H., Gatter K., Picketts D.J., Buckle V.J., Chapman L., Rhodes D., Higgs D.R. Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes. *Proc Natl Acad Sci U S A* 1999;96:13983–13988.
- [195] Xue Y., Gibbons R., Yan Z., Yang D., McDowell T.L., Sechi S., Qin J., Zhou S., Higgs D., Wang W. The ATRX syndrome protein forms a chromatin-remodeling complex with Daxx and localizes in promyelocytic leukemia nuclear bodies. *Proc Natl Acad Sci U S A* 2003;100:10635–10640.
- [196] Law M.J., Lower K.M., Voon H.P., Hughes J.R., Garrick D., Viprakasit V., Mitson M., De Gobbi M., Marra M., Morris A., Abbott A., Wilder S.P., Taylor S., Santos G.M., Cross J., Ayyub H., Jones S., Ragoussis J., Rhodes D., Dunham I., Higgs D.R., Gibbons R.J. ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell* 2010;143:367–378.
- [197] Marcaud L., Reynaud C.A., Therwath A., Scherrer K. Modification of the methylation pattern in the vicinity of the chicken globin genes in avian erythroblastosis virus transformed cells. *Nucleic Acids Res* 1981;9:1841–1851.
- [198] Grafodatskaya D., Choufani S., Ferreira J.C., Butcher D.T., Lou Y., Zhao C., Scherer S.W., Weksberg R. EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. *Genomics* 2010;95:73–83.
- [199] Kerkel K., Spadola A., Yuan E., Kosek J., Jiang L., Hod E., Li K., Murty V.V., Schupf N., Vilain E., Morris M., Haghighi F., Tycko B. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* 2008;40:904–908.
- [200] Frommer M., McDonald L.E., Millar D.S., Collis C.M., Watt F., Grigg G.W., Molloy P.L., Paul C.L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 1992;89:1827–1831.
- [201] Do C., Lang C.F., Lin J., Darbary H., Krupka I., Gaba A., Petukhova L., Vonsattel J.P., Gallagher M.P., Goland R.S., Clynes R.A., Dwork A., Kral J.G., Monk C., Christiano A.M., Tycko B. Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. *Am J Hum Genet* 2016;98:934–955.
- [202] Mendioroz M., Do C., Jiang X., Liu C., Darbary H.K., Lang C.F., Lin J., Thomas A., Abu-Amro S., Stanier P., Temkin A., Yale A., Liu M.M., Li Y., Salas M., Kerkel K., Capone G., Silverman W., Yu Y.E., Moore G., Wegiel J., Tycko B. Trans effects of chromosome aneuploidies on DNA methylation patterns in human Down syndrome and mouse models. *Genome Biol* 2015;16:263.
- [203] Flanagan J.M. Epigenome-wide association studies (EWAS): past, present, and future. *Methods Mol Biol* 2015;1238:51–63.

- [204] Lill C.M., Bertram L. Probing the epigenome by EWAS: a new era in brain disease research. *Mov Disord* 2015;30:197.
- [205] Jeong M., Guzman A.G., Goodell M.A. Genome-wide analysis of DNA methylation in hematopoietic cells: DNA methylation analysis by WGBS. *Methods Mol Biol* 2017;1633:137–149.
- [206] Hahn M.A., Li A.X., Wu X., Pfeifer G.P. Single base resolution analysis of 5-methylcytosine and 5-hydroxymethylcytosine by RRBS and TAB-RRBS. *Methods Mol Biol* 2015;1238:273–287.
- [207] Seifuddin F., Wand G., Cox O., Pirooznia M., Moody L., Yang X., Tai J., Boersma G., Tamashiro K., Zandi P., Lee R. Genome-wide Methyl-Seq analysis of blood-brain targets of glucocorticoid exposure. *Epigenetics* 2017;12:637–52.
- [208] Yang Y., Scott S.A. DNA methylation profiling using long-read single molecule real-time bisulfite sequencing (SMRT-BS). *Methods Mol Biol* 2017;1654:125–134.
- [209] Schmidt D., Wilson M.D., Spyrou C., Brown G.D., Hadfield J., Odom D.T. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* 2009;48:240–248.
- [210] Bernstein B.E., Stamatoyannopoulos J.A., Costello J.F., Ren B., Milosavljevic A., Meissner A., Kellis M., Marra M.A., Beaudet A.L., Ecker J.R., Farnham P.J., Hirst M., Lander E.S., Mikkelsen T.S., Thomson J.A. The NIH Roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010;28:1045–1048.
- [211] Dekker J., Rippe K., Dekker M., Kleckner N. Capturing chromosome conformation. *Science* 2002;295:1306–1311.
- [212] Fullwood M.J., Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem* 2009;107:30–39.
- [213] Casper J., Zweig A.S., Villarreal C., Tyner C., Speir M.L., Rosenbloom K.R., Raney B.J., Lee C.M., Lee B.T., Karolchik D., Hinrichs A.S., Haeussler M., Guruvadoo L., Navarro Gonzalez J., Gibson D., Fiddes I.T., Eisenhart C., Diekhans M., Clawson H., Barber G.P., Armstrong J., Haussler D., Kuhn R.M., Kent W.J. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* 2018;46:D762–D769.
- [214] Abugessaisa I., Kasukawa T., Kawaji H. Genome annotation. *Methods Mol Biol* 2017;1525:107–121.
- [215] Abugessaisa I., Noguchi S., Hasegawa A., Harshbarger J., Kondo A., Lizio M., Severin J., Carninci P., Kawaji H., Kasukawa T. FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci Data* 2017;4:170107.
- [216] Liu C., Marioni R.E., Hedman A.K., Pfeiffer L., Tsai P.C., Reynolds L.M., Just A.C., Duan Q., Boer C.G., Tanaka T., Elks C.E., Aslibekyan S., Brody J.A., Kuhnel B., Herder C., Almli L.M., Zhi D., Wang Y., Huan T., Yao C., Mendelson M.M., Joehanes R., Liang L., Love S.A., Guan W., Shah S., McRae A.F., Kretschmer A., Prokisch H., Strauch K., Peters A., Visscher P.M., Wray N.R., Guo X., Wiggins K.L., Smith A.K., Binder E.B., Ressler K.J., Irvin M.R., Absher D.M., Hernandez D., Ferrucci L., Bandinelli S., Lohman K., Ding J., Trevisi L., Gustafsson S., Sandling J.H., Stolk L., Uitterlinden A.G., Yet I., Castillo-Fernandez J.E., Spector T.D., Schwartz J.D., Vokonas P., Lind L., Li Y., Fornage M., Arnett D.K., Wareham N.J., Sotoodehnia N., Ong K.K., van Meurs J.B., Conneely K.N., Baccarelli A.A., Deary I.J., Bell J.T., North K.E., Liu Y., Waldenberger M., London S.J., Ingelsson E., Levy D. A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry* 2016;23:422–433.
- [217] Philibert R., Erwin C. A review of epigenetic markers of tobacco and alcohol consumption. *Behav Sci Law* 2015;33:675–690.
- [218] Do C., Shearer A., Suzuki M., Terry M.B., Gelernter J., Grealley J.M., Tycko B. Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biol* 2017;18:120.
- [219] Hatchwell E., Grealley J.M. The potential role of epigenomic dysregulation in complex human disease. *Trends Genet* 2007;23:588–595.
- [220] Michels K.B., Binder A.M., Dedeurwaerder S., Epstein C.B., Grealley J.M., Gut I., Houseman E.A., Izzi B., Kelsey K.T., Meissner A., Milosavljevic A., Siegmund K.D., Bock C., Irizarry R.A. Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods* 2013;10:949–955.
- [221] Demerath E.W., Guan W., Grove M.L., Aslibekyan S., Mendelson M., Zhou Y.H., Hedman A.K., Sandling J.K., Li L.A., Irvin M.R., Zhi D., Deloukas P., Liang L., Liu C., Bressler J., Spector T.D., North K., Li Y., Absher D.M., Levy D., Arnett D.K., Fornage M., Pankow J.S., Boerwinkle E. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet* 2015;24:4464–4479.
- [222] Dick K.J., Nelson C.P., Tsaprouni L., Sandling J.K., Aïssi D., Wahl S., Meduri E., Morange P.-E., Gagnon F., Grallert H., Waldenberger M., Peters A., Erdmann J., Hengstenberg C., Cambien F., Goodall A.H., Ouwehand W.H., Schunkert H., Thompson J.R., Spector T.D., Gieger C., Trégouët D.-A., Deloukas P., Samani N.J. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 2014;383:1990–1998.
- [223] Ronn T., Volkov P., Gillberg L., Kokosar M., Perfilyev A., Jacobsen A.L., Jorgensen S.W., Brons C., Jansson P.A., Eriksson K.F., Pedersen O., Hansen T., Groop L.,

- Stener-Victorin E., Vaag A., Nilsson E., Ling C. Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Hum Mol Genet* 2015;24:3792–3813.
- [224] Hedman A.K., Mendelson M.M., Marioni R.E., Gustafsson S., Joehanes R., Irvin M.R., Zhi D., Sandling J.K., Yao C., Liu C., Liang L., Huan T., McRae A.F., Demissie S., Shah S., Starr J.M., Cupples L.A., Deloukas P., Spector T.D., Sundstrom J., Krauss R.M., Arnett D.K., Deary I.J., Lind L., Levy D., Ingelsson E. Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for results from genome-wide association studies. *Circ Cardiovasc Genet* 2017;10.
- [225] Li J., Zhu X., Yu K., Jiang H., Zhang Y., Deng S., Cheng L., Liu X., Zhong J., Zhang X., He M., Chen W., Yuan J., Gao M., Bai Y., Han X., Liu B., Luo X., Mei W., He X., Sun S., Zhang L., Zeng H., Sun H., Liu C., Guo Y., Zhang B., Zhang Z., Huang J., Pan A., Yuan Y., Angileri F., Ming B., Zheng F., Zeng Q., Mao X., Peng Y., Mao Y., Ye P., Wang Q.K., Qi L., Hu F.B., Liang L., Wu T. Genome-wide analysis of DNA methylation and acute coronary syndrome. *Circ Res* 2017;120(11):1754–1767.
- [226] Zhang J., Liu Z., Umukoro P.E., Cavallari J.M., Fang S.C., Weisskopf M.G., Lin X., Mittleman M.A., Christiani D.C. An epigenome-wide association analysis of cardiac autonomic responses among a population of welders. *Epigenetics* 2017;12:71–76.
- [227] De Jager P.L., Srivastava G., Lunnon K., Burgess J., Schalkwyk L.C., Yu L., Eaton M.L., Keenan B.T., Ernst J., McCabe C., Tang A., Raj T., Replogle J., Brodeur W., Gabriel S., Chai H.S., Younkin C., Younkin S.G., Zou F., Szyf M., Epstein C.B., Schneider J.A., Bernstein B.E., Meisner A., Ertekin-Taner N., Chibnik L.B., Kellis M., Mill J., Bennett D.A. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci* 2014;17:1156–63.
- [228] Lunnon K., Smith R., Hannon E., De Jager P.L., Srivastava G., Volta M., Troakes C., Al-Sarraj S., Burrage J., Macdonald R., Condliffe D., Harries L.W., Katsel P., Haroutunian V., Kaminsky Z., Joachim C., Powell J., Lovestone S., Bennett D.A., Schalkwyk L.C., Mill J. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nat Neurosci* 2014;17:1164–1170.
- [229] Watson C.T., Roussos P., Garg P., Ho D.J., Azam N., Katsel P.L., Haroutunian V., Sharp A.J. Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease. *Genome Med* 2016;8:5.
- [230] Li Yim A.Y., Duijvis N.W., Zhao J., de Jonge W.J., D'Haens G.R., Mannens M.M., Mul A.N., Te Velde A.A., Henneman P. Peripheral blood methylation profiling of female Crohn's disease patients. *Clin Epigenet* 2016;8:65.
- [231] Zimmermann M.T., Oberg A.L., Grill D.E., Ovsyanikova I.G., Haralambieva I.H., Kennedy R.B., Poland G.A. System-wide associations between DNA-methylation, gene expression, and humoral immune response to influenza vaccination. *PLoS One* 2016;11:e0152034.
- [232] Hannon E., Dempster E., Viana J., Burrage J., Smith A.R., Macdonald R., St Clair D., Mustard C., Breen G., Therman S., Kaprio J., Touloupoulou T., Hulshoff Pol H.E., Bohlken M.M., Kahn R.S., Nenadic I., Hultman C.M., Murray R.M., Collier D.A., Bass N., Gurling H., McQuillin A., Schalkwyk L., Mill J. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol* 2016;17:176.
- [233] Zhang R., Miao Q., Wang C., Zhao R., Li W., Haile CN, Hao W., Zhang X.Y. Genome-wide DNA methylation analysis in alcohol dependence. *Addiction Biol* 2013;18:392–403.
- [234] Feinberg A.P., Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 1983;301:89–92.
- [235] Feinberg A.P., Koldobskiy M.A., Gondor A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet* 2016;17:284–299.
- [236] Feinberg A.P., Tycko B. The history of cancer epigenetics. *Nat Rev Cancer* 2004;4:143–153.
- [237] Calin G.A., Croce C.M. MicroRNA signatures in human cancers. *Nature reviews. Cancer* 2006;6:857–866.
- [238] Davalos V., Esteller M. MicroRNAs and cancer epigenetics: a macroevolution. *Curr Opin Oncol* 2010;22:35–45.
- [239] Kondo Y., Shinjo K., Katsushima K. Long non-coding RNAs as an epigenetic regulator in human cancers. *Cancer Sci* 2017;108:1927–1933.
- [240] Portales-Casamar E., Lussier A.A., Jones M.J., MacIsaac J.L., Edgar R.D., Mah S.M., Barhdadi A., Provost S., Lemieux-Perreault L.P., Cynader M.S., Chudley A.E., Dube M.P., Reynolds J.N., Pavlidis P., Kobor M.S. DNA methylation signature of human fetal alcohol spectrum disorder. *Epigenet Chromatin* 2016;9:25.
- [241] Werner R.J., Kelly A.D., Issa J.J. Epigenetics and precision oncology. *Cancer J* 2017;23:262–269.
- [242] Rocco J.W., Sidransky D. p16(MTS-1/CDKN2/INK4a) in cancer progression. *Exp Cell Res* 2001;264:42–55.
- [243] Brooks J.D., Weinstein M., Lin X., Sun Y., Pin S.S., Bova G.S., Epstein J.I., Isaacs W.B., Nelson W.G. CG island methylation changes near the GSTP1 gene in prostatic intraepithelial neoplasia. *Cancer Epidemiol Biomarkers Prev* 1998;7:531–536.

- [244] Dobrovic A., Simpfendorfer D. Methylation of the BRCA1 gene in sporadic breast cancer. *Cancer Res* 1997;57:3347–3350.
- [245] Esteller M., Levine R., Baylin S.B., Ellenson L.H., Herman J.G. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. *Oncogene* 1998;17:2413–2417.
- [246] Herman J.G., Latif F., Weng Y., Lerman M.I., Zbar B., Liu S., Samid D., Duan D.S., Gnarr J.R., Linehan W.M., et al. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci U S A* 1994;91:9700–4.
- [247] Mancini DN, Rodenhiser DI, Ainsworth PJ, O'Malley FP, Singh SM, Xing W, Archer TK. CpG methylation within the 5' regulatory region of the BRCA1 gene is tumor specific and includes a putative CREB binding site. *Oncogene* 1998;16:1161–1169.
- [248] Veigl M.L., Kasturi L., Olechnowicz J., Ma A.H., Lutterbaugh J.D., Periyasamy S., Li G.M., Drummond J., Modrich P.L., Sedwick W.D., Markowitz S.D. Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. *Proc Natl Acad Sci U S A* 1998;95:8698–8702.
- [249] Gossage L., Eisen T. Alterations in VHL as potential biomarkers in renal-cell carcinoma. *Nat Rev Clin Oncol* 2010;7:277–288.
- [250] Belinsky S.A., Palmisano W.A., Gilliland F.D., Crooks L.A., Divine K.K., Winters S.A., Grimes M.J., Harms H.J., Tellez C.S., Smith T.M., Moots P.P., Lechner J.F., Stidley C.A., Crowell R.E. Aberrant promoter methylation in bronchial epithelium and sputum from current and former smokers. *Cancer Res* 2002;62:2370–2377.
- [251] Klump B., Hsieh C.J., Holzmann K., Gregor M., Porschen R. Hypermethylation of the CDKN2/p16 promoter during neoplastic progression in Barrett's esophagus. *Gastroenterology* 1998;115:1381–1386.
- [252] Wong D.J., Barrett M.T., Stoger R., Emond M.J., Reid B.J. p16INK4a promoter is hypermethylated at a high frequency in esophageal adenocarcinomas. *Cancer Res* 1997;57:2619–2622.
- [253] Enokida H., Shiina H., Urakami S., Igawa M., Ogishima T., Li L.C., Kawahara M., Nakagawa M., Kane C.J., Carroll P.R., Dahiya R. Multigene methylation analysis for detection and staging of prostate cancer. *Clin Cancer Res* 2005;11:6582–6588.
- [254] Hoque M.O., Begum S., Topaloglu O., Chatterjee A., Rosenbaum E., Van Criekinge W., Westra WH, Schoenberg M, Zahurak M, Goodman SN, Sidransky D. Quantitation of promoter methylation of multiple genes in urine DNA and bladder cancer detection. *J Natl Cancer Inst* 2006;98:996–1004.
- [255] Belinsky SA, Liechty KC, Gentry FD, Wolf HJ, Rogers J, Vu K, Haney J., Kennedy T.C., Hirsch F.R., Miller Y., Franklin W.A., Herman J.G., Baylin S.B., Bunn P.A., Byers T. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. *Cancer Res* 2006;66:3338–3344.
- [256] Chen W.D., Han Z.J., Skoletsky J., Olson J., Sah J., Myeroff L., Platzer P., Lu S., Dawson D., Willis J., Pretlow T.P., Lutterbaugh J., Kasturi L., Willson J.K., Rao J.S., Shuber A., Markowitz S.D. Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. *J Natl Cancer Inst* 2005;97:1124–1132.
- [257] Rasmussen S.L., Krarup H.B., Sunesen K.G., Pedersen I.S., Madsen P.H., Thorlacius-Ussing O. Hypermethylated DNA as a biomarker for colorectal cancer: a systematic review. *Colorectal Dis* 2016;18:549–561.
- [258] Lapeyre J.N., Becker F.F. 5-Methylcytosine content of nuclear DNA during chemical hepatocarcinogenesis and in carcinomas which result. *Biochem Biophys Res Commun* 1979;87:698–705.
- [259] Feinberg A.P., Gehrke C.W., Kuo K.C., Ehrlich M. Reduced genomic 5-methylcytosine content in human colonic neoplasia. *Cancer Res* 1988;48:1159–1161.
- [260] Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene* 2002;21:5400–5413.
- [261] Chen R.Z., Pettersson U., Beard C., Jackson-Grusby L., Jaenisch R. DNA hypomethylation leads to elevated mutation rates. *Nature* 1998;395:89–93.
- [262] Madakashira B.P., Sadler K.C. DNA methylation, nuclear organization, and cancer. *Front Genet* 2017;8:76.
- [263] Mudbhary R., Hoshida Y., Chernyavskaya Y., Jacob V., Villanueva A., Fiel M.I., Chen X., Kojima K., Thung S, Bronson RT, Lachenmayer A, Revill K, Alsinet C, Sachidanandam R, Desai A, SenBanerjee S, Ukomadu C, Llovet JM, Sadler KC. UHRF1 overexpression drives DNA hypomethylation and hepatocellular carcinoma. *Cancer Cell* 2014;25:196–209.
- [264] Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CPE, van Dijk CM, Tollenaar R.A.E.M., Van Den Berg D., Laird P.W. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet* 2012;44:40–46.
- [265] Timp W., Bravo H.C., McDonald O.G., Goggins M., Umbricht C., Zeiger M., Feinberg A.P., Irazarry R.A. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med* 2014;6:61.
- [266] Gao F, Shi L., Russin J., Zeng L., Chang X., He S., Chen T.C., Giannotta S.L., Weisenberger D.J., Zada G., Mack W.J., Wang K. DNA methylation in the

- malignant transformation of meningiomas. *PLoS One* 2013;8:e54114.
- [267] Kim Y.I., Giuliano A., Hatch K.D., Schneider A., Nour M.A., Dallal G.E., Selhub J., Mason J.B. Global DNA hypomethylation increases progressively in cervical dysplasia and carcinoma. *Cancer* 1994;74:893–899.
- [268] Baer C., Claus R., Plass C. Genome-wide epigenetic regulation of miRNAs in cancer. *Cancer Res* 2013;73:473–477.
- [269] Fraga M.F., Ballestar E., Villar-Garea A., Boix-Chornet M., Espada J., Schotta G., Bonaldi T., Haydon C., Ropero S., Petrie K., Iyer N.G., Perez-Rosado A., Calvo E., Lopez J.A., Cano A., Calasanz M.J., Colomer D., Piris M.A., Ahn N., Imhof A., Caldas C., Jenuwein T., Esteller M. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet* 2005;37:391–400.
- [270] Halkidou K., Gaughan L., Cook S., Leung H.Y., Neal D.E., Robson C.N. Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer. *Prostate* 2004;59:177–189.
- [271] Song J., Noh J.H., Lee J.H., Eun J.W., Ahn Y.M., Kim S.Y., Lee S.H., Park W.S., Yoo N.J., Lee J.Y., Nam S.W. Increased expression of histone deacetylase 2 is found in human gastric cancer. *APMIS* 2005;113:264–8.
- [272] Yang X.J. The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Res* 2004;32:959–76.
- [273] Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M., Jackman S., Krzywinski M., Scott D.W., Trinh D.L., Tamura-Wells J., Li S., Firme M.R., Rogic S., Griffith M., Chan S., Yakovenko O., Meyer I.M., Zhao E.Y., Smailus D., Moksa M., Chittaranjan S., Rimsza L., Brooks-Wilson A., Spinelli J.J., Ben-Neriah S., Meissner B., Woolcock B., Boyle M., McDonald H., Tam A., Zhao Y., Delaney A., Zeng T., Tse K., Butterfield Y., Birol I., Holt R., Schein J., Horsman D.E., Moore R., Jones S.J., Connors J.M., Hirst M., Gascoyne R.D., Marra M.A. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* 2011;476:298–303.
- [274] Nguyen C.T., Weisenberger D.J., Velicescu M., Gonzales F.A., Lin J.C., Liang G., Jones P.A. Histone H3-lysine 9 methylation is associated with aberrant gene silencing in cancer cells and is rapidly reversed by 5-aza-2'-deoxycytidine. *Cancer Res* 2002;62:6456–6461.
- [275] Kleer C.G., Cao Q., Varambally S., Shen R., Ota I., Tomlins S.A., Ghosh D., Sewalt R.G., Otte A.P., Hayes D.F., Sabel M.S., Livant D., Weiss S.J., Rubin M.A., Chinnaiyan A.M. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A* 2003;100:11606–11611.
- [276] Varambally S., Dhanasekaran S.M., Zhou M., Barrette T.R., Kumar-Sinha C., Sanda M.G., Ghosh D., Pienta K.J., Sewalt R.G., Otte A.P., Rubin M.A., Chinnaiyan A.M. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 2002;419:624–629.
- [277] Ernst T., Chase A.J., Score J., Hidalgo-Curtis C.E., Bryant C., Jones A.V., Waghorn K., Zoi K., Ross F.M., Reiter A., Hochhaus A., Drexler H.G., Duncombe A., Cervantes F., Oscier D., Boultonwood J., Grand F.H., Cross N.C. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat Genet* 2010;42:722–6.
- [278] Ntziachristos P., Tsirogas A., Van Vlierberghe P., Nedjic J., Trimarchi T., Flaherty M.S., Ferres-Marco D., da Ros V., Tang Z., Siegle J., Asp P., Hadler M., Rigo I., De Keersmaecker K., Patel J., Huynh T., Utro F., Poglio S., Samon J.B., Paietta E., Racevskis J., Rowe J.M., Rabadan R., Levine R.L., Brown S., Pflumio F., Dominguez M., Ferrando A., Aifantis I. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med* 2012;18:298–301.
- [279] Dalglish G.L., Furge K., Greenman C., Chen L., Bignell G., Butler A., Davies H., Edkins S., Hardy C., Latimer C., Teague J., Andrews J., Barthorpe S., Beare D., Buck G., Campbell P.J., Forbes S., Jia M., Jones D., Knott H., Kok C.Y., Lau K.W., Leroy C., Lin M.L., McBride D.J., Maddison M., Maguire S., McLay K., Menzies A., Mironenko T., Mulderrig L., Mudie L., O'Meara S., Pleasance E., Rajasingham A., Shepherd R., Smith R., Stebbings L., Stephens P., Tang G., Tarpey P.S., Turrell K., Dykema K.J., Khoo S.K., Petillo D., Wonderegern B., Anema J., Kahnoski R.J., Teh B.T., Stratton M.R., Futreal P.A. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* 2010;463:360–363.
- [280] Pasqualucci L., Dominguez-Sola D., Chiarenza A., Fabbri G., Grunn A., Trifonov V., Kasper L.H., Lerach S., Tang H., Ma J., Rossi D., Chadburn A., Murty V.V., Mullighan C.G., Gaidano G., Rabadan R., Brindle P.K., Dalla-Favera R. Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* 2011;471:189–195.
- [281] Varela I., Tarpey P., Raine K., Huang D., Ong C.K., Stephens P., Davies H., Jones D., Lin M.L., Teague J., Bignell G., Butler A., Cho J., Dalglish G.L., Galapathige D., Greenman C., Hardy C., Jia M., Latimer C., Lau K.W., Marshall J., McLaren S., Menzies A., Mudie L., Stebbings L., Largaespada D.A., Wessels L.F.,

- Richard S., Kahnoski R.J., Anema J., Tuveson D.A., Perez-Mancera P.A., Mustonen V., Fischer A., Adams D.J., Rust A., Chan-on W., Subimerb C., Dykema K., Furge K., Campbell P.J., Teh B.T., Stratton M.R., Futreal P.A. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 2011;469:539–542.
- [282] Versteeg L., Sevenet N., Lange J., Rousseau-Merck M.F., Ambros P., Handgretinger R., Aurias A., Delattre O. Truncating mutations of hSNF5/IN1 in aggressive paediatric cancer. *Nature* 1998;394:203–6.
- [283] Papillon-Cavanagh S, Lu C, Gayden T, Mikael LG, Bechet D, Karamboulas C, Ailles L, Karamchandani J, Marchione DM, Garcia BA, Weinreb I, Goldstein D, Lewis PW, Dancu O.M., Dhaliwal S., Stecho W., Howlett C.J., Mymryk J.S., Barrett J.W., Nichols A.C., Allis C.D., Majewski J., Jabado N. Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas. *Nat Genet* 2017;49:180–185.
- [284] Wu G., Broniscer A., McEachron T.A., Lu C., Paugh B.S., Becksfort J., Qu C., Ding L., Huether R., Parker M., Zhang J., Gajjar A., Dyer M.A., Mullighan C.G., Gilbertson R.J., Mardis E.R., Wilson R.K., Downing J.R., Ellison D.W., Zhang J., Baker S.J., St Jude Children's Research Hospital-Washington University Pediatric Cancer Genome, P. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet* 2012;44:251–253.
- [285] Nikbakht H., Panditharatna E., Mikael L.G., Li R., Gayden T., Osmond M., Ho C.Y., Kambhampati M., Hwang E.I., Faury D., Siu A., Papillon-Cavanagh S., Bechet D., Ligon K.L., Ellezam B., Ingram W.J., Stinson C., Moore A.S., Warren K.E., Karamchandani J., Packer R.J., Jabado N., Majewski J., Nazarian J. Spatial and temporal homogeneity of driver mutations in diffuse intrinsic pontine glioma. *Nat Commun* 2016;7:11185.
- [286] Lu J., Getz G., Miska E.A., Alvarez-Saavedra E., Lamb J., Peck D., Sweet-Cordero A., Ebert B.L., Mak R.H., Ferrando A.A., Downing J.R., Jacks T., Horvitz H.R., Golub T.R. MicroRNA expression profiles classify human cancers. *Nature* 2005;435:834–838.
- [287] Chan J.A., Krichevsky A.M., Kosik K.S. MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer Res* 2005;65:6029–6033.
- [288] Zhang B., Pan X., Cobb G.P., Anderson T.A. microRNAs as oncogenes and tumor suppressors. *Dev Biol* 2007;302:1–12.
- [289] Deng S, Calin GA, Croce CM, Coukos G, Zhang L. Mechanisms of microRNA deregulation in human cancer. *Cell Cycle* 2008;7:2643–2646.
- [290] Lujambio A., Calin G.A., Villanueva A., Ropero S., Sanchez-Cespedes M., Blanco D., Montuenga L.M., Rossi S., Nicoloso M.S., Faller W.J., Gallagher W.M., Eccles S.A., Croce C.M., Esteller M. A microRNA DNA methylation signature for human cancer metastasis. *Proc Natl Acad Sci USA* 2008;105:13556–13561.
- [291] Toyota M., Suzuki H., Sasaki Y., Maruyama R., Imai K., Shinomura Y., Tokino T. Epigenetic silencing of microRNA-34b/c and B-cell translocation gene 4 is associated with CpG island methylation in colorectal cancer. *Cancer Res* 2008;68:4123–4132.
- [292] Palamarchuk A., Efanov A., Nazaryan N., Santanam U., Alder H., Rassenti L., Kipps T., Croce C.M., Pekarsky Y. 13q14 deletions in CLL involve cooperating tumor suppressors. *Blood* 2010;115:3916–3922.
- [293] Hosseini N., Aghapour M., Duijff P.H.G., Baradaran B. Treating cancer with microRNA replacement therapy: a literature review. *J Cell Physiol* 2018.
- [294] Esquela-Kerscher A., Trang P., Wiggins J.F., Patrawala L., Cheng A., Ford L., Weidhaas J.B., Brown D., Bader A.G., Slack F.J. The let-7 microRNA reduces tumor growth in mouse models of lung cancer. *Cell Cycle* 2008;7:759–764.
- [295] Morlando M., Fatica A. Alteration of epigenetic regulation by long noncoding RNAs in cancer. *Int J Mol Sci* 2018;19.
- [296] Schmitt A.M., Chang H.Y. Long noncoding RNAs in cancer pathways. *Cancer Cell* 2016;29:452–463.
- [297] Gupta R.A., Shah N., Wang K.C., Kim J., Horlings H.M., Wong D.J., Tsai M.C., Hung T., Argani P., Rinn J.L., Wang Y., Brzoska P., Kong B., Li R., West R.B., van de Vijver M.J., Sukumar S., Chang H.Y. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010;464:1071–1076.
- [298] Bhan A., Mandal S.S. LncRNA HOTAIR: a master regulator of chromatin dynamics and cancer. *Biochim Biophys Acta Rev Cancer* 2015;1856:151–164.
- [299] Ji P., Diederichs S., Wang W., Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H, Muller-Tidow C. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 2003;22:8031–41.
- [300] Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M, Gross M, Zornig M, MacLeod A.R., Spector D.L., Diederichs S. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* 2013;73:1180–1189.
- [301] Huang N.S., Chi Y.Y., Xue J.Y., Liu M.Y., Huang S., Mo M., Zhou S.L., Wu J. Long non-coding RNA metastasis associated in lung adenocarcinoma transcript 1 (MALAT1) interacts with estrogen receptor and predicted poor survival in breast cancer. *Oncotarget* 2016;7:37957–37965.

- [302] Rodriguez-Paredes M., Esteller M. A combined epigenetic therapy equals the efficacy of conventional chemotherapy in refractory advanced non-small cell lung cancer. *Cancer Discov* 2011;1:557–559.
- [303] Da Costa E.M., McInnes G., Beaudry A., Raynal N.J. DNA methylation-targeted drugs. *Cancer J* 2017;23:270–276.
- [304] Jones P.A., Issa J.P., Baylin S. Targeting the cancer epigenome for therapy. *Nat Rev Genet* 2016;17:630–641.
- [305] Plimack E.R., Kantarjian H.M., Issa J.P. Decitabine and its role in the treatment of hematopoietic malignancies. *Leuk Lymphoma* 2007;48:1472–1481.
- [306] Cortez C.C., Jones P.A. Chromatin, cancer and drug therapies. *Mutat Res* 2008;647:44–51.
- [307] McClure J.J., Li X., Chou C.J. Chapter six – advances and challenges of HDAC inhibitors in cancer therapeutics. *Adv Cancer Res*. 2018;138:183–211.
- [308] Azad N., Zahnow C.A., Rudin C.M., Baylin S.B. The future of epigenetic therapy in solid tumours—lessons from the past. *Nat Rev Clin Oncol* 2013;10:256–266.
- [309] McCabe M.T., Ott H.M., Ganji G., Korenchuk S., Thompson C., Van Aller G.S., Liu Y., Graves AP, Della Pietra 3rd A, Diaz E, LaFrance LV, Mellinger M, Duquette C, Tian X, Kruger RG, McHugh CF, Brandt M, Miller WH, Dhanak D, Verma SK, Tummino PJ, Creasy CL. EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* 2012;492:108–12.
- [310] Gunjan A, Singh RK. Epigenetic therapy: targeting histones and their modifications in human disease. *Future Med Chem* 2010;2:543–548.
- [311] Bond D.J., Lam R.W., Yatham L.N. Divalproex sodium versus placebo in the treatment of acute bipolar depression: a systematic review and meta-analysis. *J Affect Disord* 2010;124:228–234.
- [312] Brodie M.J. Antiepileptic drug therapy the story so far. *Seizure* 2010;19:650–655.
- [313] Chateauvieux S., Morceau F., Dicato M., Diederich M. Molecular and therapeutic potential and toxicity of valproic acid. *J Biomed Biotechnol* 2010;2010.
- [314] Baud'huin M., Lamoureux F., Jacques C., Rodriguez Calleja L., Quillard T., Charrier C., Amiaud J., Berreur M., Brounais-LeRoyer B., Owen R., Reilly G.C., Bradner J.E., Heymann D., Ory B. Inhibition of BET proteins and epigenetic signaling as a potential treatment for osteoporosis. *Bone* 2017;94:10–21.
- [315] da Motta L.L., Ledaki I., Purshouse K., Haider S., De Bastiani M.A., Baban D., Morotti M., Steers G., Wigfield S., Bridges E., Li J.L., Knapp S., Ebner D., Klamt F., Harris A.L., McIntyre A. The BET inhibitor JQ1 selectively impairs tumour response to hypoxia and downregulates CA9 and angiogenesis in triple negative breast cancer. *Oncogene* 2017;36:122–132.
- [316] Ember S.W., Lambert Q.T., Berndt N., Gunawan S., Ayaz M., Tauro M., Zhu J.Y., Cranfill P.J., Greninger P., Lynch C.C., Benes C.H., Lawrence H.R., Reuther G.W., Lawrence N.J., Schonbrunn E. Potent dual BET bromodomain-kinase inhibitors as value-added multitargeted chemical probes and cancer therapeutics. *Mol Cancer Ther* 2017;16:1054–1067.
- [317] Saenz D.T., Fiskus W., Manshoury T., Rajapakshe K., Krieger S., Sun B., Mill C.P., DiNardo C., Pemmaraju N., Kadia T., Parmar S., Sharma S., Coarfa C., Qiu P., Verstovsek S., Bhalla K.N. BET protein bromodomain inhibitor-based combinations are highly active against post-myeloproliferative neoplasm secondary AML cells. *Leukemia* 2017;31:678–87.
- [318] Suarez-Alvarez B, Morgado-Pascual JL, Rayego-Mateos S., Rodriguez R.M., Rodrigues-Diez R., Cannata-Ortiz P., Sanz A.B., Egidio J., Tharaux P.L., Ortiz A., Lopez-Larrea C., Ruiz-Ortega M. Inhibition of bromodomain and extraterminal domain family proteins Ameliorates experimental renal damage. *J Am Soc Nephrol* 2017;28:504–519.
- [319] Zhou B., Mu J., Gong Y., Lu C., Zhao Y., He T., Qin Z. Brd4 inhibition attenuates unilateral ureteral obstruction-induced fibrosis by blocking TGF-beta-mediated Nox4 expression. *Redox Biol* 2017;11:390–402.
- [320] Hoek H.W., Brown A.S., Susser E. The Dutch famine and schizophrenia spectrum disorders. *Soc Psychiatry Psychiatr Epidemiol* 1998;33:373–379.
- [321] St Clair D., Xu M., Wang P., Yu Y., Fang Y., Zhang F., Zheng X., Gu N., Feng G., Sham P., He L. Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959–1961. *J Am Med Assoc* 2005;294:557–562.
- [322] Heijmans B.T., Tobi E.W., Stein A.D., Putter H., Blauw G.J., Susser E.S., Slagboom P.E., Lumey L.H. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci Unit States Am* 2008;105:17046–17049.
- [323] Barker D.J., Eriksson J.G., Forsen T., Osmond C. Fetal origins of adult disease: strength of effects and biological basis. *Int J Epidemiol* 2002;31:1235–1239.
- [324] Brown A.S. Epidemiologic studies of exposure to prenatal infection and risk of schizophrenia and autism. *Dev Neurobiol* 2012;72:1272–1276.
- [325] Chen S.W., Zhong X.S., Jiang L.N., Zheng X.Y., Xiong Y.Q., Ma S.J., Qiu M., Huo S.T., Ge J., Chen Q. Maternal autoimmune diseases and the risk of autism spectrum disorders in offspring: a systematic review and meta-analysis. *Behav Brain Res* 2016;296:61–69.
- [326] Fang S.Y., Wang S., Huang N., Yeh H.H., Chen C.Y. Prenatal infection and autism spectrum disorders in childhood: a population-based case-control study in Taiwan. *Paediatr Perinat Epidemiol* 2015;29:307–16.

- [327] Krakowiak P, Walker CK, Tancredi D, Hertz-Picciotto I, Van de Water J. Autism-specific maternal anti-fetal brain autoantibodies are associated with metabolic conditions. *Autism Res* 2016;10(1):89–98.
- [328] Lee B.K., Magnusson C., Gardner R.M., Blomstrom A., Newschaffer C.J., Burstyn I., Karlsson H., Dalman C. Maternal hospitalization with infection during pregnancy and risk of autism spectrum disorders. *Brain Behav Immun* 2015;44:100–105.
- [329] Li M., Fallin M.D., Riley A., Landa R., Walker S.O., Silverstein M., Caruso D., Pearson C., Kiang S., Dahm J.L., Hong X., Wang G., Wang M.C., Zuckerman B., Wang X. The association of maternal obesity and diabetes with autism and other developmental disabilities. *Pediatrics* 2016;137:e20152206.
- [330] Li Y.M., Ou J.J., Liu L., Zhang D., Zhao J.P., Tang S.Y. Association between maternal obesity and autism spectrum disorder in offspring: a meta-analysis. *J Autism Dev Disord* 2016;46:95–102.
- [331] Xiang A.H., Wang X., Martinez M.P., Walthall J.C., Curry E.S., Page K., Buchanan T.A., Coleman K.J., Getahun D. Association of maternal diabetes with autism in offspring. *J Am Med Assoc* 2015;313:1425–1434.
- [332] Fregeac J., Colleaux L., Nguyen L.S. The emerging roles of MicroRNAs in autism spectrum disorders. *Neurosci Biobehav Rev* 2016;71:729–738.
- [333] Mundalil Vasu M., Anitha A., Thanseem I., Suzuki K., Yamada K., Takahashi T., Wakuda T., Iwata K., Tsujii M., Sugiyama T., Mori N. Serum microRNA profiles in children with autism. *Mol Autism* 2014;5:40.
- [334] Parikshak N.N., Swarup V., Belgard T.G., Irimia M., Ramaswami G., Gandal M.J., Hartl C., Leppa V., Ubieta L.T., Huang J., Lowe J.K., Blencowe B.J., Horvath S., Geschwind D.H. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 2016;540:423–427.
- [335] Wang Y., Zhao X., Ju W., Flory M., Zhong J., Jiang S., Wang P., Dong X., Tao X., Chen Q., Shen C., Zhong M., Yu Y., Brown W.T., Zhong N. Genome-wide differential expression of synaptic long noncoding RNAs in autism spectrum disorder. *Transl Psychiatry* 2015;5:e660.
- [336] Cortez MA, Calin GA. MicroRNA identification in plasma and serum: a new tool to diagnose and monitor diseases. *Expert Opin Biol Ther* 2009;9:703–711.
- [337] Kosik K.S. The neuronal microRNA system. *Nat Rev Neurosci* 2006;7:911–920.
- [338] Poirier L.A., Vlasova T.I. The prospective role of abnormal methyl metabolism in cadmium toxicity. *Environ Health Perspect* 2002;110(Suppl. 5):793–795.
- [339] Pilsner J.R., Liu X., Ahsan H., Ilievski V., Slavkovich V., Levy D., Factor-Litvak P., Graziano J.H., Gamble M.V. Genomic methylation of peripheral blood leukocyte DNA: influences of arsenic and folate in Bangladeshi adults. *Am J Clin Nutr* 2007;86:1179–1186.
- [340] Feil R., Fraga M.F. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* 2012;13:97–109.
- [341] Vaiserman A. Epidemiologic evidence for association between adverse environmental exposures in early life and epigenetic variation: a potential link to disease susceptibility? *Clin Epigenet* 2015;7:96.
- [342] Chater-Diehl E.J., Laufer B.I., Castellani C.A., Alberry B.L., Singh S.M. Alteration of gene expression, DNA methylation, and histone methylation in free radical scavenging networks in adult mouse Hippocampus following fetal alcohol exposure. *PLoS One* 2016;11:e0154836.
- [343] Joubert B., London S. Epigenomics and maternal smoking, with Bonnie Joubert and Stephanie London by Ashley Ahearn. *Environ Health Perspect* 2012;120.
- [344] Lee K.W., Richmond R., Hu P., French L., Shin J., Bourdon C., Reischl E., Waldenberger M., Zeilinger S., Gaunt T., McArdle W., Ring S., Woodward G., Bouchard L., Gaudet D., Smith G.D., Relton C., Paus T., Pausova Z. Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. *Environ Health Perspect* 2015;123:193–9.
- [345] Houtepen LC, van Bergen AH, Vinkers CH, Boks MP. DNA methylation signatures of mood stabilizers and antipsychotics in bipolar disorder. *Epigenomics* 2016;8:197–208.
- [346] Zahnow CA, Topper M, Stone M, Murray-Stewart T, Li H, Baylin SB, Casero Jr RA. Inhibitors of DNA methylation, histone deacetylation, and histone demethylation: a perfect combination for cancer therapy. *Adv Cancer Res* 2016;130:55–111.
- [347] Roullet FI, Lai JK, Foster JA. In utero exposure to valproic acid and autism—a current review of clinical and animal studies. *Neurotoxicol Teratol* 2013;36:47–56.
- [348] Tomson T, Battino D, Perucca E. The remarkable story of valproic acid. *Lancet Neurol* 2016;15:141.
- [349] Detich N, Theberge J, Szyf M. Promoter-specific activation and demethylation by MBD2/demethylase. *J Biol Chem* 2002;277:35791–4.
- [350] Abel T, Zukin RS. Epigenetic targets of HDAC inhibition in neurodegenerative and psychiatric disorders. *Curr Opin Pharmacol* 2008;8:57–64.
- [351] Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. *Nat Biotechnol* 2010;28:1069–78.

- [352] Schneider A, Chatterjee S, Bousiges O, Selvi BR, Swaminathan A, Cassel R, Blanc F, Kundu TK, Boutillier AL. Acetyltransferases (HATs) as targets for neurological therapeutics. *Neurotherapeutics* 2013;10:568–88.
- [353] Friso S, Udali S, De Santis D, Choi SW. One-carbon metabolism and epigenetics. *Mol Aspects Med* 2017;54:28–36.
- [354] Friso S, Girelli D, Trabetti E, Olivieri O, Guarini P, Pignatti PF, Corrocher R, Choi SW. The MTHFR 1298A>C polymorphism and genomic DNA methylation in human lymphocytes. *Cancer Epidemiol Biomarkers Prev* 2005;14:938–43.
- [355] Rampersaud GC, Kauwell GP, Hutson AD, Cerda JJ, Bailey LB. Genomic DNA methylation decreases in response to moderate folate depletion in elderly women. *Am J Clin Nutr* 2000;72:998–1003.
- [356] Ingrosso D, Cimmino A, Perna AF, Masella L, De Santo NG, De Bonis ML, Vacca M, D'Esposito M, D'Urso M, Galletti P, Zappia V. Folate treatment and unbalanced methylation and changes of allelic expression induced by hyperhomocysteinaemia in patients with uraemia. *Lancet* 2003;361:1693–9.
- [357] Caramaschi D, Sharp GC, Nohr EA, Berryman K, Lewis SJ, Davey Smith G, Relton CL. Exploring a causal role of DNA methylation in the relationship between maternal vitamin B12 during pregnancy and child's IQ at age 8, cognitive performance and educational attainment: a two-step Mendelian randomization study. *Hum Mol Genet* 2017;26:3001–13.
- [358] Irwin RE, Pentieva K, Cassidy T, Lees-Murdock DJ, McLaughlin M, Prasad G, McNulty H, Walsh CP. The interplay between DNA methylation, folate and neurocognitive development. *Epigenomics* 2016;8:863–79.
- [359] Song J, Medline A, Mason JB, Gallinger S, Kim YI. Effects of dietary folate on intestinal tumorigenesis in the *apcMin* mouse. *Cancer Res* 2000;60:5434–40.
- [360] Trasler J, Deng L, Melnyk S, Pogribny I, Hiou-Tim F, Sibani S, Oakes C, Li E, James SJ, Rozen R. Impact of *Dnmt1* deficiency, with and without low folate diets, on tumor numbers and DNA methylation in *Min* mice. *Carcinogenesis* 2003;24:39–45.
- [361] Bjork M, Riedel B, Spigset O, Veiby G, Kolstad E, Daltveit AK, Gilhus NE. Association of folic acid supplementation during pregnancy with the risk of autistic traits in children exposed to antiepileptic drugs in utero. *JAMA Neurol* 2018;75:160–8.
- [362] Castro K, Klein Lda S, Baronio D, Gottfried C, Riesgo R, Perry IS. Folic acid and autism: what do we know? *Nutr Neurosci* 2016;19:310–7.
- [363] McGarel C, Pentieva K, Strain JJ, McNulty H. Emerging roles for folate and related B-vitamins in brain health across the lifecycle. *Proc Nutr Soc* 2015;74:46–55.
- [364] Li T, Vu TH, Ulaner GA, Littman E, Ling JQ, Chen HL, Hu JF, Behr B, Giudice L, Hoffman AR. IVF results in de novo DNA methylation and histone methylation at an *Igf2-H19* imprinting epigenetic switch. *Mol Hum Reprod* 2005;11:631–40.
- [365] Mann MR, Chung YG, Nolen LD, Verona RI, Latham KE, Bartolomei MS. Disruption of imprinted gene methylation and expression in cloned preimplantation stage mouse embryos. *Biol Reprod* 2003;69:902–14.
- [366] Mann MR, Lee SS, Doherty AS, Verona RI, Nolen LD, Schultz RM, Bartolomei MS. Selective loss of imprinting in the placenta following preimplantation development in culture. *Development* 2004;131:3727–35.
- [367] Sato A, Otsu E, Negishi H, Utsunomiya T, Arima T. Aberrant DNA methylation of imprinted loci in super-ovulated oocytes. *Hum Reprod* 2007;22:26–35.
- [368] Kobayashi H, Sato A, Otsu E, Hiura H, Tomatsu C, Utsunomiya T, Sasaki H, Yaegashi N, Arima T. Aberrant DNA methylation of imprinted loci in sperm from oligospermic patients. *Hum Mol Genet* 2007;16:2542–51.
- [369] CDC, Centers for Disease Control and Prevention.
- [370] Schieve LA, Devine O, Boyle CA, Petrini JR, Warner L. Estimation of the contribution of non-assisted reproductive technology ovulation stimulation fertility treatments to US singleton and multiple births. *Am J Epidemiol* 2009;170:1396–407.
- [371] Chang AS, Moley KH, Wangler M, Feinberg AP, Debaun MR. Association between Beckwith-Wiedemann syndrome and assisted reproductive technology: a case series of 19 patients. *Fertil Steril* 2005;83:349–54.
- [372] Cox GF, Burger J, Lip V, Mau UA, Sperling K, Wu BL, Horsthemke B. Intracytoplasmic sperm injection may increase the risk of imprinting defects. *Am J Hum Genet* 2002;71:162–4.
- [373] Halliday J, Oke K, Breheny S, Algar E, D JA. Beckwith-Wiedemann syndrome and IVF: a case-control study. *Am J Hum Genet* 2004;75:526–8.
- [374] Maher ER, Brueton LA, Bowdin SC, Luharia A, Cooper W, Cole TR, Macdonald F, Sampson JR, Barratt CL, Reik W, Hawkins MM. Beckwith-Wiedemann syndrome and assisted reproduction technology (ART). *J Med Genet* 2003;40:62–4.
- [375] Orstavik KH, Eiklid K, van der Hagen CB, Spetalen S, Kierulf K, Skjeldal O, Buiting K. Another case of imprinting defect in a girl with Angelman syndrome who was conceived by intracytoplasmic semen injection. *Am J Hum Genet* 2003;72:218–9.
- [376] Sutcliffe AG, Peters CJ, Bowdin S, Temple K, Reardon W, Wilson L, Clayton-Smith J, Brueton LA, Bannister W, Maher ER. Assisted reproductive therapies and imprinting disorders—a preliminary British survey. *Hum Reprod* 2006;21:1009–11.

- [377] DeBaun MR, Niemitz EL, Feinberg AP. Association of in vitro fertilization with Beckwith-Wiedemann syndrome and epigenetic alterations of LIT1 and H19. *Am J Hum Genet* 2003;72:156–60.
- [378] Gicquel C, Gaston V, Mandelbaum J, Siffroi JP, Flahault A, Le Bouc Y. In vitro fertilization may increase the risk of Beckwith-Wiedemann syndrome related to the abnormal imprinting of the KCN1OT gene. *Am J Hum Genet* 2003;72:1338–41.
- [379] Ludwig M, Katalinic A, Gross S, Sutcliffe A, Varon R, Horsthemke B. Increased prevalence of imprinting defects in patients with Angelman syndrome born to subfertile couples. *J Med Genet* 2005;42:289–91.
- [380] Kobayashi H, Hiura H, John RM, Sato A, Otsu E, Kobayashi N, Suzuki R, Suzuki F, Hayashi C, Utsunomiya T, Yaegashi N, Arima T. DNA methylation errors at imprinted loci after assisted conception originate in the parental sperm. *Eur J Hum Genet* 2009;17:1582–91.
- [381] Poplinski A, Tuttelmann F, Kanber D, Horsthemke B, Gromoll J. Idiopathic male infertility is strongly associated with aberrant methylation of MEST and IGF2/H19 ICR1. *Int J Androl* 2009.
- [382] Market-Velker BA, Zhang L, Magri LS, Bonvissuto AC, Mann MR. Dual effects of superovulation: loss of maternal and paternal imprinted methylation in a dose-dependent manner. *Hum Mol Genet* 2010;19:36–51.
- [383] Gardener H, Spiegelman D, Buka SL. Perinatal and neonatal risk factors for autism: a comprehensive meta-analysis. *Pediatrics* 2011;128:344–55.
- [384] Grafodatskaya D, Cytrynbaum C, Weksberg R. The health risks of ART. *EMBO Rep* 2013;14:129–35.
- [385] Savage T, Peek J, Hofman PL, Cutfield WS. Childhood outcomes of assisted reproductive technology. *Hum Reprod* 2011;26:2392–400.
- [386] Szyf M, Weaver I, Meaney M. Maternal care, the epigenome and phenotypic differences in behavior. *Reprod Toxicol* 2007;24:9–19.
- [387] Weaver IC, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ. Epigenetic programming by maternal behavior. *Nat Neurosci* 2004;7:847–54.
- [388] Fish EW, Shahrokh D, Bagot R, Caldji C, Bredy T, Szyf M, Meaney MJ. Epigenetic programming of stress responses through variations in maternal care. *Ann N Y Acad Sci* 2004;1036:167–80.
- [389] Heim C, Binder EB. Current research trends in early life stress and depression: review of human studies on sensitive periods, gene-environment interactions, and epigenetics. *Exp Neurol* 2012;233:102–11.
- [390] Holman DM, Ports KA, Buchanan ND, Hawkins NA, Merrick MT, Metzler M, Trivers KF. The association between adverse childhood experiences and risk of cancer in adulthood: a systematic review of the literature. *Pediatrics* 2016;138:S81–91.
- [391] McGowan PO, Szyf M. The epigenetics of social adversity in early life: implications for mental health outcomes. *Neurobiol Dis* 2010;39:66–72.
- [392] Levine AB, Lockwood CJ, Chitkara U, Berkowitz RL. Maternal renal artery Doppler velocimetry in normotensive pregnancies and pregnancies complicated by hypertensive disorders. *Obstet Gynecol* 1992;79:264–7.
- [393] Meaney MJ. Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. *Annu Rev Neurosci* 2001;24:1161–92.
- [394] Turecki G, Meaney MJ. Effects of the social environment and stress on glucocorticoid receptor gene methylation: a systematic review. *Biol Psychiatr* 2016;79:87–96.
- [395] Klengel T, Mehta D, Anacker C, Rex-Haffner M, Pruessner JC, Pariante CM, Pace TW, Mercer KB, Mayberg HS, Bradley B, Nemeroff CB, Holsboer F, Heim CM, Ressler KJ, Rein T, Binder EB. Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nat Neurosci* 2013;16:33–41.
- [396] Wu Y, Patchev AV, Daniel G, Almeida OF, Spengler D. Early-life stress reduces DNA methylation of the Pomc gene in male mice. *Endocrinology* 2014;155:1751–62.
- [397] Vukojevic V, Kolassa IT, Fastenrath M, Gschwind L, Spalek K, Milnik A, Heck A, Vogler C, Wilker S, Demougin P, Peter F, Atucha E, Stetak A, Roozendaal B, Elbert T, Papassotiropoulos A, de Quervain DJ. Epigenetic modification of the glucocorticoid receptor gene is linked to traumatic memory and post-traumatic stress disorder risk in genocide survivors. *J Neurosci* 2014;34:10274–84.
- [398] Zannas AS, Provençal N, Binder EB. Epigenetics of post-traumatic stress disorder: current evidence, challenges, and future directions. *Biol Psychiatr* 2015;78:327–35.
- [399] Do C, Xing Z, Yu YE, Tycko B. Trans-acting epigenetic effects of chromosomal aneuploidies: lessons from Down syndrome and mouse models. *Epigenomics* 2017;9:189–207.
- [400] Strong E, Butcher DT, Singhania R, Mervis CB, Morris CA, De Carvalho D, Weksberg R, Osborne LR. Symmetrical dose-dependent DNA-methylation profiles in children with deletion or duplication of 7q11.23. *Am J Hum Genet* 2015;97:216–27.
- [401] Aref-Eshghi E, Rodenhiser DI, Schenkel LC, Lin H, Skinner C, Ainsworth P, Pare G, Hood RL, Bulman DE, Kernohan KD, Boycott KM, Campeau PM, Schwartz C, Sadikovic B. Genomic DNA methylation signatures enable concurrent diagnosis and clinical genetic variant classification in neurodevelopmental syndromes. *Am J Hum Genet* 2018;102:156–74.

- [402] Aref-Eshghi E, Schenkel LC, Lin H, Skinner C, Ainsworth P, Pare G, Rodenhiser D, Schwartz C, Sadikovic B. The defining DNA methylation signature of Kabuki syndrome enables functional assessment of genetic variants of unknown clinical significance. *Epigenetics* 2017;12:923–33.
- [403] Aref-Eshghi E, Schenkel LC, Lin H, Skinner C, Ainsworth P, Pare G, Siu V, Rodenhiser D, Schwartz C, Sadikovic B. Clinical validation of a genome-wide DNA methylation assay for molecular diagnosis of imprinting disorders. *J Mol Diagn* 2017;19:848–56.
- [404] Butcher DT, Cytrynbaum C, Turinsky AL, Siu MT, Inbar-Feigenberg M, Mendoza-Londono R, Chitayat D, Walker S, Machado J, Caluseriu O, Dupuis L, Grafo-datskaya D, Reardon W, Gilbert-Dussardier B, Verloes A, Bilan F, Milunsky JM, Basran R, Papsin B, Stockley TL, Scherer SW, Choufani S, Brudno M, Weksberg R. CHARGE and Kabuki syndromes: gene-specific DNA methylation signatures identify epigenetic mechanisms linking these clinically overlapping conditions. *Am J Hum Genet* 2017;100:773–88.
- [405] Hood RL, Schenkel LC, Nikkel SM, Ainsworth PJ, Pare G, Boycott KM, Bulman DE, Sadikovic B. The defining DNA methylation signature of Floating-Harbor Syndrome. *Sci Rep* 2016;6:38803.
- [406] Kernohan KD, Cigana Schenkel L, Huang L, Smith A, Pare G, Ainsworth P, Boycott KM, Warman-Chardon J, Sadikovic B. Identification of a methylation profile for DNMT1-associated autosomal dominant cerebellar ataxia, deafness, and narcolepsy. *Clin Epigenet* 2016;8:91.
- [407] Schenkel LC, Aref-Eshghi E, Skinner C, Ainsworth P, Lin H, Pare G, Rodenhiser DI, Schwartz C, Sadikovic B. Peripheral blood epi-signature of Claes-Jensen syndrome enables sensitive and specific identification of patients and healthy carriers with pathogenic mutations in KDM5C. *Clin Epigenet* 2018;10:21.
- [408] Schenkel LC, Kernohan KD, McBride A, Reina D, Hodge A, Ainsworth PJ, Rodenhiser DI, Pare G, Berube NG, Skinner C, Boycott KM, Schwartz C, Sadikovic B. Identification of epigenetic signature associated with alpha thalassemia/mental retardation X-linked syndrome. *Epigenet Chromatin* 2017;10:10.

Websites

- GeneImprint database: <http://www.geneimprint.com/>.
 ENCODE at the UCSC Genome Browser: <https://genome.ucsc.edu/encode/>.
 NIH Roadmap Epigenomics Project: <http://www.roadmap-epigenomics.org/>.
 International Human Epigenome Consortium (IHEC): <http://ihec-epigenomes.org/>.
 FANTOM: fantom.gsc.riken.jp/.
 Regulome DB: <http://regulomedb.org/>.
 GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes): <https://grasp.nhlbi.nih.gov/Overview.aspx>.

FURTHER READING

- Augui S, Nora EP, Heard E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat Rev Genet* 2011;12:429–42.

Human Genomic Variants and Inherited Disease: Molecular Mechanisms and Clinical Consequences

Stylianos E. Antonarakis¹, David N. Cooper²

¹Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

²Institute of Medical Genetics, Cardiff University, Cardiff, United Kingdom

6.1 INTRODUCTION

Each individual human genome is unique and varies from the reference genome at millions of sites. This genomic individuality contributes considerably to the biological identity of each person; a very small fraction of the genomic variation is pathogenic and contributes to the various disease phenotypes. The recent advances in rapid and relatively inexpensive DNA sequencing, the development of computational tools for data analysis, the functional exploration of the genome, the creation of databases of genomic variants from hundreds of thousands of individuals, the use of model organisms for the functional characterization of variants, have had a major impact on the development of genomic medicine. Genomic medicine, which is based on genomic variation, is at the heart of understanding the molecular pathophysiology of human disorders. This chapter provides examples of the different varieties of pathogenic variants in the different functional elements of the human genome.

A major aid in recognizing a high-impact pathogenic variant in the sea of neutral or slightly deleterious variants is the existence of databases of variants linked to phenotypes. The existing databases, however, are still rather small, and a major challenge in the coming years is to create, maintain and update large, accessible, and high-quality international databases [1].

A wide variety of different types of pathogenic variants occur in the human genome, with many diverse mechanisms being responsible for their generation: single base-pair substitutions in coding, regulatory and splicing-relevant regions of human genes (67.4%), as well as microdeletions (14.7%), microinsertions (6.2%), gross insertions and duplications (1.8%), repeat expansions (0.2%), combined microinsertions/deletions (“indels”) (1.4%), gross deletions (7.4%), gross insertions (1.8%), inversions, and other complex rearrangements (0.9%).

Characterized genomic variants occur not only in coding sequences but also in promoter regions, splice junctions, and within introns and untranslated regions, and noncoding RNAs. Different types of human gene variants may vary in size, from structural variants to single base-pair substitutions, but what they all have in common is that their nature, size and location are often determined either by specific characteristics of the local DNA sequence environment or by higher order features of the genomic architecture. A major goal of genomic medicine is to be able to predict the nature of the clinical phenotype through ascertainment of the genotype. The extent to which this is feasible in medical genetics is very much disease, gene, and mutation dependent. The study of variants in human genes is nevertheless of paramount importance for understanding the pathophysiology of inherited disorders, optimizing diagnostic testing and

guiding the design of emergent therapies. A major goal of molecular genetic medicine is to be able to predict the nature of the clinical phenotype through ascertainment of the genotype. The extent to which this is feasible in medical genetics is very much disease, gene, and variant dependent.

The first description of the precise molecular defect in a human disease (the sickle cell variant, a Glu to Val substitution at the sixth codon of the β -globin [*HBB*] gene) was identified by Ingram in 1956 [2], who found that the difference between hemoglobin A and hemoglobin S lies in a single tryptic peptide. His analysis was made possible by the methods developed by Sanger for determining the structure of insulin and by Edman to effect the stepwise degradation of peptides. This was followed, 40 years ago, by the characterization of the first heritable pathogenic variants in a human gene at the DNA level: gross deletions of the human α -globin (*HBA*) and *HBB* gene clusters giving rise to α - and β -thalassemia [3] and a single base-pair substitution (Lys17Term) in *HBB* gene causing β -thalassemia [4]. Since then, continuous technical advances have enabled the identification of numerous disease-related genes and the discovery of thousands of underlying pathological lesions [5]. Single base-pair substitutions (67%) and microdeletions (15.6%) are the most frequently encountered variants in the human genome, the remainder comprising an assortment of microinsertions (6.5%), indels (1.5%), gross deletions (6.6%), gross insertions and duplications (1.4%), inversions, repeat expansions (0.3%), and complex rearrangements (1.0%).

The vast majority of single nucleotide variants listed in Human Gene Mutation Database (HGMD) reside within the coding region (84%), the remainder being located in either intronic (13%) or regulatory (3%, promoter, untranslated, or flanking regions) sequences. Variants may interfere with any stage in the pathway of expression, from gene activation to synthesis and secretion of the mature protein product. The question of the proportion of possible variants within human disease genes that are likely to be of pathological significance is one that is difficult to address because it is dependent not only on the type and location of the variant but also on the functionality of the nucleotides involved (itself dependent in part upon the amino acid residues that they encode), which is often hard to assess [6–11]. In addition, some types of variants are likely to be much more comprehensively ascertained than others, making observational comparisons between mutation types an inherently hazardous undertaking.

Different types of human gene variants may vary in size, from structural variants (SVs) to single base-pair substitutions, but what they all have in common is that their nature, size, and location are often determined either by specific characteristics of the local DNA sequence environment or by higher order features of the genomic architecture [12]. This chapter attempts to provide an overview of the nature of variants causing human genetic disease and then considers their consequences for the clinical phenotype. Three online databases, which interested readers may consult, contain information on known disease-related (pathogenic) variants: the HGMD (hgmd.org), *Mendelian Inheritance in Man* (omim.org/), and *ClinVar* (ncbi.nlm.nih.gov/clinvar/). The HGMD contains 225,000 likely pathogenic variants; *ClinVar* includes 431,000 variants with interpretation, a substantial fraction of which are not pathogenic; OMIM contains only representative pathogenic variants per gene in 4150 protein-coding genes.

The advances in DNA sequencing technologies, the computational analysis of the data, the development of public databases of genomic variants of thousands of individuals and the international exchanges of prepublication data, as well as guidelines and criteria for assessing the pathogenicity of genomic variants have greatly advanced the discovery of gene–disease links, primarily on mendelian phenotypes, and has tremendously expanded the pool of pathogenic and likely pathogenic variants [13–18]. The criteria for assessing pathogenicity from the American College of Medical Genetics (ACMG) are particularly useful and are implemented in the diagnostic services worldwide [17]. The ACMG criteria categorize the variants in five classes regarding pathogenicity: benign, likely benign, variant of unknown significance, likely pathogenic, and pathogenic.

An excellent discussion on the origins, determinants, and consequences of human mutations has been recently published [19] and is recommended to the reader.

6.2 MOLECULAR MECHANISMS OF VARIANTS CAUSING HUMAN INHERITED DISEASE

6.2.1 “Neutral Variation”/DNA Polymorphisms

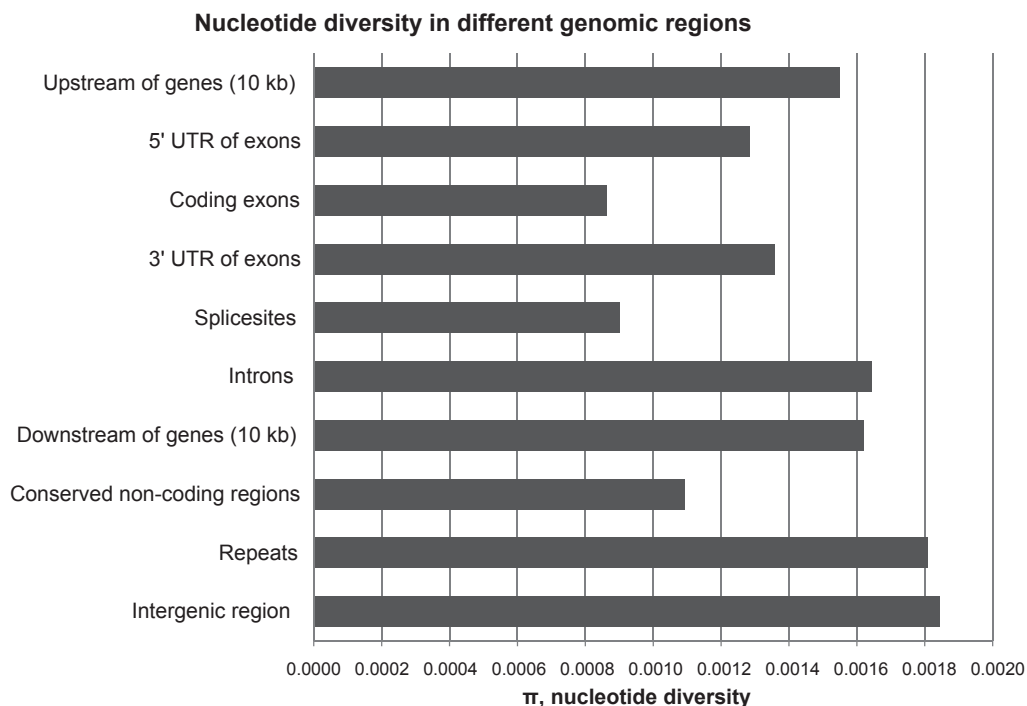
The term *polymorphism* has been defined [20] as a “Mendelian trait that exists in the population in at least two phenotypes, neither of which occurs at a frequency of less than 1%.” Polymorphisms are not therefore

rare. Indeed, there is enormous variation in the DNA sequences of any two randomly chosen human haploid genomes. Clearly, not all variations within a gene result in the abnormal expression of protein products. Indeed, single nucleotide substitutions/polymorphisms (SNPs) occur in 1:~600–1200 nucleotides in intervening sequences and flanking DNA (2005; [21–25]). These substitutions represent the most common form of DNA polymorphism that can be used as markers for specific regions of the human genome. Similarly, some single nucleotide substitutions in the coding regions of genes may also be normal (nonpathogenic) polymorphic variants even if they result in nonsynonymous substitutions of the polypeptide product [26]. For example, there are three common forms of *HBB* gene on chromosome 11p; these forms differ at five nucleotides, one of which lies within the first exon of the gene and results in a synonymous codon. The average human gene contains >120 biallelic polymorphisms, 46 of which occur with a frequency >5%, with five occurring within the coding region [27].

Some polymorphisms entail the alteration of an encoded amino acid, for example, the Lewis *Le* alleles of the *FUT3* gene [28], whereas others may introduce a stop codon that serves to inactivate the gene in question—for example, the secretor *se* allele of the *FUT2* gene present in 20% of the population [29]. However, not all polymorphisms are SNPs. Examples of other types of gene-associated polymorphisms in the human genome include triplet repeat copy number (e.g., in the *FMR1* gene; see 9.2.1.3), gross gene deletion (e.g., *GSTM1* and *GSTT1* [30]), gene duplication (e.g., *HBB2* [31]), intragenic duplication (e.g., *IVL* [32]), microinsertion/deletion (e.g., *PAI1* [33]), indel (e.g., *APOE* [34]), gross insertion (e.g., the inserted *Alu* sequence in intron 16 of the *ACE* gene [35]), inversion (e.g., the 48-kb Xq28 inversion involving the *EMD* and *FLN1* genes [36]), and gene fusion (e.g., between the *RCP* and *GCP* visual pigment genes [37]). Functional polymorphisms may occur within the coding region [38] or regulatory regions [39] of a gene or may impact on pre-messenger RNA (mRNA) splicing [40] and therefore can have consequences for protein structure/function, gene expression, or mRNA splicing. It can be seen that the spectrum of polymorphisms in the human genome is qualitatively different than the variants underlying human disease; they may vary in terms of location and frequency but otherwise they display remarkable similarities indicative of the same underlying mutational mechanisms.

It is likely that some SNPs, whether frequent or rare, alter the risk of common complex human phenotypes (“functional SNPs”). A public SNP database now contains >660 million entries (dbSNP; ncbi.nlm.nih.gov/SNP/snp_summary.cgi). An international project termed the “HapMap project” [41–43] had the objective to define the patterns of common SNP genetic variation in a sample of 270 DNAs from individuals of European, African, Chinese, and Japanese origin (hapmap.org). The data obtained from this project constitute ~2.8 million SNPs and are publicly available. The results of this and other similar projects are contributing significantly to our understanding of both common and rare human genetic disorders and traits. Furthermore, recent advances in high-throughput sequencing have led to the discovery of a large number of individually rare polymorphic variants in samples from the 1000 Genomes [44] and other projects. The protein-coding regions and splice junctions of a typical human genome (also known as the exome) contain 9000–11,000 nonsynonymous variants, ~100 nonsense codons, and 35 splice variants [45]. Analysis of 54 human genomes, sequenced by Complete Genomics (completegenomics.com), revealed 3,700,000–4,700,000 single nucleotide variants per genome; the frequency of these variants was not identical in various fractions of the genome and is related to regional evolutionary constraints. Fig. 6.1 illustrates that the protein-coding fraction of the genome, which is under evolutionary pressure, contains the smallest number of variants per kilobase (kb); by contrast, the repeat fraction of the genome and the intergenic regions, which presumably evolve neutrally, contain almost double the number of variants per kb.

The genome of each individual contains a number of likely damaging alleles for the encoded protein, and not all of them contribute to recognizable phenotypes. The sequence of exomes of 3222 British Pakistani-heritage adults from consanguineous marriages has revealed 1111 homozygous rare variants with predicted loss of function in 781 genes. On average, there were 1.6 homozygous LOF variants per individual; remarkably, these homozygous variants were found in apparently healthy people [46]. One of the first studies to assess the number of predicted deleterious variants in normal individuals has examined low-coverage whole-genome sequences from 179 individuals. Each individual carried 281–515 missense substitutions, 40–85 of which were homozygous, predicted to be highly damaging. They also carried



- Genomic variants from CompleteGenomics, 54 unrelated individuals from different ethnic groups (www.completegenomics.com)
- Genomic regions drawn from UCSC genome browser (<http://genome.ucsc.edu>)

Figure 6.1 Nucleotide diversity (equivalent to frequency of polymorphic variants) in different genomic regions. The genomic variants analyzed are from the whole genome sequences of 54 unrelated human genomes (see text).

40–110 variants classified by the Human Gene Mutation Database (HGMD) as disease-causing variants, 3–24 variants in the homozygous state [47].

Another form of polymorphic variation in our genome is the presence of variable numbers of tandem repeats. The repeat unit can be 10–60 nucleotides in length and many different alleles may exist at a given locus [48,49]. The combination of a VNTR and single nucleotide substitutions within the repeat unit results in an extremely high level of polymorphic variability, which can be used as a unique bar code to distinguish different individuals [50]. The introduction of the polymerase chain reaction (PCR) [51] permitted the rapid detection and analysis of variation in short sequence repeats (SSRs), for example (GT)_n repeats [52,53]. These are common polymorphisms that occur on average once for every 50 kb of genomic DNA. The SSRs also display many alleles and the repeat unit can be two, three, four, five, or more nucleotides. Poly(A) tracts may also be

polymorphic, exhibiting variation in the number of A residues [54]; many of these polymorphisms are localized at the ends of *Alu* repetitive elements. Another kind of polymorphism in the human genome involves the presence or absence of retrotransposons (i.e., *Alu* or LINE repetitive elements or pseudogenes) at specific locations [55,56]. Duplicational polymorphisms in some human genes, such as *HBA1*, *PRB1-4*, *HBZ*, and *CYP21/C4A/C4B*, have been known for some time [56,57]. The use of comparative genomic hybridization against BAC or oligonucleotide arrays has revealed extensive copy number polymorphism/variation (CNP or CNV) of sizeable genomic regions [58–60]. Details of many thousands of such genomic variants may be found in the following databases: CNV Project, <http://www.sanger.ac.uk/humgen/cnv>, and Database of Genomic Variants, <http://projects.tcag.ca/variation>. A first CNV map of the human genome of the 270 “HapMap” individuals revealed a total of 1440 CNV regions covering ~360 megabases

(Mb) (12% of the genome) [61]. High-resolution tiling oligonucleotide microarrays have been used to generate comprehensive genomic maps of >10,000 CNVs [62,63]. The functional significance, if any, of most of these polymorphic variants is, however, unknown. To understand the prevailing mutational mechanisms responsible for human genome structural variation, a total of 1054 large structural variants have been sequenced (589 deletions, 384 insertions, and 81 inversions from 17 human genomes). The prevailing mechanisms for these structural variations were: i/microhomology-mediated processes involving short (2–20 bp) sequences (28%), nonallelic homologous recombination (22%), and L1 retrotransposition (19%) [64].

It is clear that no single individual genome contains the full complement of functional genes [65], a paradigm shift that strikes at the heart of the concept of a “reference genome” [66].

Deletion polymorphisms are also remarkably frequent in the human genome: a typical individual has been estimated to be hemizygous for some 30–50 deletions >5 kb, spanning >550 kb in total, and encompassing >250 known or predicted genes [67,68]. Because such deletions appear to be in linkage disequilibrium with neighboring SNPs, we may surmise that they share a common evolutionary history [69].

Human DNA polymorphisms have proven extremely useful in developing linkage maps, for mapping monogenic and polygenic complex disorders, for determining the origin of aneuploidies and chromosomal abnormalities, for distinguishing normal from mutant chromosomes in genetic diagnoses, for performing forensic, paternity, and transplantation studies, for studying the evolution of the genome, the loss of heterozygosity in certain malignancies, the detection of uniparental disomy, the instability of the genome in certain tumors, recombination at the level of the genome, the study of allelic expression imbalance, and the development of haplotype maps of the genome. In studying the role of a candidate gene in a given disorder, it is imperative to distinguish between pathogenic variants that cause a clinical phenotype and the polymorphic variability of the normal genome.

6.2.2 Nonsense SNPs

The loss of a particular gene/allele is not invariably associated with a readily discernible clinical phenotype [70,71]. This assertion is supported by the

identification of more than 1000 putative nonsense SNPs (i.e., nonsense variants that have attained polymorphic frequencies) in human populations [72,73]. About half of these nonsense SNPs have been validated by dbSNP (ncbi.nlm.nih.gov/projects/SNP), a process that involves the exclusion of variants in pseudogenes and of artifacts caused by sequencing errors. Bona fide nonsense SNPs are expected either to lead to the synthesis of a truncated protein product or alternatively to the greatly reduced synthesis of the truncated protein product (if the mRNA bearing them is subject to nonsense-mediated mRNA decay [NMD]). Based on the relative locations of the nonsense SNPs and the exon-intron structures of the affected genes, Yamaguchi-Kabata et al. [74] concluded that 49% of nonsense SNPs would be predicted to elicit NMD, whereas 51% would be predicted to yield truncated proteins. Some of these nonsense SNPs have been found to occur in the homozygous state in normal populations [73], attesting to the likely functional redundancy of the corresponding genes. At the very least, genes harboring nonsense SNPs may be assumed to be only under weak selection [72].

It should be appreciated that nonsense SNPs may even occur in “essential” genes, yet still fail to come to clinical attention (or give rise to a detectable phenotype) if these genes are subject to CNV (see CNVs later) that masks any deleterious consequences by ensuring an adequate level of gene expression from additional wild-type copies either in *cis* or in *trans*. Thus, CNV might serve to “rescue” the full or partial loss of gene function brought about by the nonsense variants, thereby accounting for the occurrence of the latter at polymorphic frequencies. Consistent with this postulate [72], it was reported [72] that ~30% of nonsense SNPs occur in genes residing within segmental duplications, a proportion some threefold larger than that noted for synonymous SNPs. Genes harboring nonsense SNPs were also found to belong to gene families of higher than average size [72], suggesting that some functional redundancy may exist between paralogous human genes. In support of this idea, Hsiao and Vitkup [75] reported that those human genes that have a homolog with ≥90% sequence similarity are approximately three times less likely [76] to harbor disease-causing variants than are genes with less closely related homologs. They interpreted their findings in terms of “genetic robustness” against null variants,

with the duplicated sequences providing “back-up” by potentiating the functional compensation/complementation of homologous genes in the event that they acquire deleterious variants.

6.3 DISEASE-CAUSING VARIANTS

6.3.1 The Nature of Genomic Variants

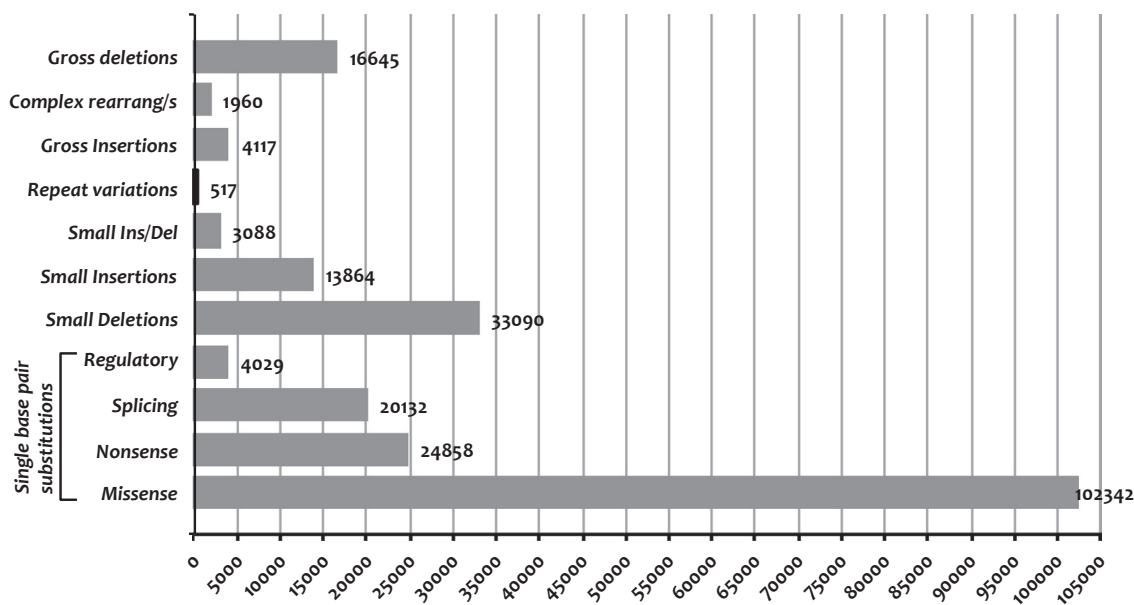
Fig. 6.2 depicts the frequencies of the various genomic variant types responsible for molecularly characterized human genetic disorders, as recorded in HGMD (<http://www.hgmd.org>) and studies [77–80]. HGMD records each variant *once*, regardless of the number of independent occurrences of that lesion. Fig. 6.2 shows the frequency of the first variant per disease recorded in MIM (omim.org/). As of July 2018, HGMD contained some 225,000 different disease-causing variants and disease-associated/functional polymorphisms in 8784 human genes

(Fig. 6.2), whereas MIM contained selected examples of allelic variants in 4150 human genes associated with a specific phenotype.

6.3.2 Nucleotide Substitutions

Single nucleotide substitutions are the most frequent pathological variants in the human genome (Fig. 6.2). Most of these alterations occur during DNA replication, which is an accurate yet error-prone multistep process. The accuracy of DNA replication depends on the fidelity of the replicative step and the efficiency of the subsequent error correction mechanisms [81]. Analysis of more than 7000 missense and nonsense variants associated with human disease has indicated that the most common nucleotide substitution for T (thymine) is to C (cytosine), for C it is to T, for A (adenine) it is to G (guanine), and for G it is to A [82]. Transitions are, therefore, much more common than transversions. Some 61% of the missense and nonsense variants currently logged in

The Human Gene Mutation Database (HGMD)



224642 pathogenic mutations in 8784 genes (13jul18)

D Cooper, PD Stenson et al, Cardiff, UK

sea3112

Figure 6.2 Spectrum of different types of human disease-causing mutations and disease-associated/functional polymorphisms logged in the HGMD as of July 2018.

HGMD are transitions (T to C, C to T, A to G, G to A) while 39% are transversions (T to A or G, A to T or C, G to C or T, C to G or A).

Among single nucleotide substitutions, there is one that clearly predominates and represents the most common type of mutational lesion in the human genome: CpG dinucleotides mutate to TpG at a frequency that is about five times higher than mutations in all other dinucleotides [82–85]. This substitution, which when it occurs on one DNA strand generates TG, and on the other, CA (the “CG to TG or CA rule”) represents a major cause of human genetic disease. This phenomenon was first observed in the factor VIII (*F8*) gene in cases of hemophilia A [85], but it was soon noted in the studies of many other genes [86]. In hemophilia A, CG to TG or CA mutations account for 46% of single nucleotide variants in unrelated patients [87]. In the HGMD (www.hgmd.org), such variants currently account for ~18% of the total number of missense and nonsense variants [76]. Among CpG dinucleotide mutations, transitions to TG or CA account for ~90% of substitutions. The mechanism of this common type of mutation appears to be methylation-mediated deamination of 5-methylcytosine (5mC). In eukaryotic genomes, 5mC occurs predominantly in CpG dinucleotides, most of which appear to be methylated (see [88] for review). 5mC then undergoes spontaneous nonenzymatic deamination to form thymine (Fig. 6.3). There is a bias in terms of the origin of CpG to TpG mutations: most occur in male germ cells (the male:female ratio is 7:1). One reason for this may be that sperm DNA is heavily methylated, whereas oocyte DNA is comparatively undermethylated [89]. Another reason may be the considerably higher number of germline cell divisions in males as compared to females [90].

Cytosine methylation also occurs in the context of CpNpG sites [91]. If we assume not only that CpNpG methylation occurs in the germline but also that 5mC deamination can occur within a CpNpG context, then it follows that methylated CpHpG sites are also very likely to constitute mutation hotspots causing human inherited disease. Initial evidence that this might indeed be the case came from the observation that disproportionately high numbers of C>T and G>A transitions occur at CpNpG sites in studies of the human genes, *NF1* [92] and *BRCA1* [93]. Further, ~9.9% of 54,625 missense and nonsense variants from 2113 genes causing inherited disease (HGMD) are C>T and G>A transitions located within CpNpG trinucleotides, approximately twofold higher proportion than would have been expected by chance alone [76]. Some 5% of missense or nonsense variants causing human inherited disease may, therefore, be attributable to methylation-mediated deamination of 5mC within a CpNpG context.

In a recent analysis, the average direct estimate of the combined rate of all mutations was 1.8×10^{-8} per nucleotide per generation [94]. Single nucleotide substitutions were found to be ~25 times more common than all other mutations, while deletions were ~3 times more common than insertions; complex mutations were very rare and the CpG context was found to increase substitution rates by an order of magnitude [94]. Rates of different kinds of mutations were also found to be strongly correlated across different loci [94].

It has been estimated that ~20% of new missense variants in humans result in a loss of function, whereas 53% have mildly deleterious effects, and 27% are effectively neutral with respect to phenotype [95]. These estimates have received independent support, at least qualitatively, from a study of human coding SNPs by

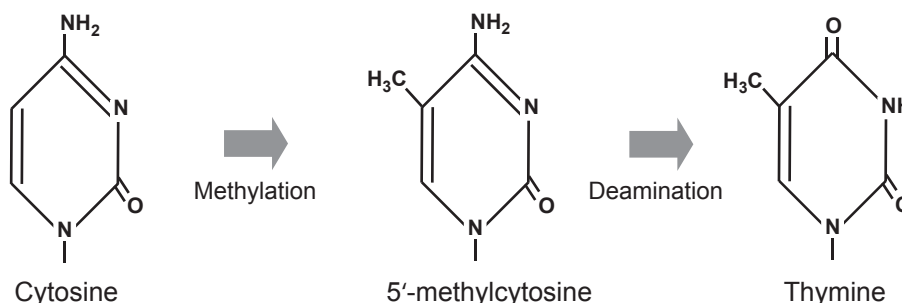


Figure 6.3 Schematic representation of cytosine, 5-methylcytosine, and thymine, and the chemical events involved in the mutational transformation of cytosine to thymine.

Boyko et al. [96], who predicted that 27%–29% of missense variants would be neutral or near neutral, 30%–42% would be moderately deleterious, with most of the rest (i.e., 29%–43%) being highly deleterious or lethal, and by Eyre-Walker et al. [97] who estimated that >50% of variants would be likely to exert only a mild effect on the phenotype.

A recent study using human osteosarcoma cell lines has shown that noncanonical (non-B) DNA conformations are capable of increasing the overall spectrum of single base-pair substitutions in a reporter gene in *cis* by exposing those DNA sequences to oxidative damage [98]. In this study, the spectrum of single base-pair substitutions was shown to be indistinguishable from that induced by other conditions known to lead to a hyperoxidative state (such as WRN deficiency and lung tumorigenesis), an observation that lends support to a model whereby DNA bases become oxidized, followed by the transfer of their oxidized state (“hole migration”) to target neighboring bases. If these observations are eventually found to be relevant in the context of “natural” chromatin during meiosis, then the impact of non-B DNA conformations on human inherited disease, both with respect to single base-pair substitutions and gross rearrangements (see Section 6.3.6), could be quite significant. Further, because non-B DNA structures can interfere with DNA replication and repair, and may serve to increase mutation frequencies in generalized fashion, they have the potential to serve as a unifying concept in studies of mutational mechanisms underlying human inherited disease.

6.3.3 Synonymous Nucleotide Substitutions

Synonymous (“silent”) variants, although not altering the amino acid sequence of the encoded protein directly, can still influence splicing accuracy or efficiency [99–104]. It has become increasingly clear that apparently silent SNPs may also become distinctly “audible” in the context of mRNA stability or even protein structure and function. Thus, three common haplotypes of the human *COMT* gene, which differ in terms of two synonymous and one nonsynonymous substitution, confer differences in *COMT* enzymatic activity and pain sensitivity [105,106]. The major *COMT* haplotypes differed with respect to the stability of the *COMT* mRNA local stem-loop structures, the most stable being associated with the lowest levels of *COMT* protein and enzymatic activity [106]. In a similar vein, synonymous SNPs in the *ABCB1* gene have been

shown to alter *ABCB1* protein structure and activity [107], possibly by changing the timing of protein folding following extended ribosomal pause times at rare codons [108]. Finally, it should be understood that although the deleteriousness of the average synonymous variants is always likely to be less than that of a nonsynonymous (missense) variants [96], the higher prevalence of synonymous variant means that they may actually make a significantly greater contribution to the phenotype than nonsynonymous variants [109].

6.3.4 Microdeletions and Microinsertions

Deletions or insertions of a few nucleotides are also fairly common as a cause of human inherited disease. Most of these are less than 20 bp in length. Indeed, the majority of microdeletions involve <5 nucleotides. In HGMD, the deletion of 1 bp accounts for 48% of small deletions while an additional 30% involve two or three nucleotides. The majority of microdeletions recorded (78%) result in an alteration of the reading frame. Most microdeletions occur in regions that contain direct repeats of 2 bp or more. The most common length of direct repeat is 3 bp (48% of direct repeats associated with short deletions [83]. The most plausible mechanism for small deletions mediated by the presence of direct repeats is the slipped mispairing model [110] (Fig. 6.4). In addition, deletions of one or a few nucleotides frequently occur in runs of the same nucleotide, for example, a poly(T) region [111]. Finally, inverted repeats and “symmetric elements” are also frequently found in the immediate vicinity of microdeletions [112,113]. Krawczak

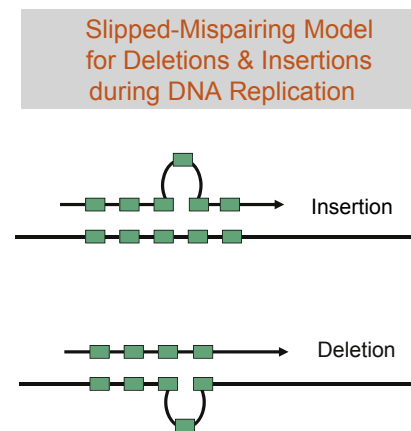


Figure 6.4 Schematic representation of the slipped mispairing model for deletions and insertions during DNA replication.

and Cooper [114] identified a consensus sequence—TG(A/G)(A/G)(G/T)(A/C)—which they claimed to represent a deletion hotspot.

Microinsertions (again up to 20 nucleotides) are rarer than microdeletions; thus, in HGMD there are three times as many microdeletions as microinsertions (Fig. 6.2). Nearly half of these involve the insertion of only one nucleotide (Fig. 6.5). As is the case with microdeletions, most microinsertions lead to alterations of the reading frame and are located in regions containing direct or inverted repeats or runs of the same nucleotide. Details of possible mechanisms of generation during replication can be found in [115]; however,

there are as-yet insufficient data available to estimate the frequency ratio of microinsertions or microdeletions in male or female germ cells. In the case of such lesions in factor VIII (*F8*) gene, 56% of microdeletions/insertions have been reported to occur in DNA regions harboring direct repeats or runs of the same nucleotide [87].

HGMD data (3767 microdeletions and 1960 microinsertions) were used to perform a meta-analysis of microdeletions and microinsertions causing inherited disease, both defined as involving ≤ 20 bp DNA [116]. A positive correlation was noted between the microdeletion and microinsertion frequencies for 564 genes in which both microdeletions and microinsertions have

HGMD Small Deletions and Insertions (1-Dec-11)

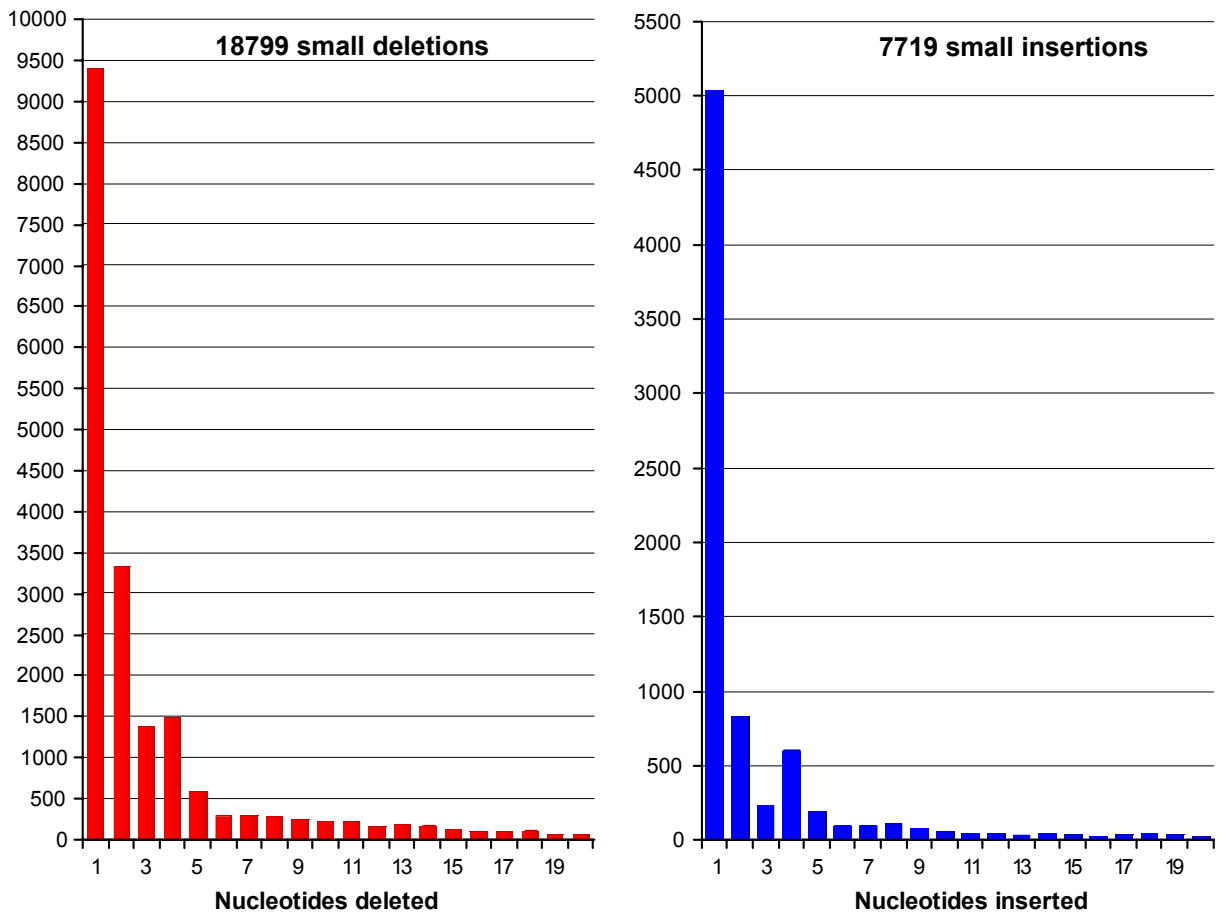


Figure 6.5 Size distribution of short (<20bp) pathogenic human gene deletions and insertions (HGMD; <http://www.hgmd.org>; January 5, 2007).

been reported. This is consistent with the view that the propensity of a given gene/sequence to undergo microdeletion is related to its propensity to undergo microinsertion. While microdeletions and microinsertions of 1 bp constitute, respectively, 48% and 66% of the corresponding totals, the relative frequency of the remaining lesions correlates negatively with the length of the DNA sequence deleted or inserted. Many microdeletions and microinsertions of >1 bp are potentially explicable in terms of slippage mutagenesis, involving the addition or removal of one copy of a mono-, di-, or trinucleotide tandem repeat. The frequency of in-frame 3-bp and 6-bp microinsertions and microdeletions was, however, found to be significantly lower than that of variants of other length, suggesting that some of these in-frame lesions may not have come to clinical attention. Various sequence motifs were found to be overrepresented in the vicinity of both microinsertions and microdeletions, including the heptanucleotide CCCCTG that shares homology with the complement of the 8-bp human minisatellite conserved sequence/chi-like element (GCWGGWGG). The “indel hotspot” GTAAGT (and its complement ACTTAC) were also found to be overrepresented in the vicinity of both microinsertions and microdeletions, thereby providing a first example of a mutational hotspot that is common to different types of gene lesions. Other motifs overrepresented in the vicinity of microdeletions and microinsertions included DNA polymerase pause sites and topoisomerase cleavage sites. Several novel microdeletion/microinsertion hotspots were noted and some of these exhibited sufficient similarity to one another to justify terming them “super-hotspot” motifs. Analysis of DNA sequence complexity also demonstrated that a combination of slipped mispairing mediated by direct repeats, and secondary structure formation promoted by symmetric elements, can account for the majority of microdeletions and microinsertions. Thus, microinsertions and microdeletions exhibit strong similarities in terms of the characteristics of their flanking DNA sequences, implying that they are generated by very similar underlying mechanisms.

A similar analysis on microdeletions and microinsertions in 19 human genes presented evidence for an elevated microdeletion rate at YYTG and an elevated microinsertion rate at TACCRC and ATMMGCC [117]. These authors also found that ~45% of microdeletions

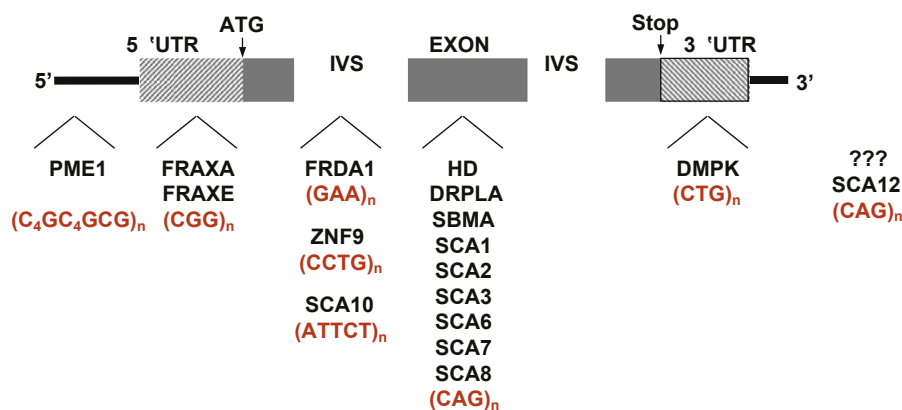
led to the removal of a repeated sequence, an event they termed “deduplication” in order to highlight the identity of the deleted sequence and the sequence abutting the site of deletion.

Another mutational mechanism, DNA triplex formation followed by DNA repair, has been proposed to explain ~ 5% of microdeletions and microinsertions at mirror repeats [118].

6.3.5 Expansion/CNV of Trinucleotide (and Other) Repeat Sequences

Another mechanism of human gene variation causing hereditary disease is the instability of repeat (mainly trinucleotide) sequences and their expansion in affected genes [119–121]. A growing number of repeat expansion disorders (in excess of 66 are now recorded in HGMD), the majority of which involve neuromuscular tissue, have been found to be due to, or associated with, the expansion of repeat sequences; of these, 53 are expansions of triplet repeats. The first such disease was fragile X, a common cause of male mental retardation, which mapped to chromosome Xq27.3. Some examples of these disorders include Huntington disease, myotonic dystrophy, spinobulbar muscular atrophy, spinocerebellar ataxia 1, spinocerebellar ataxia 3, or Machado–Joseph disease, the fragile E site, and dentatorubral pallidoluy-sian atrophy. Genetic “anticipation” (the earlier onset and increasingly severe phenotype in successive generations) is a common phenomenon in these disorders [122]. The trinucleotide involved is usually either CAG or CGG but occasionally CTG, GCG, or GAA. It can be located in the 5′ untranslated region (UTR) as in the case of the *FMR1* gene underlying fragile X, within the coding region (as in Huntington disease, spinocerebellar ataxia 1 [SCA1], SCA3, and Kennedy disease) where it encodes poly(Gln), in an intron as in Friedreich ataxia (*FXN*) and myotonic dystrophy type 2 (*ZNF9*), or in the 3′ UTR as in myotonic dystrophy type 1 (*DMPK*) (Fig. 6.6). The expansion of the triplet repeat either prevents the expression of the associated gene [123], results in a dominant gain-of-function variant mediated by the longer poly(Gln) peptide [124], or alters the RNA processing of other genes [125,126].

Trinucleotide repeats are usually polymorphic in human populations. Rarely, however, the number of trinucleotide repeats lies within a high-risk category that is termed a “premutation.” In such a case, the premutation exhibits a high probability of further expansion



Selected Repeat Expansions in Human Disorders

Figure 6.6 Location of the repeat expansion in selected human disorders.

(instability) to yield disease-related alleles (“full mutation”). In fragile X, for example, the normal polymorphic alleles of the CGG repeat contain between 10 and 50 triplets, the premutation between 50 and 200 triplets, and the full mutation more than 200 triplets [127]. Expansion of premutations to full mutations only occurs during female meiotic transmission. The probability of repeat expansion correlates with repeat copy number in the premutated allele. Because the premutation must precede the appearance of a full mutation, all mothers of affected children carry either a full mutation or a premutation [127]. Premutation alleles may also be associated with late-onset movement disorders and premature ovarian failure [128,129].

The precise mechanism of repeat expansion is unclear, although it is known that DNA polymerase progression is blocked by CTG and CGG repeats and the resultant idling of the polymerase could serve to catalyze slippage leading to repeat expansion [130]. In the case of SCA1, interruption of the CAG repeat with a CAT unit is associated with more stable trinucleotide repeat [131]. More details of these “dynamic mutations” can be found in the appropriate sections covering individual disorders, and in [132]. Short expansions of GCG trinucleotide codons encoding Ala have been observed in the *HOXD13* gene causing dominant polydactyly, and in the *PABP2* gene causing oculopharyngeal muscular dystrophy [133,134]. These mutations may be due to unequal crossing-over rather than polymerase slippage. Generally speaking, it is likely that repeat instability is

a consequence of the resolution of unusual secondary structure intermediates during DNA replication, repair, and recombination [135].

A repeat expansion of 12 nucleotides (CCCCGC-CCCCGCG) in the 5' flanking region of the *CSTB* gene causes one form of the recessive progressive myoclonus epilepsy type 1 (EPM1) [136]. This indicates that repeat sequences other than trinucleotides can expand and cause human disorders. This particular expansion silences the *CSTB* gene, probably because it alters the spacing of transcription factor binding sites from each other and/or the transcriptional initiation site [137].

A tetranucleotide repeat expansion $(CCTG)_n$ in intron 1 of the *ZNF9* gene causes myotonic dystrophy type 2 [125]. This expansion can be between 75 and 11,000 repeats in length. The expansion of the pentanucleotide repeat $(ATTCT)_n$ is responsible for the phenotype of spinocerebellar ataxia 10 (SCA10). The expansion occurs in intron 9 of the *SCA10* gene and can be up to 22.5 kb in length [138]. Expansions of even longer repeats have been reported. In Usher syndrome type 1C, for example, there is an expansion of a 45-bp VNTR in intron 5 of the *USH1C* gene (nine tandem repeats instead of the usual less than six such repeats); this expansion has been predicted to inhibit the transcription of the gene [139]. There are also cases in which a large repeat expansion is not associated with a particular phenotype, for example, the expansion of an AT-rich 33-mer repeat in the dictamycin-sensitive fragile site 16B [140].

6.3.6 Mechanisms of Gross Genomic Rearrangement

Structural variation in the human genome is characterized by a number of different types of gross rearrangements including deletions, duplications, insertions (termed CNVs), as well as inversions, and translocations. Four major mutational mechanisms account for these SVs: nonallelic homologous recombination (NAHR), nonhomologous end joining (NHEJ), replication-based mechanisms, and L1-retrotransposition [62–64].

NAHR: Sequence analysis of the breakpoints of 1054 SVs identified in the genomes of 17 healthy human individuals revealed that NAHR accounts for 22% of insertions and deletions as well as 69% of inversions [64]. The majority of these SVs are likely to represent neutral polymorphisms but ~1% may be disease associated. Some apparently neutral SVs appear to predispose to further structural rearrangements, such as deletions and duplications, which in turn give rise to disease [141–145]. Thus, for example, heterozygosity for the ~970-kb inversion polymorphism of the *MAPT* locus at 17q21.3 predisposes to the NAHR events that underlie the 17q21.31 microdeletion syndrome [146,147]. It may be that inversion heterozygosity perturbs the pairing of homologous chromosomes during meiosis, which then promotes interchromosomal NAHR between the inversion-flanking low copy repeats (LCRs), thereby giving rise to the 17q21.3 microdeletion.

During meiosis, NAHR between sequences that are nonallelic (i.e., paralogous) can result in recurrent deletions and duplications that cause specific genomic disorders. Liu et al. [148] studied two patient cohorts with reciprocal genomic disorders localized to chromosome 17p11.2: the deletion-associated Smith–Magenis syndrome and the duplication-associated Potocki–Lupski syndrome. They reported that complex rearrangements (those with more than one breakpoint) were more prevalent in copy number gains (17.7%) than in copy number losses (2.3%), an observation which supports a role for replicative mechanisms in the formation of complex rearrangements. With respect to the NAHR-mediated recurrent rearrangements, the crossover frequency was found to be positively associated with the flanking LCR length and inversely influenced by the inter-LCR distance. It would, therefore, appear that the probability of ectopic chromosome synapsis increases with increasing LCR length, with ectopic synapsis being a prerequisite for ectopic crossing-over.

Recent findings also indicate that NAHR represents a major mechanism underlying unbalanced recurrent translocations, which are mediated by either interchromosomal LCRs or segmental duplications located on nonhomologous chromosomes [149].

NHEJ: The defining characteristic of NHEJ is the ligation of double-strand break (DSB) ends without the requirement for extensive homology, in stark contrast to the situation pertaining with homologous recombination. The presence of terminal microhomologies (typically 1–3 bp) facilitates NHEJ, but this appears not to be an absolute requirement; only 30%–50% of all SVs in the human genome have originated through microhomology-mediated NHEJ events [62,150].

Although some NHEJ events would have resulted from the repair of DSBs that originated quasi-randomly, there are also many well-documented cases in which the locations of the NHEJ-initiating DSBs appear to be highly dependent on the local DNA sequence environment. The role of the local DNA sequence context in generating NHEJ-mediated germline mutations is exemplified by the constitutional t(11;22)(q23;q11), the most common type of recurrent non-Robertsonian translocation in humans [151,152]. The breakpoint sequences of both chromosomes are characterized by several hundred base pairs of inverted AT-rich repeats; similar sequences have also been identified at the breakpoints of other nonrecurrent translocations [153]. It would appear that the NHEJ of two ends from different DSBs requires those ends to be physically located in the immediate vicinity. Indeed, DSBs tend to undergo translocations with those chromosomes with which they share nuclear space [154]. This provides strong support for the “contact-first” hypothesis, which proposes that interactions between different DSBs can only take place if they are colocalized at the time of DNA damage [155]. Consistent with this hypothesis, close spatial proximity has been observed between several frequent translocation partners [156,157].

A number of recombination-predisposing motifs and non-B DNA-forming sequences have been found to be overrepresented at NHEJ breakpoints, indicative of the sequence-directed nature of many NHEJ-mediated rearrangements [158,159]. It has also been observed that at least one of the breakpoints of NHEJ-mediated rearrangements is often located within repetitive elements (such as LTRs, LINE or *Alu* elements) and sequence motifs capable of causing DSBs have been frequently

identified in the vicinity of the breakpoints of these NHEJ-mediated rearrangements [160]. The breakpoints of many nonrecurrent CNVs mediated by NHEJ map to LCRs, suggesting that LCRs can promote genomic instability by inducing certain chromatin secondary structures.

Replication-based mechanisms: Replication slippage or template switching during replication account for both small and large deletions and duplications with terminal microhomologies. Recently, relevant replication-based models including serial replication slippage (SRS) [161–163], fork stalling and template switching (FoSTes) [164], and microhomology-mediated break-induced replication [165] (MMBIR), which were collectively termed microhomology-mediated replication-dependent recombination by Chen et al. [166], have been used to explain the generation of a diverse range of complex genomic rearrangements [164,167].

DNA replication stalling-induced chromosome breakage has also been found to be an important mechanism causing deletions at chromosomal ends. Different types of telomeric deletions have been described [168]: type A terminal deletions are formed by chromosomal ends that are stabilized by the capture of a telomere from another source, whereas type B deletions are actually interstitial deletions toward the chromosomal ends. By contrast, type C deletions describe the process by which chromosomal ends are stabilized by telomere healing, namely the telomerase-dependent de novo addition of telomeres at nontelomeric sites. Terminal deletions associated with inverted duplications [169] can be classified as either type A or type C. Hannes et al. [170] cloned the breakpoints of nine chromosome 4p terminal deletions. All nine cases were shown to be type C terminal deletions. Bioinformatic analysis of the breakpoint-flanking regions involved in these nine cases, together with 12 previously fully characterized type C terminal deletions, led to the realization that there is an enrichment in secondary structure-forming sequences and replication stalling site motifs in these regions compared with a randomly selected sequence dataset [170].

Certain sequence features, such as microsatellites and transposon-rich regions, can serve to induce replication stalling, thereby acting as potential sources of genome instability [171,172]. On this basis, Koszul et al. [173] proposed a two-step mechanism to account for the generation of large segmental duplications: “First, a replication fork pauses and collapses generating a

chromosome breakage. Second, the double-strand break can be processed into a new replication fork either intra- or inter-molecularly by a break-induced replication-like mechanism that does not necessarily need a long sequence homology.” It was this “microhomology-dependent BIR” model that was subsequently deployed to explain disease-causing CNVs. In MMBIR, replication ends with the engagement of a misaligned template instead of reannealing to its original template; the synthesis of the second strand then follows the synthesis of the first (see review [166]). In practice, mutations due to SRS/FoSTes are often indistinguishable from those due to MMBIR. Indeed, the two terms have sometimes been used interchangeably [174,175].

Ankala et al. [176] have recently proposed another mechanism, the aberrant firing of replication origins, to explain a number of complex lesions in a series of 62 intragenic nonrecurrent rearrangements within various genes (mainly in the *DMD* gene). While repetitive sequence elements were noted in only four individual cases, microhomologies (2–10 bp) were observed at breakpoint junctions in 56% of the cases studied; further, insertions ranging from 1 to 48 bp were noted in 16 of the 62 cases. The sequence proximal to the breakpoints in six individual Duchenne muscular dystrophy (DMD) cases was characterized by tandem repetitions of short segments (5–20 bp). The repeated replication of template sequences proximal to the mapped deletion breakpoints was taken as evidence of attempts by the replication machinery to bypass a stalled replication fork. This mutational mechanism, based on the replication rescue model originally suggested by Doksani et al. [177], constitutes a novel type of template slippage event. Indeed, it can be seen that microhomologies at CNV breakpoints may be attributed to microhomology-mediated end joining (MMEJ), a replication repair mechanism, rather than to a recombination-based mechanism.

A recent review on the mechanisms underlying structural variant formation in genomic disorders [178] is also recommended to the reader.

6.3.7 Gross Deletions

Gross deletions are common causes of certain disorders and rare in others. In most of the X-linked disorders, for example, large deletions account for ~5% of molecular defects. In other disorders, however, such as steroid sulfatase deficiency, large deletions of the *STS* gene account

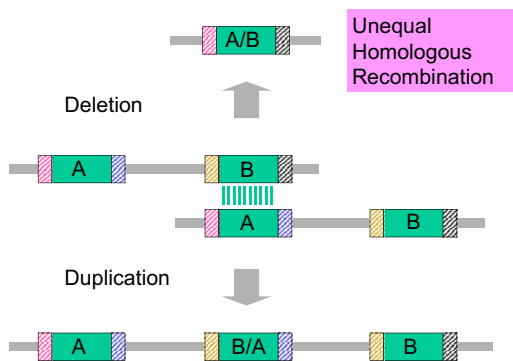


Figure 6.7 Homologous unequal recombination between similar regions of sequences A and B. The recombination events cause either deletions or duplications. In the case of a deletion, a hybrid sequence is generated with the first part from sequence A and the second from sequence B. The middle sequence in the duplication product is also a hybrid sequence; the first part is from sequence B and the second from sequence A.

for 84% of patients [179]. The same is true for disorders such as DMD, growth hormone deficiency, and α -thalassaemia [180–182].

A considerable number of large deletions probably are generated by mispairing of homologous sequences and unequal recombination (Fig. 6.7). One of the best examples of homologous unequal recombination is the case of α -globin genes on chromosome 16p. As a result of a recent evolutionary duplication of the α -globin genes, extensive regions of sequence homology exist between the two closely linked α -genes. Unequal crossover results in either the deletion of one α -gene or the creation of a fusion hybrid gene [183]. The reciprocal product chromosomes carry three α -genes and are not associated with a clinical phenotype [184]. Another example of a fusion gene resulting from an unequal crossover is the case of Hemoglobin Lepore, characterized by a hybrid gene between the δ - and β -globin genes on chromosome 11p [185]. In the case of steroid sulfatase deficiency, the deletion can be as large as 1 Mb [186]. In Kallmann syndrome, translocation can occur as a result of unequal mispairing of X and Y homologous sequences [187].

Several genetic disorders are due to large deletions (or duplications) caused by unequal crossing-over of homologous sequences. Fig. 6.8 depicts various examples that include a 1.5 Mb deletion of 17p12 in hereditary neuropathy with liability to pressure palsies (HNPP) [188], deletion of 1.5 Mb of 17q11.2 in neurofibromatosis type 1 [189], deletion of 1.6 Mb of

7q11.23 in Williams syndrome [190], deletion of 5 Mb of 17p11.2 in Smith–Magenis syndrome [191], deletion of either 3 Mb or more rarely 1.5 Mb of 22q11 in DiGeorge and velo-cardio-facial syndromes [192,193], and 4 Mb deletions of 15q in Prader–Willi and Angelman syndromes [194]. A recurrent deletion of ~ 0.5 Mb of 17q21.3, which may be mediated by a common inversion polymorphism, has also been described [195–198]. For a review of chromosomal “duplicons,” the LCRs that mediate deletions and duplications, see Ref. [199]. It has been estimated that $\sim 5\%$ of the human genome is duplicated either intra- or interchromosomally [200]. The large deletions or duplications (see below) due to duplicon crossover are also termed “genomic disorders.” A review of such genomic disorders may be found in Ref. [201].

In many cases of large deletion, homologous unequal crossover occurs between repetitive elements such as *Alu* sequences [202]. The *Alu* repeat is the most abundant repetitive element, with $\sim 1.5 \times 10^6$ copies in the human genome [202,203]. The element is ~ 300 bp in length and consists of two similar regions separated by a short A-rich region. Unequal crossover can occur between *Alu* sequences oriented in either the opposite or the same direction. In addition, unequal crossing over events have been noted between *Alu* elements and nonrepetitive DNA sequences without homology to *Alus*. The best examples of *Alu*-*Alu* recombination occur in the genes encoding the low-density lipoprotein receptor (*LDLR*) which underlies familial hypercholesterolemia and complement component 1 inhibitor (*C1I*; [204,205]). All but one of the breakpoints associated with *LDLR* gene deletions occur within *Alu* repeats. By contrast, deletions in other *Alu*-rich genes (e.g., *GLA1*) do not necessarily involve *Alu* repetitive elements [206]. This notwithstanding, *Alu*-mediated recombination between nonallelic *Alu* sequences is a fairly frequent cause of gene deletion causing human genetic disease [207–216]. It should be appreciated that the importance of *Alu* sequences in the context of mediating genomic deletions does not lie simply with their sheer abundance; *Alu* elements also possess inherent recombination–predisposing properties [217].

Nonhomologous (illegitimate) recombination occurs between two DNA sites that share minimal sequence homology of a few base-pairs. This type of recombination during meiosis or alternatively, slipped mispairing during DNA replication mediated by short (2–8)

Unequal crossover

Deletions / Duplications / Inversions

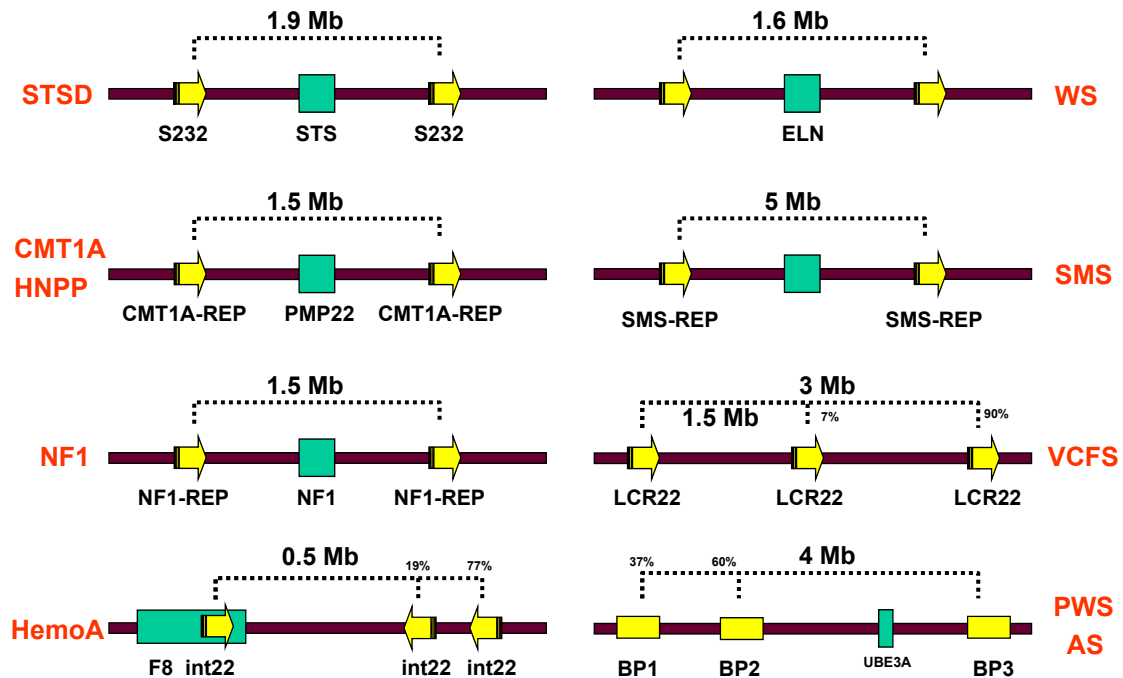


Figure 6.8 Genes, duplcons, and diseases. Unequal crossover between homologous sequences (duplcons) produce either deletions or duplications of the DNA between the duplcons. The duplcons are shown by *arrows* or by *clear boxes*. Genes included in the duplications/deletions are shown as *dark boxes*. *AS*, Angelman syndrome; *CMTA1*, Charcot-Marie-Tooth type A1; *HemoA*, hemophilia A; *HNPP*, hereditary neuropathy with liability to pressure palsies; *NF1*, neurofibromatosis 1; *PWS*, Prader-Willi syndrome; *SMS*, Smith-Magenis syndrome; *STSD*, steroid sulfatase deficiency; *VCFS*, velo-cardio-facial syndrome; *WS*, Williams syndrome.

nucleotide direct repeats flanking the deletions, is a common finding in many instances of large gene deletions [218]. Such deletions have been studied, for example, in hemophilia A; a compilation of 46 junctions from large deletions revealed that ~50% shared 2- to 6-bp homology at the breakpoint junction, as compared with only 17% in which the deletion was due to *Alu-Alu* recombination [219]. Similar results have been reported from the intron 7 deletion hotspot in the *DMD* gene; 8/9 deletion breakpoints examined were found to be flanked by DNA sequences with minimal homology [220].

It has also been proposed that alternative DNA conformations may trigger genomic rearrangements through recombination–repair activities. Distance measurements have indicated the significant proximity of alternating purine–pyrimidine and oligo(purine–pyrimidine) tracts to breakpoint junctions in 222 gross deletions and

translocations, respectively, involved in human diseases. In 11 deletions analyzed, breakpoints were explicable by non-B DNA structure formation [221].

The Gross Rearrangement Breakpoint Database (GRaBD; uwcm.ac.uk/uwcm/mg/grabd/) was established primarily for the analysis of the sequence context of translocation and deletion breakpoints in a search for characteristics that might have rendered these sequences prone to rearrangement [222]. GRaBD, which contains 397 germline and somatic DNA breakpoint junction sequences derived from 219 different rearrangements underlying human inherited disease and cancer, represents a large but not comprehensive collection of sequenced gross gene rearrangement breakpoint junctions. Analysis of these breakpoints has extended our understanding of illegitimate recombination by highlighting the importance of secondary structure formation between single-stranded

DNA ends at breakpoint junctions. For example, potential secondary structure was noted between the 5' flanking sequence of the first breakpoint and the 3' flanking sequence of the second breakpoint in 49% of rearrangements, and between the 5' flanking sequence of the second breakpoint and the 3' flanking sequence of the first breakpoint in 36% of rearrangements [159]. In addition, deletion breakpoints were found to be AT-rich, whereas translocation breakpoints were GC-rich. Alternating purine-pyrimidine sequences were found to be significantly overrepresented in the vicinity of deletion breakpoints while polypyrimidine tracts were overrepresented at translocation breakpoints [158].

Finally, several examples of pathogenic large genomic deletions caused by the prior L1-mediated insertion of L1 [223,224], *Alu* [225,226], or SVA [227] insertions (see later) have been reported, as well as the first cases of L1-driven pseudogene insertion causing human genetic disease [228,229].

6.3.8 Large Retrotranspositional Insertions

A less common but nevertheless still fascinating mechanism of human gene mutation is the de novo insertion of repetitive elements via retrotransposition. The

phenomenon was first observed in humans in the factor VIII (*F8*) gene in two unrelated de novo cases of severe hemophilia A [230]. Truncated LINE (long interspersed) repetitive elements were introduced into exon 14 of the factor VIII (*F8*) gene where they caused disruption of the reading frame. The inserted elements contained a poly(A) tract and caused a target site duplication of >12 nucleotides. Further analysis of these insertions revealed that, in one case, the inserted element was an exact but truncated copy of a full-length LINE element with open reading frames (ORFs) found at chromosome 22q11 [231]. The master source gene produces an mRNA that is probably reverse transcribed (possibly via a reverse transcriptase encoded by itself) and the double-stranded nucleic acid is then reinserted into an A-rich region of the genome (Fig. 6.9). LINEs probably integrate into genomic DNA by a process called target-primed reverse transcription [232]. The proposed mechanism of LINE retrotransposition is as follows: An active LINE is transcribed in the nucleus and is subsequently transported to, and translated in, the cytoplasm. The two LINE-encoded proteins, ORF1 and ORF2, complex with LINE transcripts in ribonucleoprotein particles. The complexes are then transported

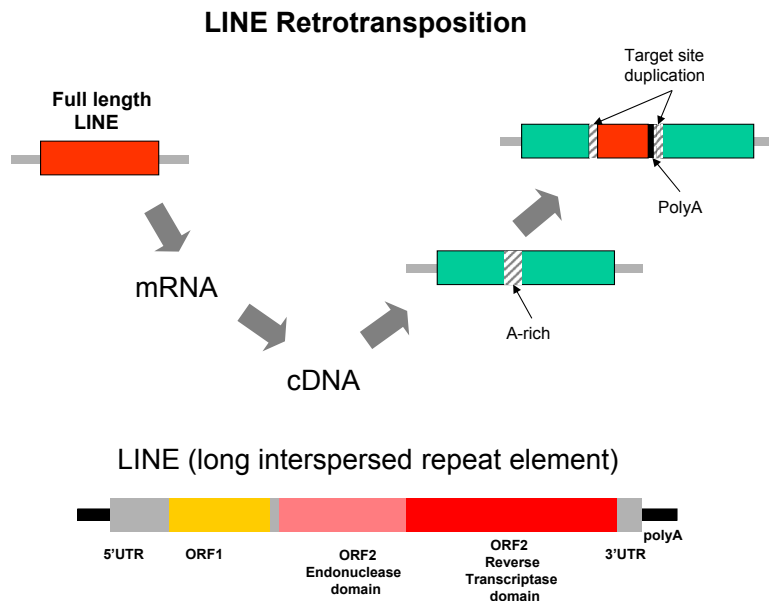


Figure 6.9 Schematic representation of LINE retrotransposition. A master retrotransposon (full length LINE) from one chromosomal location is transcribed to mRNA; then reverse transcribed to double-stranded DNA and inserted into an adenine-rich region of another chromosomal location. The transposon has a poly(A) tail and produces a target site duplication.

to recipient DNA sequences where target-primed reverse transcription occurs. The new, integrated LINE copy is usually truncated at its 5' end. Over evolutionary time, L1s have shaped mammalian genomes through a number of different mechanisms. First, they have greatly expanded the genome both by their own retrotransposition and by providing the machinery necessary for the retrotransposition of other mobile elements, such as *Alu* sequences or SVA elements [163]. Second, they have shuffled non-L1 sequence throughout the genome by a process termed transduction. Accidents of retrotransposition can cause disease, and a number of such insertions have been reported to date [232,233]. It is noteworthy that insertions of these elements within introns of genes or flanking regions are probably not associated with disease, but instead represent rare, private polymorphisms [234].

Similar retrotranspositions that involve members of the *Alu* sequence family have also been reported in several genes (examples include *Alu* insertions into the *NF1* gene causing type 1 neurofibromatosis, into the factor IX (*F9*) gene causing hemophilia B, and into the cholinesterase (*BCHE*) gene in a case of acholinesterasemia; [235–237]). It is likely that LINES provide the molecular machinery necessary for the retrotransposition of *Alus*. One study using mutation analysis of the *F9* gene has estimated the frequency of retrotransposition to be such that it occurs somewhere in the genome of ~1 in every 17 children born [238].

Some 17% of a collection of gross insertions, all ≥ 276 bp in length, were due to LINE-1 (L1) retrotransposition involving different types of elements (L1 *trans*-driven *Alu*, L1 direct, and L1 *trans*-driven SVA) [163]. A meta-analysis of 48 recent L1-mediated retrotranspositional events known to have caused human genetic disease revealed that 26 were L1 *trans*-driven *Alu* insertions, 15 were direct L1 insertions, 4 were L1 *trans*-driven SVA insertions, and 3 were associated with simple poly(A) insertions [239]. The systematic study of these lesions, when combined with previous in vitro and genome-wide analyses, allowed several conclusions regarding L1-mediated retrotransposition to be drawn: (a) ~25% of L1 insertions are associated with the 3' transduction of adjacent genomic sequences, (b) ~25% of the new L1 inserts are full-length, (c) poly(A) tail length correlates inversely with the age of the element, and (d) the length of target site duplication in vivo is rarely longer than 20 bp. This analysis also suggested

that some 10% of L1-mediated retrotranspositional events are associated with significant genomic deletions in humans. Chen et al. [240] reported an indel in the *CFTR* gene that involved the insertion of a short 41-bp sequence with partial homology to a retrotranspositionally competent LINE-1 element. These authors dubbed such insertions of ultra-short LINE-1 elements “hyphen elements.”

Several instances of the clustering of pathogenic L1-mediated insertion events have also been observed. Thus, three independent *Alu* insertions have been found to be integrated into a 104 bp region of the *FGFR2* gene [241,242], two independent L1 insertions have been reported to have inserted into an 89 bp region of exon 44 of the dystrophin (*DMD*) gene [243,244], while six different insertions were found in a 1.5-kb region of the *NF1* gene [245]. It should also be noted that independent L1-retrotransposition elements can integrate at precisely the same chromosomal sites. Thus, two markedly different *Alu* Ya5a2 elements became integrated at precisely the same site in the *F9* gene causing severe hemophilia B [236,246], whereas an SVA element and an *Alu* sequence were inserted at the same site within the coding region of the *BTK* gene [247]. These observations are consistent with some genomic locations being exquisitely prone to L1-retrotransposition [239].

6.3.9 Large Insertion of Repetitive and Other Elements

The insertion of nonretrotransposons, namely β -satellite repeats, has been observed in the human genome. The insertion of 18 copies of the 68-bp monomer of the β -satellite repeat in exon 11 of the *TMPRSS3* gene on chromosome 21 caused one form of recessive nonsyndromic deafness DFNB10 [248]. This may have been mediated by invasion of the genomic DNA by a small polydispersed circular DNA.

A patient with a sporadic case of Pallister–Hall syndrome has been shown to have experienced a de novo nucleic acid transfer from the mitochondrial to the nuclear genome. This variant, a 72-bp insertion into exon 14 of the *GLI3* gene, creates a premature stop codon and predicts a truncated protein product. Both the mechanism and the cause of the mitochondrial-nuclear transfer are however, unknown [249]. Further examples of pathological mitochondrial-nuclear sequence transfers have been subsequently identified in the *USH1C* gene [163] and the *PAFAH1B1* gene [250].

Gross insertions (>20bp) comprise <1% of disease-causing variants. In an attempt to study these insertions in a systematic way, 158 gross insertions ranging in size between 21bp and ~10kb were identified from the HGMD; their study has revealed extensive diversity in terms of the nature of the inserted DNA sequence and has provided new insights into the underlying mutational mechanisms [163]. Some 70% of gross insertions were found to represent sequence duplications of different types (tandem, partial tandem, or complex). In the context of a 26-bp insertion into the *ERCC6* gene, the authors also speculated as to whether they had found evidence for another mechanism of human genetic disease, involving the possible capture of DNA oligonucleotides [163].

6.3.10 Inversions

The most common inversion found to date is that associated with the factor VIII (*F8*) gene, which occurs via intrachromosomal recombination mediated by a 9.5-kb sequence that is repeated three times in the last megabase of Xqter; once in intron 22 of the *F8* gene and twice ~400 kb telomeric to the first [251,252] (Fig. 6.10). Most inversions, which are high-frequency independent recurring events, involve the distal sequence. The vast

majority of inversions occur in male germ cells [253], perhaps because intrachromosomal recombination is inhibited by the presence of homologous X chromosomes (the male:female ratio was estimated to be ~300:1). Almost all mothers of inversion hemophilia A cases are carriers of the abnormality. DNA diagnosis of the molecular lesion in severe hemophilia A has been greatly facilitated by the frequent occurrence of this common inversion of the *F8* gene (45% of individuals with severe hemophilia A). The frequency of de novo *F8* gene inversion has been estimated to be 7.2×10^{-6} per gamete per generation. Another example of inversion has been described in the *IDS* gene (also on Xq) in ~13% of cases of Hunter syndrome [254]. Inversions of DNA sequences have also been reported in the β -globin gene cluster on 11p and in the *APOA1-APOC3-APOA4* gene cluster on 11q [255,256].

A meta-analysis of inversions of ≥ 5 bp but <1kb has been performed by Chen et al. [162]. Of the 21 mutations studied, 19 were found to be compatible with a model of intrachromosomal serial replication slippage in *trans* (SRStrans) mediated by short inverted repeats. Eighteen (one simple inversion, six inversions involving sequence replacement by upstream or downstream sequence [DSS],

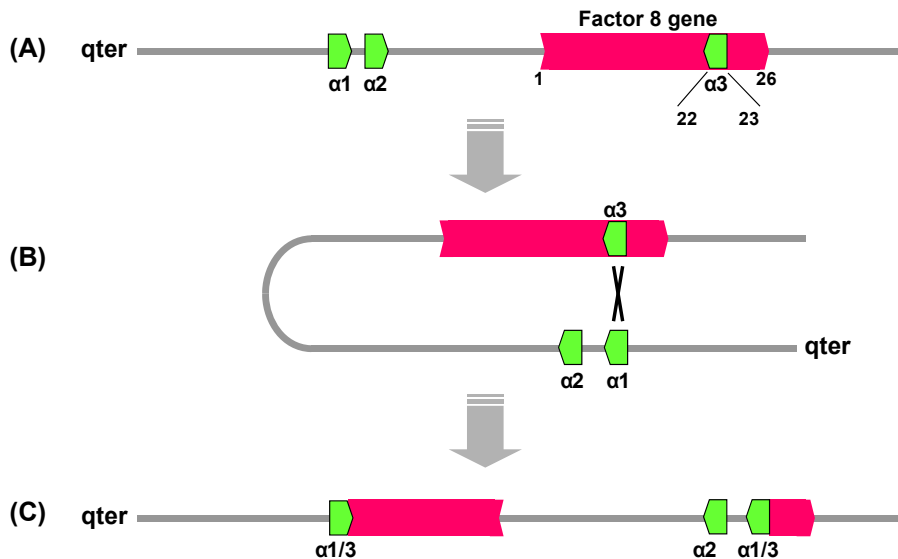


Figure 6.10 Common inversion of the factor VIII (*F8*) gene in severe hemophilia A. (A) Schematic representation of the most distal 1 Mb of Xq. Regions α_1 , α_2 , and α_3 represent 9.5-kb highly homologous DNA elements. The orientations of these sequences are shown by arrows. (B) Intrachromosomal recombination between elements α_1 and α_3 . (C) The crossover results in the inversion of exons 1–22 of the *F8* gene.

five inversions involving the partial reinsertion of removed sequence, and six inversions that occurred in a more complicated context) of these were found to be consistent with either two steps of intrachromosomal SRStrans or a combination of replication slippage in *cis* plus intrachromosomal SRStrans. The remaining lesion, a 31-kb segmental duplication associated with a small inversion in the *SLC3A1* gene, was explicable in terms of a modified SRS model incorporating the concept of “break-induced replication.” This study has, therefore, lent broad support to the idea that intrachromosomal SRStrans can account for a variety of complex gene rearrangements involving inversions.

6.3.11 Duplications

Duplications of whole genes or exons have contributed very significantly to the evolution of the human genome [56]. Indeed, most gene clusters (e.g., β -globin, growth hormone, Hox) owe their origin to gene duplications that have occurred during vertebrate evolution. Furthermore, the presence of similar domains in proteins (e.g., immunoglobulin-like domains in many transmembrane proteins) is due to duplications of certain exons.

Occasionally, however, duplications may also be the cause of genetic disorders. The most frequent mechanism of duplication is homologous unequal crossover as described for large deletions. In fact, most large duplications generated as the reciprocal product of a deletion resulting from homologous unequal crossover. Duplications are less common, however, than their theoretically reciprocal deletions (e.g., see Ref. [257], for the *DMD* gene). This may be due to the nonpathogenicity of a duplication (e.g., α -globin genes [184]), elimination of duplications as is the case for the *HPRT1* gene, or to the fact that not all mechanisms that lead to deletions also produce duplications. A large and common duplication has been identified in cases of Charcot–Marie–Tooth disease type 1A [258]. This duplication involves 1.5 Mb of DNA on chromosome 17p containing the peripheral myelin protein 22 (*PMP22*) gene. It results from homologous unequal crossover events between 24-kb repeats that flank the duplicated region. The reciprocal deletion product of this recombination event is responsible for a completely different clinical phenotype: HNPP (Fig. 6.8). Another notable duplication of at least 500 kb that includes the *PLP1* gene is a frequent cause of Pelizaeus–Merzbacher disease [259]. The pathogenetic mechanism of these duplications involves unequal crossing-over in meiosis mediated by “duplicons” in the genome [199].

The molecular defect in the majority of cases with ectrodactyly type SHFM3 on chromosome 10q24, is an approximately 0.5-Mb tandem duplication. The precise pathogenetic mechanism of this duplication is unknown [260]. Additional gene duplications causing recognizable syndromes include the *APP* duplication causing early-onset Alzheimer disease [261], the *SNCA* duplication and Parkinson disease [262], and the triplication of ~605-kb segment containing the *PRSS1* gene in families with hereditary pancreatitis [263].

6.3.12 CNV in Association With Disease

CNVs are a form of genomic diversity that involves DNA sequences ≥ 1 kb in length that are present in the human genome in a variable number of copies. CNVs may be recurrent (which arise by NAHR) or nonrecurrent (which arise by nonhomologous end joining and replication-based mechanisms) [264]. Such gross duplications/deletions not only are rather abundant but also often occur at polymorphic frequencies. Conrad et al. [62] generated a comprehensive map of >8500 validated CNVs >500 bp (detected in 41 Europeans/West Africans) that together cover a total of 112.7 Mb (3.7% of the genome). These authors estimated that 39% of the validated CNVs overlapped 13% of RefSeq genes (NCBI mRNA reference sequence collection). Further, they concluded that the CNVs detected resulted in the “unambiguous loss of function” of alleles for 267 different genes.

It has been estimated that on average, 73–87 genes vary in terms of their copy number between any two individuals [265]. This high degree of interindividual variability with regard to gene copy number has challenged traditional definitions of wild type and “normality” and even the very concept of a “reference genome” itself. High-resolution breakpoint mapping is a prerequisite for the accurate assessment of CNV size, the identification of the genes and regulatory elements affected, and hence, for the determination of the consequences of CNV for gene expression [266] and the phenotypic sequelae [267–270]. This notwithstanding, it is already becoming clear that these consequences may go far beyond the physical bounds of a given CNV. For example, a CNV involving the human *HBA* gene has a dramatic influence on the expression of the *NME4* gene some 300 kb distant [271]. In addition, a 5.5-kb microduplication of a conserved noncoding sequence with demonstrated enhancer function, ~110 kb downstream

of the bone morphogenic protein 2 (*BMP2*) gene, has been found to cause brachydactyly type 2A in two families [272].

It may well be that the precise extent and/or location of many CNVs will vary between individuals, thereby further increasing both the mutational and phenotypic heterogeneity. The extent to which CNVs are likely to contribute to the diversity of human phenotypes, including “single gene defects,” genomic disorders, and complex disease, is increasingly being recognized. Thus, CNV of the *FCGR3B* genes is a determinant of susceptibility to immunologically mediated glomerulonephritis [273]. CNVs in the *CCL3L1* and *DEFB4* genes have also been found to be associated with increased susceptibility to HIV infection and Crohn disease, respectively [274,275], whereas rare CNVs associated with various complex phenotypes have been identified in studies of schizophrenia [276], epilepsy, and severe early-onset obesity [277]. CNVs are now being widely recruited to genome-wide association studies (GWAS) with the aim of assessing their influence on human disease causation/susceptibility [267,278,279].

To date, several dozen human disease conditions have been identified, which are either caused by CNVs or whose relative risk is increased by CNVs [267,280]. Remarkably, an excess of both rare and de novo CNVs has been identified in patients with psychiatric disorders and obesity [276,277,281–285]. These findings point to genetic heterogeneity in these conditions, thereby illustrating the likely complexity inherent in identifying all disease-causing CNVs. Shlien et al. [286] reported a highly significant increase in CNV number among patients with Li–Fraumeni syndrome, carriers of inherited *TP53* variants. Hence, it would appear that heritable genetic variants have the potential to modulate the rate of germline CNV formation.

It is already clear that the disease relevance of CNVs represents a continuum, stretching from “neutral” polymorphisms on the one hand to directly pathogenic copy number changes on the other [267]. Between these two extremes may lie those CNVs that are capable of acting as predisposing (or protective) factors in relation to complex disease [287,288]. Thus, for example, a 117-kb deletion encompassing the *UGT2B17* gene has been found to be associated with an increased risk of osteoporosis [289]. Some germline CNVs appear to predispose to disease even although no known genes reside within their boundaries [290,291]. Importantly, a 520-kb

microdeletion has been identified at 16p12.1, which predisposes to various neuropsychiatric phenotypes as a single copy number variant and aggravates neurodevelopmental disorders if it co-occurs together with other large deletions and duplications [288]. It remains to be seen whether “CNV equivalents,” <1 kb in size, which actually occur more frequently than true CNVs (>1 kb) [62], will also be relevant to disease. What is already clear is that, over the coming years, an increasing number of important CNV-disease associations are going to come to light [270].

6.3.13 Gene Conversion

Gene conversion is the modification of one of two alleles by the other. It involves the nonreciprocal correction of an “acceptor” gene or DNA sequence by a “donor” sequence, which itself remains physically unchanged. In most known instances of gene conversion as a cause of human genetic disease, the functional gene has been wholly or partially converted to the sequence of a highly homologous and closely linked pseudogene, which therefore acts as the donor sequence [292]. Probable examples include the genes for steroid 21-hydroxylase (*CYP21* [293]), polycystic kidney disease (*PKD1* [294]), neutrophil cytosolic factor p47-*phox* (*NCF1* [295]), immunoglobulin λ -like polypeptide 1 (*IGLL1* [296]), glucocerebrosidase (*GBA* [297]), von Willebrand factor (*VWF* [298]), and phosphomannomutase (*PMM2* [299]). These gene/pseudogene pairs are all closely linked with the exception of the *VWF* gene (12p13) and its pseudogene (22q11–q13) and the *PMM2* gene (16p13) and its pseudogene (18p). Together, these two exceptions would seem to establish a precedent for the occasional occurrence of gene conversion between unlinked loci in the human genome.

An in silico analysis of the DNA sequence tracts involved in 27 well-characterized nonoverlapping gene conversion events in 19 different genes reported in the context of inherited disease was recently performed [300]. It was noted that gene conversion events tended to occur within (C+G)- and CpG-rich regions and that sequences with the potential to form non-B DNA structures (and which might be involved in the generation of double-strand breaks that could, in turn, serve to promote gene conversion) occurred disproportionately within maximal converted tracts and/or short flanking regions. Maximal converted tracts were also found to be enriched in a truncated version of the chi-element

(a TGGTGG motif), immunoglobulin heavy chain class switch repeats, translin target sites, and several novel motifs including (or overlapping) the classic meiotic recombination hotspot, CCTCCCCT [300]. Finally, it was found that gene conversions tended to occur in genomic regions that had the potential to fold into stable hairpin conformations. Taken together, these findings support the concept that recombination-inducing motifs, in association with alternative (non-B DNA) conformations, can promote recombination in the human genome.

The large number of duplicated gene sequences in the human genome implies that a considerable number of disease-associated variants could originate via interlocus gene conversion. A genome-wide computational approach to identify disease-associated variants derived from interlocus gene conversion events recently revealed hundreds of known pathological variants that could have been caused by interlocus gene conversion [301]. In addition, several dozen high-confidence cases of inherited disease variants resulting from interlocus gene conversion were identified in ~1% of all genes analyzed. About half of the donor sequences associated with such variants were functional paralogous genes, suggesting that epistatic interactions or differential expression patterns would determine the impact upon fitness of a single amino acid substitution between duplicated genes. In addition, Casola et al. [301] identified thousands of hitherto undescribed deleterious variants that could potentially arise via interlocus gene conversion. It would therefore appear that the impact of interlocus gene conversion upon the spectrum of human inherited disease may be considerably greater than has hitherto been appreciated.

Although variants that are detrimental to the fitness of individuals are expected to be rapidly purged from the population by natural selection, some pathological variants are nevertheless retained at high frequencies in human populations. Several hypotheses have been proposed to account for this apparent paradox (high new mutation rate, genetic drift, overdominance, or recent changes in selective pressure). However, there is an additional process that appears to contribute to the spreading of deleterious variants: GC-biased gene conversion (gBGC), a process associated with recombination that tends to favor the transmission of GC-alleles over AT-alleles. Necsulea et al. [302] have shown that the spectrum of amino acid-altering polymorphisms

in human populations exhibits the footprints of gBGC. This pattern is not explicable in terms of selection and is evident with all nonsynonymous variants, including those predicted to be detrimental to protein structure and function as well as those that have been implicated in the causation of human genetic disease. These results indicate that gBGC meiotic drive contributes to the spreading of deleterious variants in human populations.

6.3.14 Insertion–Deletions (Indels)

A relatively rare type of mutation causing human genetic disease is the *indel*, a complex lesion that appears to represent a combination of microdeletion and microinsertion. One example is the nine deleted base pairs encoding codons 39–41 of the $\alpha 2$ -globin (*HBA2*) gene that were replaced by eight inserted bases that serve to duplicate the adjacent downstream sequence (DSS) [303]. Indels constitute a fairly infrequent type of lesion causing human genetic disease; ~1.5% of lesions in HGMD fall into this category.

Several indel hotspots have been noted in a meta-analysis of HGMD data on 211 different indels underlying genetic disease [304]. A GTAAGT motif was found to be significantly overrepresented in the vicinity of the indels studied. The change in complexity consequent to a mutation was also found to be indicative of the type of repeat sequence involved in mediating the event, thereby providing clues as to the underlying mutational mechanism. The majority of indels (>90%) were explicable in terms of a two-step process involving established mutational mechanisms. Indels equivalent to double base-pair substitutions (22% of the total) were found to be mechanistically indistinguishable from the remainder and may therefore be regarded as a special type of indel.

6.3.15 Other Types of Complex Rearrangement

Complex mutational events that involve combined gross duplications, deletions, and/or insertions of DNA sequence have been not infrequently observed and together constitute ~1% of entries in HGMD. One example of this type of gene defect is a 10.9-kb deletion coupled with a 95-bp inversion in the factor IX (*F9*) gene causing hemophilia B [305]. The molecular characterization of this type of lesion is often extremely complicated, and in most cases, the underlying mutational mechanisms could not be readily inferred.

Recently, however, a meta-analysis of 21 complex gene rearrangements derived from the HGMD revealed that all but one could be accounted for by a model of SRS, involving twin or multiple rounds of replication slippage [161]. Thus, of the 20 complex gene rearrangements, 19 (seven simple double deletions, one triple deletion, two double mutational events comprising a simple deletion and a simple insertion, six simple indels that may constitute a novel and noncanonical class of gene conversion, and three complex indels) were compatible with the model of SRS in *cis*; by contrast, the remaining indel in the *MECP2* gene appears to have arisen via interchromosomal replication slippage in *trans*.

A novel type of complex genomic rearrangement, comprising intermixed duplications and triplications of genomic segments, has recently been described at both the *MECP2* and *PLP1* loci [306]. These complex rearrangements share a common genomic organization, viz., a duplication-inverted triplication-duplication (DUP-TRP/INV-DUP), in which the triplicated segment is inverted and located between directly oriented duplicated genomic segments. The DUP-TRP/INV-DUP structures appear to be mediated by inverted repeats, up to >300 kb apart.

6.3.16 Multiple Simultaneous Mutations

Transient hypermutability is a general mutational mechanism with the potential to generate multiple synchronous mutations, a phenomenon probably best exemplified by “closely spaced multiple mutations” (CSMMs). From a collection of human inherited disease-causing multiple mutations, Chen et al. [307] retrospectively identified numerous potential examples of pathogenic CSMMs that exhibited marked similarities to the CSMMs reported in other systems. These examples included (1) eight multiple mutations, each comprising three or more components within a sequence tract of <100 bp (*CBS*, *MPZ*, *OPN1LW*, and *STK1* genes), (2) three possible instances of “mutation showers” in the *PTCH1*, *FANCA*, and *KNG1* genes, respectively, and (3) numerous highly informative “homocoordinate” mutations (multiple mutations involving the same mutation type).

Recently, a remarkable phenomenon has been reported in multiple cancer samples: The presence of tens to hundreds of genomic rearrangements involving spatially localized genomic regions [308]. These complex rearrangements primarily affected a single

chromosome, although in some cases, multiple apparently concomitant alterations affected several different chromosomes. Stephens et al. convincingly argued that these massive, yet spatially localized, genomic rearrangements must have resulted from a single catastrophic event (which they termed “chromothripsis”) rather than from a series of progressive and hence independent alterations. Chromothripsis also appears to be capable of explaining the generation of some complex de novo structural rearrangements in the germline [309]. An illustrative example pertains to a highly complex chromosomal rearrangement, identified in a child with severe congenital abnormalities, which comprised at least 12 de novo breakpoint junctions and involved chromosomes 1, 4, and 10. These breakpoints were clustered in small genomic regions of up to 3.5 Mb in size on each chromosome. Reconstruction of the derivative chromosomes indicated that the breakpoints formed concordantly oriented pairs on the reference genome. Both intra- and interchromosomal junction sequence features were compatible with those commonly associated with NHEJ [309]. The insights generated from the seminal work of Stephens and his colleagues may well help us to understand the mutational mechanisms underlying some previously reported germline complex rearrangements [310]. For example, with respect to the de novo mutational event on chromosome 2 in a patient with Waardenburg syndrome and other congenital defects, the original authors suggested that all five breaks might have occurred simultaneously but were unable to explain why the breakpoints had occurred within a single chromosome [311]. In the light of our emerging knowledge of chromothripsis, the idea that this complex rearrangement could have been generated in such a way as to be compatible with the NHEJ repair of simultaneously generated DSBs becomes quite attractive. Liu and colleagues subsequently investigated 17 subjects with various development abnormalities by means of high-resolution genome analysis [312]. Constitutional multiple copy number changes, including deletions, duplications, and/or triplications, as well as inversions were observed in all cases. Strikingly, in each case, all rearrangements occurred within a single chromosome; in 15 of the 17 cases, the rearrangements were localized to the distal half of the affected chromosomal arms. FISH and breakpoint junction data indicated that all additional copies of the duplicated and triplicated

segments appear to be randomly joined, forming a large “breakpoint junction cluster” on 9q21. By analogy with the phenomenon of chromothripsis, the observation of these extremely complex rearrangements in a single chromosome was also described as a chromosome catastrophe event [312]. However, this kind of chromosomal change cannot be easily explained by the previously described NHEJ repair of simultaneously generated DSBs. Instead, Liu and colleagues envisaged the involvement of a replicative mechanism in the generation of this complex chromosome catastrophe event comprising multiple duplications and/or triplications; they regarded MMBIR as the most likely underlying mechanism. They further suggested that a potential replication fork collapse at 9q21 could account for the breakpoint clustering therein [312]. A catalogue of published cases of germline chromothripsis has been provided in ref [313]. Most affected individuals present with developmental delay and dysmorphic features.

6.3.17 Molecular Misreading

Long runs of adenines (and perhaps other mononucleotides or dinucleotides) promote a phenomenon termed “molecular misreading” by which DNA replication/RNA transcription and/or translation result in erroneous products with different numbers of (A)s derived from the original DNA sequence [314]. In a family with hypobetalipoproteinemia, a deletion of one C in the A₅CA₃ coding sequence of the *APOB* gene results in a run of (A)₈. The affected individual, however, did not have severe disease, because some ApoB protein was made. This was the result of molecular misreading in which ~10% of the resulting mRNAs contained (A)₉ instead of the expected (A)₈; this partially restored the reading frame thereby templating the synthesis of low amounts of normal ApoB [315]. Similarly, a family with mild to moderately severe hemophilia A with a deletion of one T within the coding A₈TA₂ sequence of the *F8* gene has been reported. The partial “correction” of the phenotype was due to restoration of the reading frame because of molecular misreading in which ~5% of the resulting RNAs contained (A)₁₁ instead of the expected (A)₁₀. In this family, there was also evidence for ribosomal frameshifting during translation of the mutant RNA [316].

Another example of this phenomenon was observed in the *APC* gene. A T-to-A transversion is present in the coding A₃TA₄ sequence of the *APC* gene in 6% of

Ashkenazi Jews, and in ~28% of Ashkenazim with a family history of colorectal cancer. This variant creates a small hypermutable region, indirectly causing cancer predisposition because there are many somatic cells in which stretches of (A)₉ occur instead of the expected (A)₈; the (A)₉ results in frameshifting and a truncated dysfunctional APC [317]. Interestingly, in the neurofibrillary tangles, neuritic plaques, and neuropil threads in the cerebral cortex of Alzheimer disease and Down syndrome, abnormal forms of β -amyloid precursor protein and ubiquitin B have been observed. These aberrant proteins were produced because of +1 frameshifting that resulted from a deletion of AG in a sequence GAGAG that occurred in the coding regions of both genes (*APP* and *UBB*, respectively). This dinucleotide deletion was again the result of molecular misreading during transcription or posttranscriptional editing of RNA [318]. This mechanism is likely to yield a considerable quantity of abnormal RNA molecules and protein products in somatic cells [319].

6.3.18 Germline Epimutations

Epimutations are modifications of DNA that constitute clonally heritable (yet potentially reversible) alterations in the transcriptional status of a gene that lead to the abnormal silencing of that gene. Epimutations are not mutations *sensu stricto* because they do not alter the gene’s nucleotide sequence; however, germline epimutations of the *MLH1* gene have been reported in individuals with multiple cancers [320] and in the *MLH1* and *MSH2* genes in hereditary nonpolyposis colorectal cancer [321]. These heritable inactivating epimutations are characterized by monoallelic hypermethylation of the *MLH1* or *MSH2* genes and, to all intents and purposes, are functionally equivalent to conventional mutations. A maternal epimutation in the *GNAS* gene has also been reported as a cause of Albright osteodystrophy and parathyroid hormone resistance [322]. With the determination of the human methylome [91] and the recent recognition that DNA sequence polymorphisms can exert an effect on gene function via allele-specific methylation in *cis* [323], the number of recognized epimutations should rise quite significantly in the coming years. If eventually shown to be both of pathological significance and heritable, some examples of histone modification [324,325] or RNA editing [326,327] could also turn out to represent “honorary mutations.”

6.3.19 Frequency of Disease-Producing Variants

Mutation frequency within genes: The frequency of different molecular defects is not the same for every gene and every disorder. Indeed, human disease genes exhibit very considerable allelic heterogeneity in terms of their mutational spectra; for some genes, a few predominant disease alleles predominate whereas for others, there is a wide range of disease alleles, each relatively rare [328]. The mutational spectrum depends very largely on the DNA sequence characteristics of the gene in question (e.g., the presence of repeat units or homologous sequences), and the function of, and evolutionary constraints experienced by, its encoded protein [329]. For some genes, deletions predominate; for others, one particular type of lesion such as an inversion may be especially common. Some genes exhibit mainly frameshifts and stop codons associated with a specific disorder, whereas others manifest mainly missense variants for a given phenotype, or expansions of trinucleotide repeats.

Disease variants are nonuniformly distributed within genes [330]. Such variants were found to be statistically overrepresented in conserved domains, and underrepresented in variable regions, even after allowing for the amino acid site variability of domains over long-term evolutionary history. This finding suggests that there is a nonadditive influence of amino acid site conservation on the observed intragenic distribution of disease variants.

Mutation frequency within human populations: Population genetic considerations are also likely to be very important in determining why some variants occur frequently, either within a patient cohort or in the population at large (see Frequency of Inherited Disorders Database, <http://archive.uwcm.ac.uk/uwcm/mg/fidd/>; FINDbase, <http://www.findbase.org/>). Selection, migration, and genetic drift are all likely to play a role as well as the mutation rate [331–333]. Thus, the mutational spectrum of the *PAH* gene underlying phenylketonuria appears to result from a range of different factors including founder effect, range expansion and migration, genetic drift, and possibly also heterozygote advantage [334]. Selection can also serve to maintain deleterious variants at high frequencies in particular populations by overdominant selection (heterozygote advantage). Good examples of this phenomenon are provided by a reduction in risk of severe malaria associated with female heterozygotes and male hemizygotes for variants in the

X-linked *G6PD* gene [335,336], for individuals heterozygous for the β -globin (*HBB*) sickle cell variant, Glu6Val [337], and for individuals heterozygous and homozygous for α^+ -thalassemia [338]. Intriguingly, however, the protection against malaria afforded by sickle cell disease and α^+ -thalassemia when inherited individually is lost when the two conditions are coinherited [339]. Other possible examples of heterozygote advantage include an elevated cortisol response in heterozygous carriers of *CYP21A* variants [340], higher values for hemoglobin, serum iron, and transferrin saturation in women heterozygous for *HFE* gene variants [341], resistance to prion infection conferred by a common prion protein (*PRNP*) polymorphism [342], resistance to severe sepsis in heterozygous carriers of the factor V Leiden polymorphism, Arg506Gln [343], and increased keratinocyte cell survival in individuals heterozygous for *GJB2* gene variants [344]. Resistance to cholera toxin [345], protection against bronchial asthma [346], and resistance to *Pseudomonas aeruginosa* infection [347] have all been mooted as possible bases for overdominant selection in heterozygous carriers of *CFTR* gene variants. However, cystic fibrosis heterozygotes have been shown to secrete chloride at the same rate as individuals lacking *CFTR* gene variants [348].

A number of genetic diseases are known to be particularly prevalent in Jewish populations [349,350]. The presence of four distinct lysosomal storage diseases at significant frequencies among Ashkenazi Jews has often been considered as providing evidence for a selective advantage accruing to heterozygotes in this population. However, evidence in support of the idea of genetic drift appears to be more compelling [351,352].

Selection may also act at an extremely early stage to boost the frequency of some variants that are deleterious at a later stage in development. For example, gain-of-function missense variants in the fibroblast growth factor receptor 2 (*FGFR2*) gene responsible for Apert syndrome have been shown to confer a selective advantage on spermatogonial cells by promoting the clonal expansion of mutant cells [353,354].

6.3.20 Functional Characteristics of Human Disease Genes

Human disease genes appear to be distinguishable from “nondisease genes” (in reality, the latter can only be defined as genes that are not yet known to cause inherited disease) in terms of a range of features including

gene structure, gene expression, physicochemical properties, protein structure, and evolutionary conservation [38,329,355–360]. Thus, human disease genes are characterized by the greater length of their encoded amino acid sequences, a larger number of longer introns, a broader range of tissue expression, and a wider phylogenetic distribution [360,361]. Human disease genes are also known to be unevenly distributed between the human chromosomes [362,363]. Further, synonymous nucleotide substitutions appear to occur at a higher rate in human disease genes, a finding that may reflect increased mutation rates in the chromosomal regions in which disease genes are found [363]. It may be that disease genes are more prevalent in genomic regions that experience elevated rates of mutation [364]. Another possible explanation is that the disease gene set may contain a disproportionately lower number of genes expressed in the germline [363]. This is because variants in such genes might be expected to be more effectively repaired by transcription-coupled repair (transcription-coupled repair in the germline appears to account for the strand asymmetry that the human genome exhibits in terms of inherited variants; [365,366]). Strand asymmetries with respect to the mutation rate may, however, also arise through the influence of DNA replication origins [367], recombination [368,369], and strand-biased repair [370].

6.3.21 Mutation/Variant Nomenclature

Some consistency in the way in which variants are described is essential for the accurate and unambiguous reporting and curation of genomic variation data. Most guidelines on how to describe mutational changes in human genes are to be found in [371,372] and on the Human Genome Variation Society (HGVS) Website (hgvs.org/mutnomen/). A program, *Mutalyzer*, is available to check sequence variation nomenclature (lovd.nl/mutalyzer) using a human genome reference sequence and following the current recommendations of the HGVS; *Mutalyzer* is capable of handling most types of variants, including nucleotide substitutions, deletions, duplications, insertions, indels, and splice-site changes [373].

6.3.22 Mutations in Gene Evolution

Mutations in human gene pathology and evolution represent two sides of the same coin in that those same mutational mechanisms that have frequently been

implicated in human pathology have also been involved in potentiating evolutionary change [56]. Regardless of whether they are advantageous, disadvantageous, or neutral, these mutational changes and their putative underlying causal mechanisms are very similar. It is now clear that the gene has often been a dynamic entity over evolutionary time, not a static one. Indeed, during vertebrate evolution, many genes have undergone gross rearrangement as a result of the action of a variety of mutational processes including insertion, inversion, duplication, repeat expansion, translocation, or deletion. What links pathology and evolution is the underlying genomic architecture with its hitherto largely unexplored vocabulary of structural elements, and different types and patterns of repetitive DNA sequences [201]. It can thus be seen that the mutational spectra of germline mutations responsible for inherited disease, somatic mutations underlying tumorigenesis, polymorphisms (either neutral or functionally significant), and differences between orthologous gene sequences, exhibit remarkable similarities implying that they are very likely to have causal mechanisms in common.

6.4 CONSEQUENCES OF MUTATIONS

6.4.1 Variants Affecting the Amino Acid Sequence of the Predicted Protein but Not Gene Expression

Many missense variants (i.e., nucleotide substitutions that result in an amino acid substitution) cause hereditary disease in humans. Missense variants are of importance, in understanding the structure or function of a protein because they usually occur in amino acid residues of structural or functional significance [11]. Occasionally, however, not only is the mutated residue not conserved in mouse, but also the substituting residue in humans is identical to its wild-type counterpart in the orthologous (e.g., murine) gene [374]. It is thought that the most likely explanation for the majority of these cases of fixation of disease variants in mice is *compensatory variant*. The chimpanzee genome has been found to harbor a number of examples of potentially compensated mutations (PCMs), defined as human disease-causing or disease-associated missense variants for which the substituting amino acid is identical to the wild-type amino acid residue at the orthologous position in chimpanzee [375]. The absence of strongly

deleterious consequences of a specific PCM in chimpanzee would be explicable either by virtue of the very different (simian) environment or by dint of hitherto unidentified variants (“compensatory variants”) in the chimpanzee genome that have served to epistatically buffer the PCM [375]. It should be noted, however, that the PCM may only have become seriously disadvantageous in the human lineage, either as a consequence of other lineage-specific genetic changes or due to changes in the human environment and/or lifestyle [376,377]. In this case, it would not have been necessary for the chimpanzee PCM to be compensated for at any time, which is why the qualifier “potentially” is important when PCMs are defined on the basis of current genetic and clinical data.

It is sometimes difficult to establish a causative link between a missense variant and a disease phenotype [378]. The absence of the variant in a large sample (usually 200 individuals) from the same ethnic group as the patient serves to exclude the possibility of a common polymorphism. Amino acid substitutions in evolutionarily conserved residues can also be good candidates for true pathogenicity [11]. If the function of the protein is known, assessment of the effect of the missense variant can be performed by *in vitro* mutagenesis and functional assay. Finally, the introduction of the variant into an entire organism (e.g., transgenic mice) and the study of its systemic effects provide one of the best means to assess its contribution to a particular clinical phenotype. Amino acid substitutions can be shown to reduce or abolish the physiological function of a protein; for example, missense variants have been identified in factor VIII that abolish thrombin cleavage, which is necessary for its activation [379], interfere with binding to other proteins such as vWF [380], or create or abolish *N*-glycosylation sites [381]. In other proteins, variants have been identified, for example in DNA binding domains, catalytic domains, transmembrane domains, ATP-binding regions, receptor-ligand contact sites, phosphorylation, or other chemical modification sites. Missense variants may also affect protein folding causing a dramatic change in secondary and tertiary structures such that the protein can no longer fulfill its physiological function.

A classic example of a missense variant in the active site of an enzyme is provided by α_1 -antitrypsin, Pittsburgh, found in an individual with a fatal bleeding disorder [382]. The underlying variant in the α_1 -antitrypsin

(*SERPINA1*) gene is Met358Arg within the active site of the molecule. The substitution by Arg alters the substrate specificity of α_1 -antitrypsin by converting its “bait loop” (which is specific for elastase) to one that was specific for thrombin. In effect, the molecule lost its anti-elastase activity and became a serine protease inhibitor capable of inhibiting thrombin and factor Xa.

Variants involving gains of glycosylation have generally been considered rare, and the pathogenic role of the new carbohydrate chains has never been formally established [383]; however, the three children identified with Mendelian susceptibility to mycobacterial disease were homozygous with respect to a missense variant in the *IFNGR2* gene that created a new *N*-glycosylation site in the interferon (IFN) γ R2 chain. The resulting additional carbohydrate moiety was found to be both necessary and sufficient to abolish the cellular response to IFN γ . From 10,047 HGMD variants in 577 genes encoding proteins trafficked through the secretory pathway, 142 candidate missense variants (~1.4%) in 77 genes (~13.3%) for potential gain of *N*-glycosylation were identified. Six mutant proteins were shown to bear new *N*-linked carbohydrate moieties. Thus, it may be that an unexpectedly, high proportion of variants causing human genetic disease do so via the creation of new *N*-glycosylation sites. Indeed, the pathogenic effects of these variants may be a direct consequence of the addition of *N*-linked carbohydrate.

Missense variants can result in disease by (1) elimination or reduction of the physiological activity/role of the protein; (2) gain of function by which the amino acid substitution creates new functional capabilities of the protein in biochemical and developmental processes in which the protein either does not participate or has a different role; (3) change of the target function of another protein as in the case of the variant in the protein C cleavage site at Arg 506 of coagulation factor V, which is associated with thrombophilia [384], or in the case of a variant in the thrombin cleavage site of factor VIII that eliminates normal activation of factor VIII [379], or in the case of severe obesity from childhood and R236G in the human pro-opiomelanocortin (*POMC*) gene that disrupts the dibasic cleavage site between β -melanocyte-stimulating hormone (β -MSH) and β -endorphin [385]; and (4) participation of the mutant polypeptide in protein complexes, which renders the entire complex abnormal or nonfunctional, as in the case of the triple helical structure of certain collagens in which

incorporation of one abnormal collagen chain results in “protein suicide” or an abnormal structure that degrades rapidly [386].

Missense variants have a multitude of different effects on protein structure and function including (1) introduction of larger residues within the hydrophobic protein core leading to adverse interactions between residues [387,388], (2) introduction of buried charged residues [387–389], (3) disruption of protein–protein interactions [390], (4) disruption of hydrogen bonding [388,389], (5) interference with DNA binding [388], (6) breakage of disulfide covalent linkages [388], (7) variant of catalytic residues [389,391], (viii) perturbation of metal binding [388], (9) loss of post-translational modification sites [392], (10) gain of intrinsic disorder [387,393], (11) loss of stability [394,395], and (xii) disruption of quaternary structure [388,396].

Without in-depth analytical studies, missense variants are often difficult to distinguish from polymorphisms with little or no clinical significance, either in the context of candidate gene sequencing studies [397] or in the context of exome sequencing studies [398]. In the “genomic era,” a substantial amount of human genetic variation will become amenable to high-throughput analysis in the form of SNPs, and many of these SNPs will influence directly the structure, function, or expression of genes and the RNAs/proteins they encode. Prior knowledge as to which SNPs are most likely to be clinically relevant would greatly enhance the power of studies that aim to identify disease genes through the genotypic screening of patients in both families and populations. Inclusion of structural/functional information could be especially important in the elucidation of multifactorial disease, where genetic heterogeneity and complex interactions between genes and environment have so far limited the success of genetic epidemiological studies [399]. Recently, several predictive models have been developed that use a number of different biophysical parameters to estimate the likely functional impact of an amino acid substitution on the structure and function of a protein [8,394,400,401]. These models have been used to distinguish reasonably and successfully between pathological substitutions, functional polymorphisms, and neutral polymorphisms. Vitkup et al. [402] have claimed that variants at arginine and glycine residues are together responsible for ~30% of cases of genetic disease, whereas random variants at tryptophan and cysteine have the highest probability of causing disease.

6.4.2 Variants Affecting Gene Expression

Variants that do not result in amino acid substitution invariably affect gene expression, that is, transcription, RNA processing, and maturation, translation, or protein stability. Total or partial gene deletions, insertions, inversions, and other gross rearrangements obviously result in the loss of gene expression. These types of variants are usually less frequent unless the genomic sequence environment of specific genes (e.g., presence of repeats) predisposes to such lesions. Disorders with high frequencies of gross rearrangements include α -thalassemia, DMD, steroid sulfatase deficiency, and hemophilia A. Some partial gene deletions that eliminate one or a few exons in-frame result in milder clinical phenotypes because gene expression is not totally eliminated; the resulting protein may lack an amino acid domain that is not critical for its function [403].

6.4.3 Promoter (Transcription Regulatory) Variants

Microlesions within proximal gene regulatory regions currently comprise only ~1.7% of known variants causing or associated with human inherited disease (see HGMD). Their relative rarity may be in part because not all regulatory elements occur immediately 5′ to the genes that they regulate. Indeed, many such elements are located within the first exon, within introns [404] or within 5′ or 3′ UTRs. In the same vein, upstream ORFs (uORFs), present in 50% of human genes, often impact on the expression of the primary ORFs; indeed, both pathogenic and common variants have been reported within uORFs that can modulate or even abolish the expression of the downstream gene [405,406].

Variants in known promoter motifs usually lead to reduced (or occasionally increased) mRNA levels. Such variants have been studied in the TATA box of the β -globin (*HBB*) gene [407]. Other disease-associated nucleotide substitutions occurring within DNA motifs that bind transcription factors include those located in the CACCC motif of the β -globin (*HBB*) gene influencing transcription factor EKLF binding [408,409], several motifs in the γ -globin (*HBG*) genes [410], the CCAAT motif of the *F9* gene influencing C/EBP binding [411], the SP1 motif of the *LDLR* gene promoter [412], the HNF-1 binding site in the *PROC* gene [413], and the binding site for the transcription factor Oct-1 in the lipoprotein lipase (*LPL*) gene [414]. These few examples are only representatives of a total of over 2200 known

Mammalian Splice Site Consensus Sequences

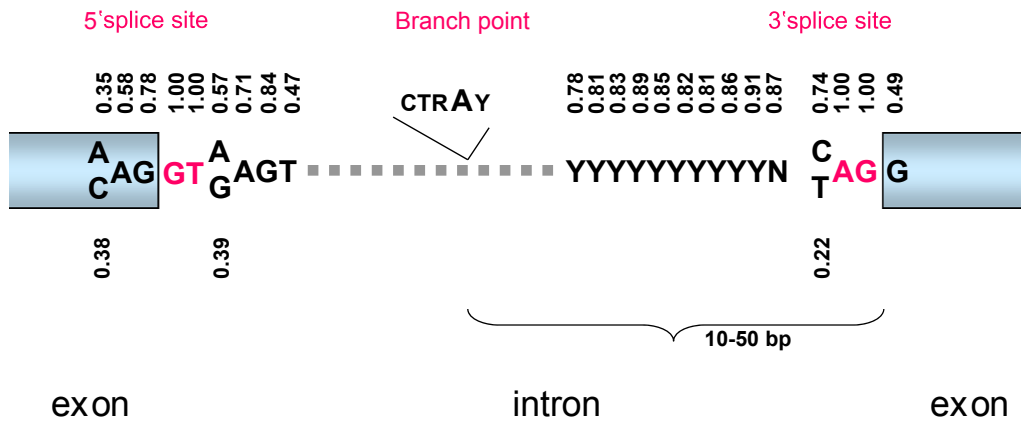


Figure 6.11 Consensus sequences for the donor (5' splice) and acceptor (3' splice) sites and the branchpoint. Numbers above or below the nucleotides correspond to frequencies of a given nucleotide in a large number of mammalian splice-site sequences. Note that the dinucleotides GT and AG (in red) at the beginning and end of the intron are invariant.

promoter variants listed in HGMD and causing human genetic disease. The importance of these mutants lies in the specific DNA sequences thereby implicated in binding to transcription factors. Although most of the known variants reduce the levels of mRNA production, some substitutions actually increase it. Examples include various lesions in the promoters of the γ - and α -globin (*HBG1* and *HBG2*) genes that cause hereditary persistence of fetal hemoglobin (HPFH) due to the inappropriate continuation of δ -globin (*HBD*) gene expression into adult life [415] and a gain-of-function (creates a GATA1 binding site) regulatory SNP, which is located in a nongenic region between the α -globin genes and their upstream regulatory elements [416]. An increase in the distance of promoter elements from the transcriptional start site may also result in gene silencing. Such an example has been found in the promoter elements of the *CSTB* gene in EPM1 [137]. Variants that alter the transcriptional regulation of gene expression have been reviewed in [417].

The concomitant change in local DNA sequence complexity surrounding a substituted nucleotide is directly related to the likelihood of a regulatory variant coming to clinical attention [77]. This finding is consistent with the view that DNA sequence complexity is a critical determinant of gene regulatory function and may reflect the internal axial symmetry that

frequently characterizes transcription factor binding sites. Polymorphisms in the promoter region that are associated with differential levels of gene expression may predispose to common disorders. For example, a G>An SNP, at nucleotide -6 relative to the transcriptional initiation site of the angiotensin (*AGT*) gene, influences the basal level of transcription and may predispose to essential hypertension [418]. In excess of 400 disease-associated promoter polymorphisms are listed in HGMD plus >700 functional promoter polymorphisms that significantly increase or decrease promoter activity but which have not yet been associated with a clinical phenotype.

6.4.4 mRNA Splicing Mutants

Single base-pair substitutions in splice junctions constitute at least 10% of all variants causing human inherited disease. There are, however, a wide variety of variants within both introns and exons that can affect normal RNA splicing (see Ref. [419] for review). The different mechanisms by which disruption of pre-mRNA splicing play a role in human disease have been reviewed in Ref. [420]. The most commonly found variants occur in the dinucleotides GT and AG found at the beginning and end of the donor (5') and acceptor (3') consensus splice sequences (see Fig. 6.11 for the consensus splice elements and Fig. 6.12 for the different kinds of

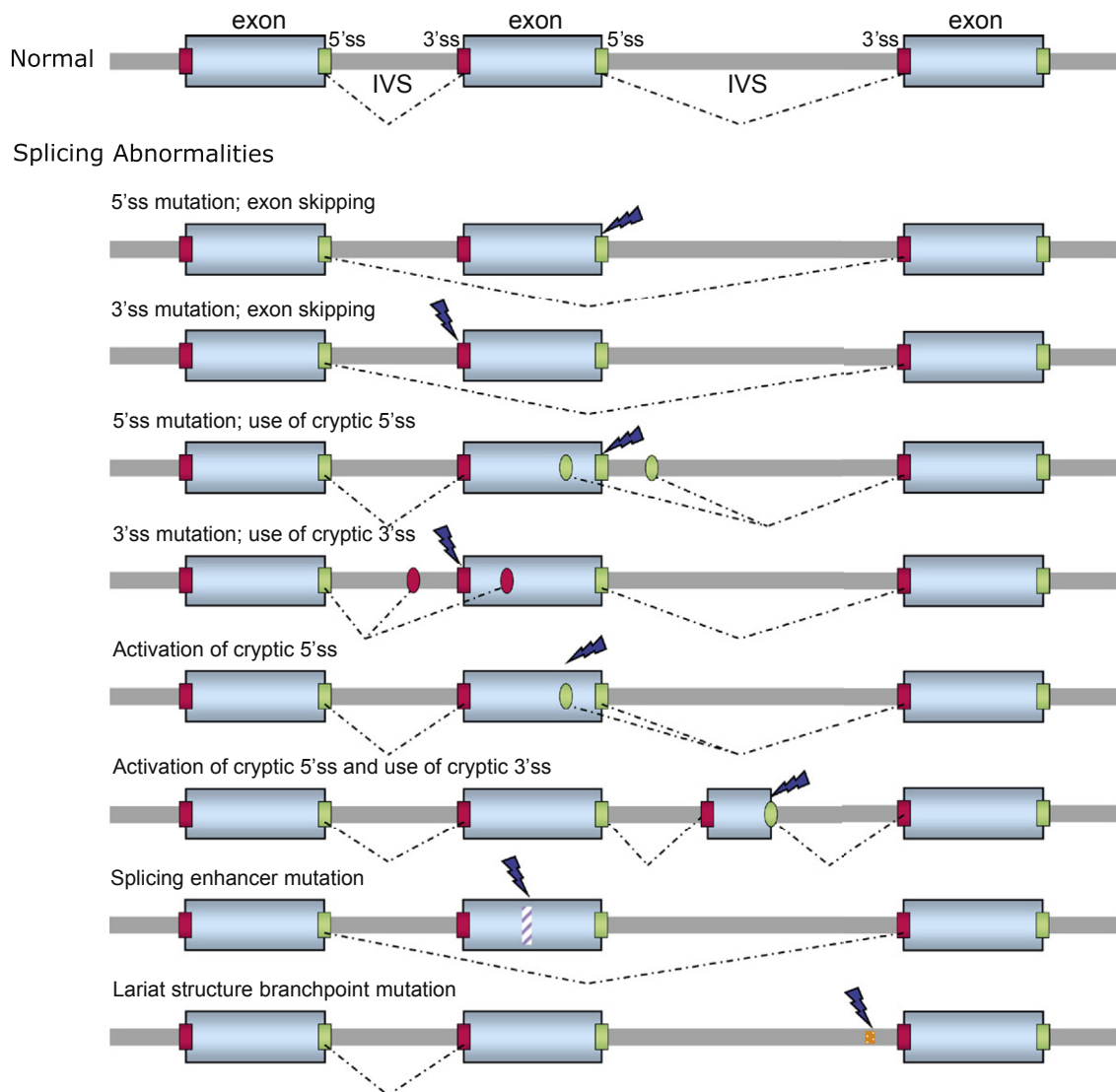


Figure 6.12 Examples of splicing abnormalities in introns of human genes. Exons are shown as *blue boxes*; introns as *lines* between exons. Green squares denote the normal 5' (donor) splice sites; red squares represent the normal 3' (acceptor) splice sites. Green and red circles denote cryptic 5' and 3' splice sites, respectively. The *broken blue wedge* represents the site of mutation.

RNA splicing abnormalities). Almost all of these variants cause either exon skipping or cryptic splice-site utilization resulting in the severe reduction or absence of normally spliced mRNA. In addition, variants in nucleotides +3, +4, +5, +6, -1, and -2 of the consensus donor splice site have also been observed (Fig. 6.13), with variable severity of the RNA splicing defect. Similarly, likely pathogenic variants in positions -3 and

the polypyrimidine tract of the consensus acceptor splice site have been noted (Fig. 6.13). In the majority of these cases, some normal splicing occurs and the defect is not severe. Utilization of cryptic splice sites leads to the production of abnormal mature mRNA with premature stop codons or to the inclusion of additional amino acids after translation (see Ref. [83] for examples and references cited therein).

HGMD Mutations in splice sites (14Jul17)

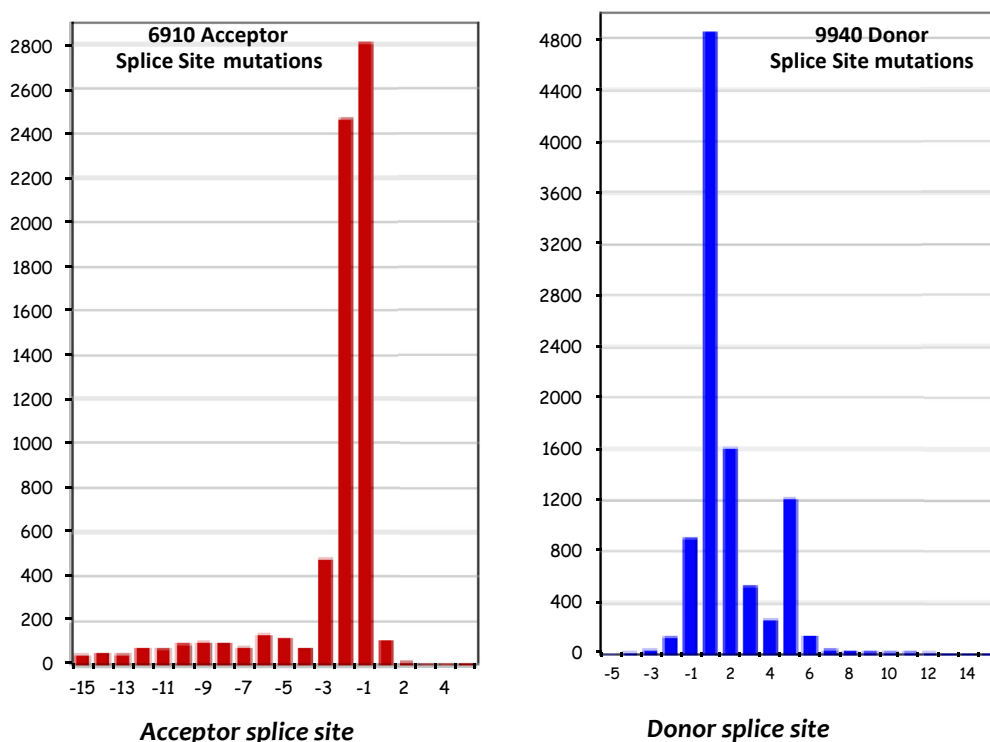


Figure 6.13 Mutations in the consensus sequences of splice junctions recorded in the HGMD.

Using a neural network for splice-site recognition, Krawczak et al. [421] performed a meta-analysis of 478 disease-associated splicing variants, in 38 different genes, for which detailed laboratory-based mRNA phenotype assessment had been performed. Inspection of the ± 50 -bp DNA sequence context of the variants revealed that exon skipping was the preferred phenotype when the immediate vicinity of the affected exon-intron junctions was devoid of alternative splice sites. By contrast, in the presence of at least one such motif, cryptic splice-site utilization became more prevalent. This association was, however, confined to donor splice sites. Outside the obligate dinucleotide, the spatial distribution of pathological variants was found to differ significantly from that of SNPs. Although disease-associated lesions clustered at positions -1 and $+3$ to $+6$ for donor sites and -3 for acceptor sites, SNPs were found to be almost evenly distributed over all sequence positions considered. When all putative missense variants

in the vicinity of splice sites were extracted from the HGMD for the 38 studied genes, a significantly higher proportion of changes at donor sites (37/152; 24.3%) than at acceptor splice sites (1/142; 0.7%) was found to reduce the neural network signal emitted by the respective splice site. It was estimated that some 1.6% of disease-causing missense substitutions in human genes are likely to affect the mRNA splicing phenotype [216].

Other kinds of pathogenic variants in introns include those that cause the activation of cryptic splice sites (by altering a sequence so as to make it more similar to an authentic consensus splice site) or by creation of new splice sites [422]. In both instances, new intron splice patterns occur with consequent introduction of stop codons or abnormal peptides after translation. These variants do not completely abolish normal splicing and are therefore not associated with the absence of normal mature mRNA. A variant in a lariat structure branch-point [423] has been found in the *LICAM* gene in a

patient with X-linked hydrocephalus [424]. By contrast, another variant in intron 5 of the type 2 neurofibromatosis (*NF2*) gene created a consensus branchpoint sequence and led to the activation of a cryptic exon [425].

Some 98.7% of all splice sites in human genes conform to consensus sequences that include the invariant dinucleotides GT and AG at the 5' and 3' ends of the introns, respectively [426]. Noncanonical sequences (e.g., GA-AG, GC-AG, and AT-AC) do however, occur at human splice junctions, albeit much less frequently (<0.02%, 0.69%, and 0.05%, respectively). Some of these noncanonical splice sites are nevertheless known to be used with high efficiency and may be conserved over quite long stretches of evolutionary time. Such sites have occasionally come to clinical attention when they have harbored variants causing human inherited disease [427]. Moreover, the utilization of a cryptic noncanonical donor splice site within exon 1 of the *HRPT2* gene in a case of familial isolated primary hyperparathyroidism as a consequence of a causative lesion in intron 1 of the gene has been reported [428]. RNA isolated from EBV-transformed lymphoblastoid cell lines derived from the patients was used to demonstrate the consequences at the level of the mRNA phenotype (the loss of 30 bases from the mRNA transcript).

Single base-pair substitutions within “splicing enhancer” sequences may also perturb splicing by promoting exon skipping; examples include a variant in intron 3 of the growth hormone (*GH1*) gene causing short stature [429] and a variant in exon 5 of the adenosine deaminase (*ADA*) gene causing ADA deficiency [430]. In patients with frontotemporal dementia with parkinsonism, three heterozygous variants in a cluster of four nucleotides +13 to +16 of exon 10 of the microtubule-associated protein tau (*MAPT*) gene destabilized a potential stem-loop structure that probably is involved in regulating the alternative splicing of exon 10. This caused more frequent use of the 5' splice site and an increased proportion of tau transcripts that include exon 10. The increase in exon 10⁺ mRNA increased the proportion of tau protein containing four microtubule-binding repeats, which is consistent with the neuropathology described in families with this type of frontotemporal dementia [431]. One variant found in the *ATM* gene causing ataxia-telangiectasia was a deletion of four nucleotides (GTAA) in intron 20 within an intron-splicing processing

element (ISPE) that is complementary to U1 snRNA. This element mediates accurate intron processing and interacts specifically with U1 snRNP particles [432]. Finally, the intronic prothrombin (*F2*) gene 19911A>G polymorphism influences splicing efficiency by altering a known functional pentamer CAGGG motif [433]. Some nonsense variants cause skipping of one or more exons, presumably during pre-mRNA splicing in the nucleus. This phenomenon has been termed “nonsense-mediated altered splicing” but its underlying mechanism is unclear. The first such variant was described in the *FBN1* gene in Marfan syndrome [434]. It is now recognized that any nucleotide substitution within exons (nonsense, missense, or translationally silent synonymous point variant) that disrupts a splicing enhancer or silencer (ESE, enhancer splicing element; composite exonic regulatory element of splicing) or creates an exon splicing silencer (ESS) may affect either the pattern or the efficiency of mRNA splicing [99,435–437] (Fig. 6.14). In exon 12 of the *CFTR* gene, about one quarter of synonymous variations result in exon skipping and, hence lead to the synthesis of an inactive CFTR protein [438]. For reviews on the effects of exonic variants on splicing, and additional examples of such pathogenic variants, see Ref. [439]. It has been estimated that pathogenic effects of ~20% of variants in the *MSH2* gene result from missense variants that perturb splicing through the disruption of ESE sites. Similarly, the pathogenic effects of ~16% of missense variants in the *MLH1* gene are thought to be ESE related [440]. Further, recent studies of exon splicing enhancer variants have suggested that as many as 25% of known missense and nonsense variants causing human inherited disease may alter functional splicing signals within exons [441,442]. If this estimate is accurate, it suggests a much more widespread role for aberrant mRNA processing in causing human inherited disease than has hitherto been appreciated.

Splice-mediated insertional inactivation involving an *Alu* repeat was first reported by Mitchell et al. [443]. Analysis of the ornithine δ -aminotransferase (*OAT*) mRNA of a patient with gyrate atrophy revealed a 142 nucleotide insertion at the junction of exons 3 and 4. An *Alu* sequence is normally present in intron 3 of the *OAT* gene, 150bp downstream of exon 3. The *Alu* sequence found in the cDNA was identical to this one, except that the patient was homozygous for a C→G transversion in the right arm of the *Alu* repeat which served to create a new

Exon Skipping due to mutations in enhancer splicing elements

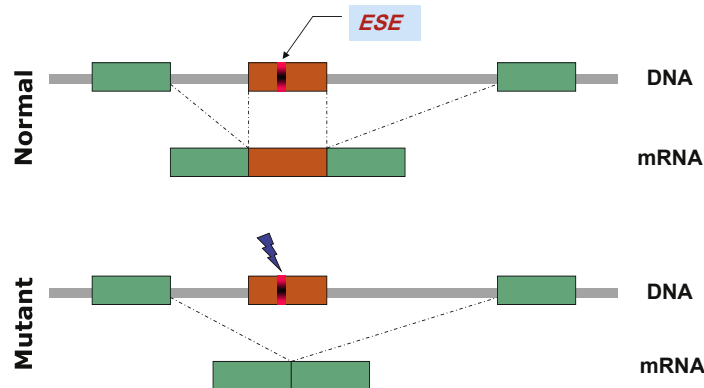


Figure 6.14 Exon skipping due to nonsense, missense, and silent mutations in enhancer splicing elements (ESE). This element is shown as a darkened segment of the middle exon.

5' splice site. This activated an upstream cryptic 3' splice site [the poly(T) complement of the *Alu* poly(A) tail followed by an AG dinucleotide] and a new “exon,” containing the majority of the right arm of the *Alu* sequence, was recognized by the splicing apparatus and incorporated into the mRNA. The splice-mediated insertion of an *Alu* sequence in reverse orientation has also been reported in the *COL4A3* gene causing Alport syndrome [444]. Deep intronic variants, located at some considerable distance from splice sites and known splicing-related sequence elements, generally appear to comprise <1% of known splicing variants [445–447]. Such lesions often create novel splice sites thereby activating cryptic exons (“pseudoexons”). It should be appreciated that the <1% figure is very likely to be an underestimate owing to the inherent difficulty in detecting splicing variants located outside of (and distant from) exon–intron splice junctions. Thus, for example, when the *NF1* gene was methodically screened for variants that altered splicing, 5% of the identified lesions that altered splicing were deep intronic variants [448]. Among disease-causing lesions, inclusion of a pseudoexon as a consequence of cryptic splice-site activation appears to be the most common consequence of deep intronic variant [449]. If we also consider the deep intronic polymorphic variants that have the potential to confer susceptibility to disease [450–452], it is very likely that splicing-relevant intronic variation will have been seriously underascertained thus far. A recent review contains comprehensive tables listing the majority of the known deep intronic pathogenic variants [453].

6.4.5 RNA Cleavage-Polyadenylation Mutants

A number of examples of RNA cleavage-polyadenylation variants have now been described [454]. Those reported occur in the sequence AAUAAA, which is 10–30 nucleotides upstream of the polyadenylation site and is important for the endonucleolytic cleavage and polyadenylation of the mRNA. A Variant in this sequence of the β -globin (*HBB*) gene results in mild thalassemia [455]. In these cases, normal polyadenylation and cleavage occurs at a level ~10% of normal. Alternative AAUAAA sites downstream of the mutated ones are used, resulting in larger mRNAs that are highly unstable. Other variants near the poly(A) cleavage sequence may result in mRNA destabilization; one such variant has been described 12 bp upstream of the AAUAAA sequence of the *HBB* gene in a patient with β -thalassemia [456].

The G>A mutation at the 3'-terminal nucleotide of the 3' UTR of the *F2* (prothrombin) gene mRNA gives rise to an elevated prothrombin plasma level and represents a common genetic risk factor for the occurrence of thromboembolic events. This variant creates an inefficient 3' end cleavage signal and represents a gain-of-function variant, causing increased cleavage site recognition, increased 3' end processing and increased mRNA accumulation and protein synthesis [457,458].

6.4.6 Variants in MicroRNA-Binding Sites

MicroRNAs (miRNAs) posttranscriptionally downregulate gene expression by binding to complementary sequences on the 3' UTRs of their cognate mRNAs,

thereby inducing either mRNA degradation or translational repression. Over 700 human miRNAs have so far been identified but many more probably still remain to be discovered. These miRNAs are each likely to down-regulate a large number of different target mRNAs.

The first reported pathological variant in an miRNA-binding site was a G→A transition in a binding site for miR-189 within the 3' UTR of the *SLITRK1* gene of two apparently unrelated Tourette syndrome patients [459]. Experimental confirmation of the functional effect of this variant came from the demonstration that, in the presence of miRNA-189, in vitro constructs bearing the 3' UTR variant served to increase repression of a reporter gene by comparison with the wild type. A further example of a functional miRNA-target site variation involves an SNP in the 3' UTR of the human *AGTR1* gene; although the variant allele is not downregulated by miR155, it has been associated with hypertension in numerous studies [460]. An increasing number of genetic variants located in microRNA target sites are being reported, either causing or associated with an increased risk of inherited disease [461–465].

6.4.7 Variants in Non-Protein-Coding Genes

In contrast to the plethora of variants identified in protein-coding genes, the identification of variants in non-protein-coding genes is still very much in its infancy [466]. A number of disease-causing or disease-associated variants have already been reported in various small nucleolar RNA genes [467,468] and miRNA genes [469–472]. In addition, pathogenic variants have also been documented in the longer noncoding RNA genes *XIST*, *TERC*, *H19*, and *RMRP* (see HGMD for details). Numerous pathogenic variants numerous have been described in the untranslated *RMRP* long noncoding RNA gene that cosegregate with the cartilage hair hypoplasia phenotype [473].

A putative pathological variant has been described in a “gene” encoding a paternally expressed antisense transcript of the *GNAS* complex locus (*GNASAS*) [474], whereas a functional polymorphism has been reported within an enhancer at the 30 end of the *CDKN2BAS* “gene,” which encodes an antisense RNA transcript [475]. A *CRYGEPI* pseudogene-reactivating variant associated with hereditary cataract formation [476] probably also falls into this category. The above examples are likely to comprise only the tip of a fairly large iceberg that still remains essentially unexplored. Thus,

for example, both SNP and CNV are likely to impact significantly on miRNA gene expression with myriad potential pathological consequences [461].

Pathogenic variants in microRNAs have also been observed. A pathogenic single nucleotide variant in the seed sequence of the human miR-96 gene causes nonsyndromic autosomal dominant progressive hearing loss (DFNA50) [470]. A mouse mutant from ENU mutagenesis (Dmdo mouse) has been identified with a pathogenic variant in the seed sequence of the mouse miR-96 gene. Heterozygous mice show progressive hearing loss, while homozygous have no cochlear responses [477]. Another example is the pathogenic variant in the seed region of miR-184 which is responsible for familial severe keratoconus combined with early-onset anterior polar cataract. The mutant form fails to compete with miR-205 for overlapping target sites on the 3' UTRs of *INPPL1* and *ITGB4* [478]. Furthermore, germline hemizygous deletions of MIR17HG, encoding the miR-17~92 polycistronic miRNA cluster, cause microcephaly, short stature, and digital abnormalities [479].

Pathogenic variants that cause Mendelian phenotypes have also been observed in small nuclear RNAs (snRNAs). The gene encoding the U4atac snRNA, a component of the minor U12-dependent spliceosome, is mutated in individuals with microcephalic osteodysplastic primordial dwarfism type I (MOPD I), a severe developmental disorder characterized by extreme intra-uterine growth retardation and multiple organ abnormalities [480,481].

A recent review [482] provides the majority of the known cases of pathogenic variants in the noncoding genes.

6.4.8 Variants in Noncoding Regions of Functional Significance

By adopting a gene-centric view, we have until now largely ignored the extensive non-protein-coding portion of the human genome in our quest for variants of pathological significance. As a consequence, we have not only seriously underestimated the extent of the functional component of the genome but may also have overlooked many variants within this genomic “dark matter” [483]. In both the human and the mouse genomes, many noncoding regions exhibit a similar level of evolutionary conservation to that evident in protein-coding regions [484,485]. As yet, however, little is known of the effect that variants in these regions

might have on either the phenotype or on overall fitness. Studies of the most evolutionarily conserved noncoding regions have yielded results that are consistent with the view that most variants in noncoding regions are only slightly deleterious [485,486].

To obtain a first, necessarily rather crude, estimate of the contribution of variation in human noncoding sequences to phenotypic and/or disease traits, Visel et al. [487] performed a meta-analysis of ~1200 SNPs that have been identified as the most significantly associated variants in published GWAS. They found that, in 40% of cases, neither the SNP in question nor its associated haplotype block overlapped with any known exons. These authors, therefore, concluded that in at least one-third of detected disease associations, variation in noncoding sequence rather than coding sequence could have causally contributed to the trait in question. We suspect that this could be because the common disease-common variant hypothesis [488] may be much more likely to apply to noncoding sequence than to coding sequence, owing to the selection constraints impacting on sufficiently frequent functional variation in the latter. In similar vein, others have also estimated that 39%–43% of trait/disease-associated SNPs in GWAS are located within intergenic regions [489,490]. This notwithstanding, it should be appreciated that any given variant apparently detected within a noncoding region may actually reside within a hitherto undiscovered exon [491]. We should, however, also be aware that rare variants, in *cis* to those found to be associated with a given disease or trait in GWAS studies, may simply by chance give rise to “synthetic associations” that are then attributed to much more common variants [492].

In the context of identifying genetic variants responsible for human inherited disease, Cooper et al. [5] have argued that it will become increasingly important to consider functional elements in the genome (the “functionome”) rather than simply genes per se. We use the term “functionome” here to describe the totality of the biologically functional nucleotide sequences in the human genome, irrespective of whether they are associated with genes. Because conserved noncoding sequences in the human genome appear to be ~10-fold more abundant than known genes [493], it is likely that (1) currently known variants within coding regions are unlikely to be fully representative of the universe of pathological variants and (2) a whole new grouping of disease-causing variants may await identification and

characterization. Once again, a paradigm shift in our thinking may well be required if we are to maximize the potential of the emerging high-throughput technology to detect new (hitherto latent) types of human genomic pathogenic variants.

The above notwithstanding, it is rather unlikely that the functional non–protein-coding portion of the human genome will prove to be quite as mutation-dense as the protein-coding portion. For most inherited disorders, the mutation detection rate is already fairly high (49%), although this success rate is often achieved by combining different mutation detection methodologies, for example, to screen for exon deletions and CNVs as well as more subtle lesions [494]. At least some of the “missing lesions” may nevertheless be found by screening extragenic functional elements.

A recent review on the noncoding genetic variants in human disease is also recommended to the reader [495]. Another recent review further discusses the noncoding pathogenic variation and particularly structural variations such as deletions, inversions or duplications that have the potential to disturb normal chromatin folding [496]. These abnormalities may lead to the repositioning or disruption of the chromatin topological associating domains (TADs [497]) and the relocation of enhancer elements with altered expression of the corresponding genes. Several recent studies highlight this as important disease mechanisms in developmental disorders. Therefore, the regulatory landscape of the genome has to be taken into consideration when investigating the molecular pathophysiology of human disease. Lupianez et al. [498] presented evidence that human limb malformations are caused by deletions, inversions or duplications that alter the structure of TADs. This in turn resulted in misplacement of limb enhancers of the *EPHA4* gene that drive ectopic gene expression. In addition, the creation of new TADs (neo-TADs) due to genomic duplications, results in altered gene expression and limb malformation phenotype [499].

6.4.9 Cap Site Variants

Transcription of the mRNA is initiated at the so-called cap site, which is protected from exonucleolytic degradation by the addition of α -methylguanine. An A-to-C transversion at the cap site of the β -globin (*HBB*) gene was found in a patient with β -thalassemia [500]. It is not, however, clear if this variant causes reduced transcription or abnormal initiation of transcription because C

is found in 6% of transcriptional initiation sites [501] (the most common nucleotide [76%] at position +1 is A). A functional (C/A) polymorphism of the transcriptional initiation site has been noted in the *APOH* gene; the rarer A allele displayed a carrier frequency of 0.12 and was associated with markedly reduced plasma β 2 glycoprotein I [502].

6.4.10 Variants in 5' UTRs

Sequence motifs in the 5' UTRs of genes are thought to play a role in controlling the translation of the encoding mRNA. The phenotypic effects of lesions in 5' UTRs and their clinical consequences have been reviewed [503]. Pathogenic variants in the iron response element (IRE) in the 5' UTR of the ferritin (*FTH1*) gene interfere with the post-transcriptional regulation of ferritin synthesis by decreasing the affinity of IRE for IRE-binding protein [504]. In contrast, decreases in the steady-state level of β -globin (*HBB*) mRNA have been noted in association with a single base deletion at position +10, a G-to-A substitution at position +22, a C-to-G transversion at position +33, and a 4-bp deletion (AAAC) at position +(40–43) in the *HBB* 5' UTR [505–507].

6.4.11 Variants in 3' Regulatory Regions

Sequences in the 3' regulatory regions (3' RRs) of genes are known to be involved in controlling mRNA cleavage/polyadenylation and determining mRNA stability, nuclear export, intracellular localization, and translational efficiency. Although such regions are rich in regulatory elements, relatively few pathological variants have been reported [454,508]. Although only ~0.2% of likely pathogenic variants currently logged in HGMD are located within 3' RRs, this is likely to represent a rather conservative estimate of their actual prevalence. A typical example is the G→A transition 69 nucleotides downstream of the polyadenylation site of the δ -globin (*HBD*) gene causing δ -thalassemia [509]; the variant occurs within a GATA motif and serves to increase the binding affinity of the sequence for erythroid-specific DNA binding protein.

In an attempt to study 3' RR variants systematically, Chen et al. [454,510] performed a systematic analysis of disease-associated variants in the 3' RRs of human protein-coding genes. A total of 121 3' RR variants in 94 human genes were collated including 17 variants in the upstream core polyadenylation signal sequence (UCPAS), 79 in the upstream sequence (USS) between

the translational termination codon and the UCPAS, 6 in the left arm of the “spacer” sequence between the UCPAS and the pre-mRNA cleavage site, 3 in the right arm of the “spacer” sequence or downstream core polyadenylation signal sequence, and 7 in the DSS of the 3'-flanking region. All the UCPAS variants and the rather unusual cases of *DMPK*, *SCA8*, *FCMD*, and *GLA* variants were found to exert a significant effect on the mRNA phenotype and the majority cause monogenic disease. By contrast, most of the remaining variants were polymorphisms, were found to exert a comparatively minor influence on mRNA expression, but may predispose to, protect from, or modify complex clinical phenotypes. The systematic study of these lesions permitted the identification of consistent patterns of secondary structural change that promise to allow the discrimination of nonfunctional USS variants from their functional counterparts.

6.4.12 Translational Initiation Codon Variants

Pathogenic variants in the ATG translational initiation codon have been reported in quite a wide variety of disorders (e.g., Ref. [511]). Instances of substitutions in all three nucleotides have been observed in β -thalassemia, Norrie disease, albinism, phenylketonuria, McArdle disease, and Albright osteodystrophy among others. Indeed, a total of 405 likely pathogenic variants within ATG translational initiation codons are recorded in HGMD representing ~0.7% of the total number of reported coding sequence variants causing human inherited disease. Almost invariably, the variant leads to severe reduction of steady-state mRNA levels similar to that associated with nonsense variants. The mutant mRNA is presumably not translated. The first AUG codon occurs in the context of the so-called Kozak consensus sequence GCCA/GCCAUGG, which is thought to be recognized by the 40S ribosomal subunit [512]. Variants at the initiator methionine ATG may completely abolish translation; however, there are alternative possibilities, namely utilization of the mutant ATG with much reduced efficiency or translational initiation at the next available ATG codon. A C/T polymorphism immediately 5' to the ATG codon within the Kozak sequence of the *CD40* gene is thought to influence translation efficiency [513].

Some diseases are caused by variants that perturb the initiation step of translation by changing the context around the start AUG codon or introducing upstream

AUG codons (see Ref. [514] for review). The scanning mechanism provides a framework for understanding the effects of these changes in mRNAs. The scanning mechanism refers to the entry of the small ribosomal subunit at the (usually capped) 5' end of the mRNA and linear migration until an AUG codon is encountered. Mutational mechanisms such as (1) reinitiation at an internal start codon (e.g., thrombopoietin, *TPO*) and (2) leaky scanning (as in the case of the *Rx/rax* gene underlying the mouse eyeless variant) probably account for such cases.

Naturally occurring variants in the GCCA/GCCAUGG motif include (for the numbering of the mutant nucleotide, the A of the AUG codon is +1; see references in Ref. [514]): (1) +4 G-to-A in the androgen receptor (*AR*) gene in a family with partial androgen insufficiency; (2) -1 C-to-T transition in the α -tocopherol transfer protein (*TTPA*) gene in a family with vitamin E deficiency; (3) a 2 nt deletion causes an A-to-C change at position -3 of the α -globin gene (*HBA*) in a patient with α -thalassemia; (4) -3 A-to-T transversion in the mouse *Pax6* gene causes defects in eye development; and (5) -3 G-to-C somatic variant in the *BRCA1* gene in one case of highly aggressive sporadic breast cancer. It is not surprising that most of the naturally occurring variants involve positions -3 and +4, the positions wherein experimentally induced mutations have the strongest effect.

A meta-analysis of 405 unique (HGMD-derived) single base-pair substitutions, located within the ATG translation initiation codons of 255 different genes, reported to cause human genetic disease has been performed [515]. Although these lesions comprised only 0.7% of coding sequence variants in HGMD, they nevertheless were 3.4-fold overrepresented as compared to other missense variants. The distance between a translation initiation codon and the next downstream in-frame ATG codon was significantly greater for genes harboring ATG codon variants than for the remainder of genes in HGMD (control genes). This suggests that the absence of an alternative ATG codon in the vicinity of an ATG translation initiation codon increases the likelihood that a given ATG mutation will come to clinical attention. An additional 42 single base-pair substitutions in 37 different genes were identified in the vicinity of ATG translation initiation codons (positions -6 to +4, comprising the Kozak consensus sequence). These substitutions were, however, not evenly distributed, being

significantly more abundant at position +4. Finally, contrary to the authors' initial expectation, the match between the original translation initiation codon and the Kozak consensus sequence was significantly better (rather than worse) for genes harboring ATG codon mutations than for the HGMD control genes [515].

6.4.13 Termination Codon ("Nonstop") Variants

"Nonstop" variants are single base-pair substitutions that occur within translational termination (stop) codons, which can lead to the continued and inappropriate translation of the mRNA into the 3'-UTR. The classic example of a termination codon mutant is the case of the α_2 -globin Constant Spring, with a variant in the normal stop codon; this substitution leads to incorporation of an additional 31 amino acid residues in the α_2 -globin polypeptide chain [516]. The resulting protein is unstable and does not interact properly with the β -globin chains of hemoglobin. Some 119 variants within Term codons (in 87 different genes) have been recorded in HGMD, representing ~0.2% of all missense/nonsense variants.

A meta-analysis of these 119 nonstop variants noted a paucity of alternative in-frame stop codons in the immediate vicinity (0–49 nucleotides downstream) of the mutated stop codons as compared with their control counterparts [517]. This implies that at least some nonstop variants with alternative stop codons in close proximity will not have come to clinical attention, possibly because they will have given rise to stable mRNAs (not subject to nonstop mRNA decay) that are translatable into proteins of near-normal length and biological function. A significant excess of downstream in-frame stop codons was, however, noted in the range 150–199 nucleotides from the mutated stop codon [517]. The authors speculated that recruitment of an alternative stop codon at greater distance from the mutated stop codon might trigger nonstop mRNA decay, thereby decreasing the amount of protein product and yielding a readily discernible clinical phenotype.

6.4.14 Frameshift Variants

A large number of frameshift variants have been described in numerous disease-related genes. All lead to altered translational termination with abnormal polypeptide chains after the frameshifts; severe phenotypes are usually seen [518]. Frameshifts occur with

microdeletions or microinsertions and exon skipping. The mechanisms underlying these variants were discussed earlier in this chapter.

6.4.15 Nonsense Variants

Nonsense variants give rise to premature termination of translation and truncated polypeptides. They account for ~11% of all described gene lesions causing human inherited disease and ~20% of disease-associated single base-pair substitutions affecting gene coding regions [519]. Pathological nonsense variants resulting in TGA (38.5%), TAG (40.4%), and TAA (21.1%) occur in different proportions to naturally occurring stop codons [519]. Of the 23 different nucleotide substitutions giving rise to nonsense variants, the most frequent are CGA→TGA (21%; resulting from methylation-mediated deamination) and CAG→TAG (19%) [519]. The differing nonsense mutation frequencies are largely explicable in terms of variable nucleotide substitution rates such that it is unnecessary to invoke differential translational termination efficiency or differential codon usage. Nonsense variants are usually associated with a reduction in the steady-state level of cytoplasmic mRNA [520]. This mechanism of “NMD” is responsible for the degradation of mRNAs that contain a premature termination codon at a position at least 50 nt upstream of an exon–exon boundary [521] but it is not universal [522]. A recent study of 4584 protein-truncating variants has introduced a better predictor of NMD than the 50 bp rule [523]. One or more parameters could be affected: the transcription rate, the efficiency of mRNA processing or transport to the cytoplasm, or mRNA stability.

In the majority of described instances of nonsense variants, the resulting disorders are recessive in nature as a consequence of the haploinsufficiency resulting from the NMD-induced absence of the truncated proteins (which ensures that such polypeptides do not interfere with the function of the wild-type protein). Nonsense variants that do not elicit NMD can, however, give rise to a dominant negative condition (e.g., variants in the *SOX10* gene causing Waardenburg Shah syndrome [524]). Because for NMD to be activated, the nonsense variant must reside at least 50–55 nt upstream of an exon–exon boundary, it follows that the precise location of the nonsense variant could be an important factor in predicting the pathogenicity of that lesion. By way of example, nonsense variants within the last exon of the human β -globin (*HBB*) gene do not elicit NMD.

As a consequence, the truncated β -globin product has near-normal abundance, fails to associate properly with α -globin, and hence gives rise to a dominantly inherited form of α -thalassemia [31]. Different nonsense variants within the same gene may thus be associated with different clinical phenotypes depending on whether or not NMD is activated. Another example of this is provided by a nonsense variant (Q37X) in the *DAX1* gene of an adrenal hypoplasia congenita patient; this lesion is associated with a milder-than-expected clinical phenotype on account of the expression of a partially functional, amino terminal-truncated DAX1 protein synthesized from an alternative in-frame translational start site at Met83 [525]. In a recent meta-analysis, the proportion of known disease-causing nonsense variants predicted to elicit NMD was found to be significantly higher than among nonobserved (potential) nonsense variants, implying that nonsense variants that elicit NMD are more likely to come to clinical attention [519]. In practical terms, the observation of greatly reduced or absent cytoplasmic mRNA associated with nonsense variants has important implications for mutation screening. Thus, attempts to obtain mRNA for RT-PCR and mutation detection may result in amplification of nucleic acid from only the non-nonsense variant-bearing allele. Nonsense variants in the factor VIII (*F8*) gene (hemophilia A) and fibrillin (*FBN1*) gene (Marfan syndrome) have been associated with the skipping of exons containing these variants [252,434] and this observation has now been extended to other genes; exon skipping is either complete or partial. The mechanism underlying this phenomenon is unknown although a number of intriguing models have been proposed [526].

Some genes are characterized by numerous nonsense variants but relatively few if any missense variants (e.g., *CHM*), whereas other genes exhibit many missense variants but few if any nonsense variants (e.g., *PSEN1*). Genes in the latter category have a tendency to encode proteins characterized by multimer formation [519]. Consistent with the operation of a clinical selection bias, genes exhibiting an excess of nonsense variants are also likely to display an excess of frameshift variants [519]. Recently, an example of the spontaneous read-through of a premature termination codon was reported in a patient who was a compound heterozygote for two nonsense variants in the *LAMA3* gene (R943X/R1159X) [527]. The patient, who presented with junctional epidermolysis bullosa, was expected to die as a consequence

of harboring these nonsense variants but was “rescued” by spontaneous read-through of the R943X-bearing allele. This patient’s full-length R943X-bearing *LAMA3* mRNA escaped nonsense-mediated decay, thereby ensuring near-normal *LAMA3* mRNA and laminin- α 3 protein levels. The genetic context of the *LAMA3* variant R943X was found to be close to a hypothetical consensus sequence for optimal premature termination codon read-through.

6.4.16 Unstable Protein Mutants

Missense (nonsynonymous) variants can cause abnormal protein folding and are, therefore, associated with reduced expression owing to instability of the protein. Reviews of variants that affect protein stability can be found in Refs. [528,529]. For proteins that circulate in body fluids, most variants are associated with “CRM negative” status in which the amount of protein correlates with the amount of activity or “CRM reduced” status in which the amount of activity is still lower than the amount of protein produced. Many such variants have been seen in factor VIII causing mild/moderate hemophilia A [87].

The nature of the biophysical properties of amino acid substitutions in p53 that increase their likelihood of coming to clinical attention has been explored [530]; these include solvent inaccessibility, the number of adverse steric interactions introduced, and a reduction in H-bond number. This study was extended by modeling in silico all amino acid replacements that could potentially have arisen from an inherited single base-pair substitution in five human genes encoding arylsulfatase A (*ARSA*), antithrombin III (*SERPINC1*), protein C (*PROC*), phenylalanine hydroxylase (*PAH*), and transthyretin (*TTR*) [401]. A total of 9795 possible mutant structures were modeled and 20 different biophysical parameters assessed. Comparison with the HGMD-derived spectra of 469 clinically detected missense variants indicated that several types of mutation-associated change affected protein function, including the energy difference between wild-type and mutant structures, solvent accessibility of the mutated residue, and distance from the binding/active site. These parameters are considered to be important in protein folding which adds support to the view that many missense variants come to clinical attention by virtue of their consequences for protein folding and stability [531,532].

6.4.17 Variants in Remote Gene Regulatory Elements

In the β -globin gene cluster, a regulatory region ~10 kb upstream of the ϵ -globin (*HBE*) gene has been identified that is capable of directing a high level of position-independent β -globin gene expression [533]. This region, termed the locus control region (LCR), is thought to organize the entire 60-kb β -globin gene cluster into an active chromatin domain and to enhance the transcription of individual globin genes [534]. A similar LCR is also present in the α -globin gene cluster and other gene clusters [535]. Deletions of the LCR in the β -globin gene cluster result in silencing of the β -globin (*HBB*) and other genes of the cluster, even although the coding regions of these genes are still intact [415]. A particular 25-kb deletion, known as Hispanic $\gamma\delta\beta$ -thalassemia, which deletes sequences 9.5–39-kb upstream of the ϵ -globin (*HBE*) gene including the LCR renders the *HBB* gene 60 kb downstream of the deletion nonfunctional [536]. This extraordinary effect of the deletion of the LCR is thought to be due to an altered (DNase I-resistant) state of chromatin associated with nonfunctional genes. Several other examples of similar deletions in the LCR of the α -globin gene cluster have been reported [537].

Several other examples of remote regulatory elements have come to attention as a consequence of their ablation by gross deletions located at some considerable distance (from 10 kb to several megabases) from the genes whose expression they disrupt [160]. For instance, a 960-kb deletion of noncoding sequence, lying between 1.477 Mb and 517 kb upstream of the *SOX9* gene gives rise to the acampomelic form of campomelic dysplasia [538]. Such pathological deletions, however, are not necessarily always so large. Indeed, a 7.4-kb deletion, located 283 kb upstream of the *FOXL2* gene, has been identified as a cause of blepharophimosis syndrome; it disrupts a long noncoding RNA (*PISRT1*) as well as eight conserved noncoding sequences [539]. For some conditions, such lesions may actually occur quite frequently, as in the case of the *SHOX* gene where ~22% of Leri-Weill syndrome patients and ~1% of individuals with idiopathic short stature harbor a microdeletion spanning the upstream enhancer region that leaves the coding region of the *SHOX* gene intact [540].

During the past few years, a number of other examples of likely pathogenic variants in remote promoter elements have been reported. These include a total of

nine variants within a 1-kb region (termed the long-range or limb-specific enhancer) ~979 kb 5' to the transcriptional initiation site of the sonic hedgehog (*SHH*) gene [541] and a T>C transition 1.44 Mb upstream of the *SOX9* gene associated with cleft palate/Pierre Robin sequence [542]. Far upstream polymorphic variants that influence gene expression and that are relevant to disease are also beginning to be documented. Thus, for example, the C>T functional SNP 14.5 kb upstream of the *IRF6* gene, associated with cleft palate, alters the binding of transcription factor AP-2 α [543]. Similarly, a functional SNP ~6 kb upstream of the α -globin-like *HBM* gene serves to create a binding site for the erythroid-specific transcription factor GATA1 and interferes with the activation of the downstream α -globin genes [416]. A functional SNP ~335 kb upstream of the *MYC* gene increases the risk of colorectal and prostate cancer by increasing the expression of the *MYC* gene by altering the binding strength of transcription factors TCF4 and/or TCF7L2 to a transcriptional enhancer [544–547]. Finally, in the context of pointing out the shortcomings of the gene-centric approach to mutation detection, we should be aware that functional SNP rs4988235, located 13.9 kb upstream of the lactase (*LCT*) gene and associated with adult-type hypolactasia, actually resides deep within intron 13 (c.19171+326C>T) of the *MCM6* gene [548–550]. Given that up to 5% of quantitative trait loci for gene expression lie >20 kb upstream of transcriptional initiation sites [551], many more far upstream polymorphic variants that influence gene expression are likely to be identified in the coming years.

Rather fewer pathological variants are known to be located at a considerable distance downstream of human genes. One example is the C>G transversion 2528 nt 3' to the term codon of the *CDK5R1* gene, which has been postulated to play a role in nonspecific intellectual disability [552]. Perhaps more dramatic is the A>G SNP (rs2943641), 565,981 bp 3' to the Term codon of the *IRS1* gene, which is associated with type 2 diabetes, insulin resistance, and hyperinsulinemia; the G allele was found to be associated with a reduced basal level of IRS1 protein [553].

In the light of the above, it can be seen that the underascertainment of disease-associated variants within regulatory regions is likely to be quite substantial but can potentially be rectified by emerging high-throughput entire genome sequencing protocols.

6.4.18 Cellular Consequences of Trinucleotide Repeat Expansions

Trinucleotide repeat expansion has been discussed earlier. In the case of fragile X, the (CGG) $_n$ repeat is located in the 5' UTR of the *FMR1* gene, and its expansion to full mutation results in hypermethylation of the promoter region, loss of transcription, and hence silencing of the gene [554]. Loss of the encoded protein, fragile X mental retardation protein (FMRP), which is thought to play a role in dendritic mRNA transport and translation, is responsible for the classic fragile X syndrome phenotype. Gene inactivation can also be caused by altering the spacing of promoter elements from the transcriptional start site as in the case of the 12-mer repeat expansion in the *CSTB* gene [137].

When the trinucleotide repeat lies within the gene-coding region as in Huntington disease, its expansion results in an abnormal protein with a gain of function due to the enlargement of the polyglutamine tract. Mutant huntingtin exerts its pathological effects via abnormal protein aggregation, transcriptional dysregulation, mitochondrial dysfunction, excitotoxicity, and abnormal cellular trafficking, leading to neuronal loss particularly in the dorsal substratum [555].

Another example of a gain-of-function mutation is provided by the expansion of the CTG repeat in the 3' UTR of the *DMPK* gene causing type 1 myotonic dystrophy (DM1). This does not abolish transcription but rather causes nuclear retention of RNA transcripts leading to the transcriptional dysregulation of other genes [556]. CTG expansion appears to lead to the sequestration of cellular RNA-binding proteins, which in turn gives rise to the abnormal splicing of multiple transcripts [557]. DM1 thus exemplifies a disease whose mechanistic basis lies at the RNA level.

6.4.19 Variants That Give Rise to Inappropriate Gene Expression

HPFH and hereditary persistence of α -fetoprotein (HPAFP) are two clinical conditions that are prototypes for the inappropriate expression of γ -globin (*HBG1* and *HBG2*) and α -fetoprotein (*AFP*) genes, respectively. Normally, the levels of fetal hemoglobin (HbF; $\alpha 2\gamma 2$) in adult life are very low, as there is a switch from fetal to adult hemoglobin during the perinatal period. Similarly, AFP is produced at high level in fetal liver but declines rapidly after birth. In HPFH and HPAFP, however, the levels of HbF and

AFP, respectively, are inappropriately high in adult life. This is often due to single nucleotide substitutions in the promoter regions of the *HBG2*, *HBG1*, or *AFP* genes. A considerable number of variants that occur in the region -114 to -202 of the γ -globin genes have been characterized and presumably cause persistent expression of their corresponding genes [415]. A similar situation has been observed with a -119 variant in the *AFP* gene [558]. These variants occur within DNA binding motifs for transcriptional regulators. A very interesting mutational mechanism has been proposed for facioscapulohumeral muscular dystrophy (FSHD), a common autosomal dominant myopathy associated with a typical pattern of muscle weakness. Most FSHD patients carry a large deletion of a 4q35-located polymorphic D4Z4 macrosatellite repeat array and present with fewer than 11 repeats whereas normal individuals possess between 11 and 150 repeats [559]. An almost identical D4Z4 repeat array is present at 10q26 [560] and the high sequence homology between these two arrays can cause difficulties in molecular diagnosis. Each 3.3-kb D4Z4 repeat contains a *DUX4* (double homeobox 4) gene that, among others, is activated on contraction of the 4q35 D4Z4 array due to the induction of chromatin remodeling of the 4qter region. An increasing number of 4q subtelomeric sequence variants are now recognized, although FSHD only occurs in association with three “permissive” haplotypes, each of which are associated with a polyadenylation signal located immediately distal of the last D4Z4 repeat [561]. This poly(A) signal stabilizes any *DUX4* mRNAs transcribed from this most distal D4Z4 repeat in FSHD muscle cells. Synthesis of both the *DUX4* transcripts and the protein in FSHD muscle cells induces significant cell toxicity. *DUX4* is a transcription factor that targets several genes, which results in a deregulation cascade that inhibits myogenesis, sensitizes cells to oxidative stress, and induces muscle atrophy, thereby recapitulating many of the key molecular features of FSHD [562].

6.4.20 Position Effect in Human Disorders

In several instances, a DNA alteration is found well outside the putative gene that is primarily involved with a disease. Variants acting by “positional effect” are those in which the transcription unit and minimal promoter of the gene remain intact, but there is a nearby alteration that influences gene expression [563]. These positional

effect DNA lesions may involve distal promoter regions, enhancer/silencer elements, or changes in the local chromatin environment. The positional effect could be up to several megabases away from the gene of interest. The examples of the LCR in the β -globin gene cluster and the transcriptional repressor D4Z4 in FSHD are provided elsewhere in this chapter. Most of the position effects are due to chromosomal rearrangements that frequently lead to alteration of the chromatin environment of the gene. Possible mechanisms which may lead to a positional effect include the following: (1) The rearrangement separates the transcription unit from distant *cis*-regulatory elements (enhancer removal results in gene silencing, whereas silencer removal results in inappropriate gene activation); (2) juxtaposition of the gene with an enhancer element from another part of the genome; (3) removal of an insulator or boundary element may also lead to inappropriate gene silencing; (4) enhancer competition of DNA sequences that were juxtaposed to the gene; (5) positional effect variegation in which the chromosomal rearrangement causes the juxtaposition of an euchromatic gene with a region of heterochromatin.

Some examples of positional effect mutations due to translocation breakpoints include genes *PAX6* in aniridia [564], *SOX9* in campomelic dysplasia [565,566], *POU3F4* in X-linked deafness [567], *HOXD* complex in mesomelic dysplasia [568], *FOXL2* in blepharophimosis/ptosis/epicanthus inversus syndrome (BPES) [569,570], and the *SHH* gene in preaxial polydactyly [571]. In these cases, the translocation breakpoints may be in excess of a megabase away from the inappropriately expressed/silenced gene. Indeed, in one example of campomelic dysplasia, the breakpoint maps ~ 1.3 Mb downstream of the *SOX9* gene, making this the longest-range position effect so far found [566]. For a recent review of position effect mutations, see [572].

It is likely that in the majority of cases, the position effect involves a highly conserved *cis*-acting regulatory element. These *conserved noncoding elements* (CNCs; also termed multiple-species conserved sequences; conserved nongenic sequences; the most highly conserved are also called ultraconserved elements) comprise approximately 1%–2% of the human genome and represent potential targets for pathogenic variants [573–577]. An example of such a lesion is provided by the 52-kb deletion of a large noncoding region downstream of the sclerostin (*SOST*) gene in patients with van Buchem

disease, leading to altered expression of the *SOST* gene [578]. The deletion disrupts a bone-specific enhancer element that drives *SOST* gene expression.

Pathogenic variants may also occur in nonconserved elements that could become functional after the introduction of the mutant sequence. This pathogenetic mechanism has been described underlying a variant form of α -thalassemia. Affected individuals from Melanesia have a gain-of-function regulatory single nucleotide polymorphism (rSNP) in a nongenic region between the α -globin genes and their upstream regulatory elements. The rSNP creates a new promoter-like element that interferes with the normal activation of all downstream α -like globin genes [416].

6.4.21 Position Effect by an Antisense RNA

An individual with an inherited α -thalassemia has been described who has a deletion that results in a truncated, widely expressed gene (*LUC7L*) becoming juxtaposed to a structurally normal α -globin (*HBA2*) gene. Although it retained all of its local and remote *cis*-regulatory elements, expression of the *HBA2* gene was nevertheless silenced and its CpG island became completely methylated at an early stage during development. The antisense RNA of the *LUC7L* gene appears to have been responsible for the silencing of the *HBA2* gene [579].

6.4.22 Abnormal Proteins Due to Fusion of Two Different Genes

The translation of fusion genes results in novel proteins with different or abnormal properties from their parent polypeptides. Fusion genes are either the result of homologous unequal crossing-over or the junction sequences at breakpoints of chromosomal translocations. Hemoglobin Lepore, a fusion of δ - and β -globin genes, is the prime example of the first mechanism. Other examples of abnormal fusion genes due to unequal crossover include the case of glucocorticoid-suppressible hyperaldosteronism (GSH), an autosomal dominant form of hypertension, caused by oversecretion of aldosterone [580]; some GSH patients have hybrid genes between *CYP11B1* and *CYP11B2*, two highly homologous cytochrome P450 genes on 8q22. The hybrid gene contains the regulatory elements of *CYP11B1*, expressed in the adrenal gland, and the 3' coding region of *CYP11B2*, which is essential for aldosterone synthesis. Another example is the case of abnormalities of color vision resulting from fusion of the green and red color pigment

(*RCP*, *GCP*) genes [581]. Recombination between the Kallmann gene on Xp22.3 (*KALX*) and its homolog (*KALY*) at Yp11.21 results in a fusion gene that is transcriptionally inactive and is associated with Kallmann syndrome secondary to an X; Y translocation. Finally, Francis et al. [582] identified a large atypical hemolytic uremic syndrome family in whom a deletion occurred through MMEJ rather than by NAHR. The deletion resulted in the formation of a *CFH/CFHR3* hybrid gene. The protein product of this gene, a 24 short consensus repeat protein, was found to be secreted at slightly lower levels than wild-type factor H, but the decay accelerating and cofactor activities of this protein were significantly impaired. A growing number of hematological malignancies are associated with abnormal fusion proteins, the genes of which are found at the breakpoints of chromosomal translocations. One of the first reported examples was the case of fusion of the *BCR* and *ABL* genes in the *t*(9;22) known as Philadelphia (Ph) chromosome in chronic myelogenous leukemia. The *BCR* gene is on chromosome 22 and the *ABL* gene is on chromosome 9; after the translocation junction, a fusion gene is created with the promoter elements of the *ABL* gene and the 3' half of the *BCR* gene [583]. A new abnormal protein is detected in the leukemia cells, the abnormal function of which probably contributes to the malignant phenotype. Another example is the case of Ewing sarcoma (a solid tumor of bone) in which an 11;22 translocation results in a fusion of the *FLI1* gene on 11q24 with the *EWS* gene on 22q12 [584]; for a classic review, see [585]. Fusion genes can be readily identified by PCR and can serve either as diagnostic indicators for relapse in the disorders concerned or as indicators of the need for an alternative therapeutic regimen.

6.4.23 Mutations in Genes Involved in Mismatch Repair Associated With Genomic Instability in the Soma

The study of somatic mutation is extremely important both for the study of cancer [586] and for other diseases such as paroxysmal nocturnal hemoglobinuria [587]. Variants that lead to abnormal or abolished function of genes encoding for proteins involved in DNA mismatch repair are of particular importance because they lead to accumulation of variants throughout the genome. For example, some forms of hereditary nonpolyposis colon cancer (HNPCC), which may account for up to 10% of colon carcinoma, are due to variants in genes such as

MSH2 or *MLH1* that encode mismatch repair proteins [588–590]. In families with variants in these genes, the DNA of tumor tissue shows considerable instability as detected by the generation of new alleles for numerous DNA polymorphic markers [591]. One of the genes affected by the genomic instability is that encoding the type II transforming growth factor- β (TGF- β) receptor (*TGFB2R*), which has a run of 10 adenines in its coding region. This run of As is altered, resulting in a frame-shift and absence of the receptor, which in turn releases the cell from TGF- β inhibitory effects and contributes to malignancy [592]. The discovery and further study of genes of the mutation repair system will enhance our understanding of both germline and somatic variants.

6.4.24 Comparison of Germline and Somatic Mutational Spectra

To date, relatively few studies have attempted to compare the germline and somatic mutational spectra for the same genes. This notwithstanding, the mutational mechanisms underlying single base-pair substitutions [593,594], microdeletions, and microinsertions [594–596], and even gross gene rearrangements [597,598] often appear to exhibit similarities between the germline and the soma. Ivanov et al. [599] performed a comparison of somatic, germline, shared (found in both soma and germline), and somatic recurrent mutational spectra for 17 human tumor suppressor genes, which focused on missense single base-pair substitutions and microdeletions/microinsertions. The somatic and germline mutational spectra for these genes were similar in relation to CG>TA transitions but differed with respect to the frequency of AT>GC, AT>TA, and CG>AT substitutions. Shared missense variants were found to be characterized by higher mutability rates, greater physicochemical differences between the wild-type and mutant residues, and a tendency to occur in evolutionarily conserved residues and within CpG/CpHpG oligonucleotides. Mononucleotide runs of ≥ 4 bp were identified as hotspots for shared microdeletions/microinsertions.

6.4.25 Mosaicism

Germline mosaicism is a relatively frequent mechanism of inherited disease and provides an explanation for the inheritance pattern in cases where multiple affected offspring are born to clinically and phenotypically normal parents [600]. It arises through the occurrence of a mutation de novo in a germline cell or one of its

precursors during the early embryonic development of the parent. Because mitotic divisions predominate in both spermatogenesis and oogenesis, most germline mutations are likely to be mitotic rather than meiotic in origin. *Somatic mosaicism* results from mutations occurring during mitotic cell divisions in the embryo with subsequent clonal expansion of the affected cells [601]. The clinical effect of somatic mosaicism depends critically upon the developmental stage at which the mutation occurs. Thus, a mutation that occurs very early on in embryonic development is likely to affect many somatic tissues. By contrast, mutations occurring rather later may give rise to a phenotype that is confined to a single body region or even to a single organ. Somatic mosaicism arising at a very early embryonic stage can involve both somatic cells and germ cells. Such individuals (*gonosomal mosaics*) are at risk of having affected children. Recent data support the postulate that the frequency of mosaicism is increased in cancer predisposition syndromes characterized by high new mutation rates, suggestive of a direct relationship between the mutation rate in the soma and that in the germline [602].

Somatic mosaicism could also contribute to noncancer phenotypes [603], and may emerge as a substantial determinant of phenotypic variability and disease. A notable mosaic abnormality involving isochromosomes is that of 12p, which causes Pallister-Killian syndrome (OMIM 601803). Isochromosome 12p is only observed in the mosaic state, presumably because such an abnormality is lethal constitutionally. The evaluation of 10,362 patients with a custom-designed, exon-targeted whole-genome oligonucleotide array has detected somatic mosaicism in a total of 57 cases (0.55%) including 20 cases of mosaic CNVs [604]. Illustrative examples of disorders due to somatic variants include the Proteus syndrome due to recurrent somatic E17L mutations in the *AKT1* gene [605], fibroadipose hyperplasia caused by somatic activating mutations in *PIK3CA* [606], and hemimegalencephaly due to somatic variants in the *PIK3CA*, *AKT3*, and *MTOR* genes [607].

6.4.26 Human Mutation Rates

Recent studies have estimated the human mutation rate per nucleotide per generation to be between 7.6×10^{-9} and 2.2×10^{-8} [608,609]. This equates to an average of 50–100 de novo mutations in a newborn genome, which corresponds to ~ 0.86 de novo amino acid altering mutation. The genome-wide CNV mutation rate has also

been estimated using SNP microarrays; at a resolution of ~30 kb, Itsara et al. observed nine *de novo* CNVs from 772 transmissions, corresponding to a mutation rate of 1.2×10^{-2} CNVs per genome per transmission [610].

Sex differences in mutation rates may have a variety of different underlying causes. For *premeiotic mutations*, the single most important factors are likely to be the much higher number of cell divisions during spermatogenesis as compared to oogenesis and the fact that the number of male germ cell divisions experienced is age dependent [611]. However, the likelihood of a given mutation having originated in a particular parent is often dependent on the nature of the mutation in question. In general, point mutations tend to display a paternal bias, arising during spermatogenesis, while gross deletions tend to occur predominantly in females having originated during oogenesis [612,613]. The ratio of the male-to-female nucleotide substitution mutation rates has been estimated to be around 6 [90] but may be as high as 20 [614], rather higher than expectation based on the ratio of the relative numbers of male versus female germline cell divisions, and consistent with most mutations being replication driven. The complete genomes from two parent–offspring trios have been sequenced in 2011 to >22-fold mapped depth; a total of 49 and 35 germline *de novo* mutations were identified in two parent–offspring trios, respectively. In one family, 92% of the *de novo* mutations originated from the paternal germline, whereas the equivalent figure from the other family was 36% of *de novo* mutations [615]. The study of the *de novo* mutations in 78 Icelandic parent–offspring trio after whole genome sequence at high coverage has shown that with an average father's age of 29.7, the average *de novo* mutation rate is 1.20×10^{-8} per nucleotide per generation [616]. The mutation rate of single nucleotide polymorphisms was dominated by the age of the father at conception of the child. There is an increase of about two mutations per year of paternal age. The authors have estimated that paternal mutations double every 16.5 years [616]. The analysis of 11,020 *de novo* mutations from the whole genome in 250 Dutch families showed that the *de novo* mutations in offspring of older fathers are not only more numerous but also occur more frequently in early-replicating, genic regions [617]. The most likely cause of the paternal age effect is the increasing number of cell divisions in the male germline [611]. While oocytes are produced early in a woman's life and have a fixed number of genome

replications, spermatogenic stem cells undergo continuous genome replication throughout a man's life. It has been estimated that the male germline has a history of 160 genome replications in a 20-year-old man, reaching 610 genome replications in a 40-year-old man [611]. Meta-analysis of 6570 mutations in the British DDD study showed that the number of *de novo* mutations increases with the fathers' age by 2.87 per year (95% confidence interval 2.11–3.64) [618]. Interestingly, when the sequence analysis was at much higher coverage of >500 reads per nucleotide, 3.8% of mutations categorized as *denovo* were mosaic in the parental germline. Two of the mutational signatures [619], previously termed Signatures 1 (25% of *de novo* mutations) and 5 (75% of *de novo* mutations), explain the majority of the observed mutational pattern [618]. Signature 1 is characterized by C:G>T:A mutations at CpG dinucleotides, while signature 5 is predominately characterized by T:A>C:G mutations [619]. In a different study of deep sequencing of samples from 50 trios (father, mother, affected child) with an alleged 107 *de novo* mutations, it was found that seven (6.5%) of these presumed germline *de novo* mutations were in fact present as mosaic mutations in the blood of the offspring and were therefore likely to have occurred postzygotically [620]. The DDD study estimated that developmental disorders caused by *de novo* mutations have an average birth prevalence of 1 in 213 to 1 in 448, depending on parental age, and given the current global demographics, this equates to almost 400,000 such children born per year [620]. Another study of 7216 autosomal DNMs with resolved parent of origin from whole-genome sequencing of 816 parent–offspring trios has found that the number of DNMs in offspring increases not only with paternal age, but also with maternal age [621]. The mutation signatures were not the same of the mutations in spermatogenesis than in oogenesis. There was an enrichment of maternal *de novo* mutations with motifs of APOBEC-mediated mutagenesis, which result from aberrant DNA double-strand break repair.

6.4.27 Concepts of Dominance and Recessiveness in Relation to the Underlying Variants

A genetic character is held to be *dominant*, if it is manifest in the heterozygous state and *recessive* if it is not. Thus, for a truly dominant condition, homozygotes should be clinically and phenotypically indistinguishable from

heterozygotes [622]. If this is not so, and the homozygote is more seriously affected, then the respective alleles may be regarded as *semidominant* [623].

In general, most recessive alleles are loss-of-function alleles and include gross gene deletions and rearrangements, frameshift variants, nonsense variants, and so on. By contrast, dominant alleles are often associated with gain of function, either due to dominant negative variants (which interfere with and hence abrogate the function of the wild-type allele) or dominant positive variants (which confer increased, constitutive, novel, or toxic activity on the mutant protein). Examples of dominant negative variants are to be found in the *GHI* [624] and *KIT* [625] genes, while dominant positive variants have been reported in the *PMP22* [626], *GNAS1* [627], *DMPK* [628], and *SERPINA1* [382] genes. It should be noted that loss-of-function variants (e.g., *TERT* [629] and *RUNX2* [630]) can also be associated with dominantly inherited conditions in cases where a 50% reduction in the level of the protein product is sufficient to impede function.

For X-linked diseases, it is probably inappropriate to use the terms “dominant” and “recessive” because males are hemizygous and females often display variable expressivity of their heterozygous variants due to skewed X-inactivation or clonal expansion [631].

6.4.28 Genetic Architecture of Complex Diseases

The study of the genetic architecture of complex disease has revolved around the discussion of two apparently opposing models: the common disease–common variant (CD/CV) hypothesis and the multiple rare variant or common disease–rare variant (CD/RV) hypothesis [488]. Because the CD/CV model conceptually underpinned the HapMap Project, GWAS that have used HapMap data have tended to interrogate the association of common SNPs (MAF >5%) with complex diseases and traits. Initial GWAS data, therefore, strongly supported the involvement of common variants, especially common SNPs in complex phenotypes [490]. However, such studies have succeeded in explaining only a small fraction of the heritability of complex phenotypes [632] and this “missing heritability” has tended to challenge the validity of the CD/CV hypothesis. Perhaps not surprisingly, more recent data are revealing contributions from both common and rare variants to complex phenotypes. Thus, although common SNPs can explain a greater

proportion of the heritability than was initially appreciated [633], support for a role for rare variants has also been accumulating from studies of rare SNPs [634,635] and rare CNVs [277,636]. This suggests that the genetic architecture of complex phenotypes is likely to comprise both common and rare variants.

6.5 GENERAL PRINCIPLES OF GENOTYPE–PHENOTYPE CORRELATIONS

Given knowledge of a specific clinical phenotype, to what extent can the underlying causal genotype be inferred? Conversely, given knowledge of a specific genotype, to what extent is it possible to infer the likely clinical phenotypic consequences (in terms, for example, of the penetrance, age of onset, and severity of the disease)? The study of the genotype–phenotype relationship is essentially an exploration of the actual correspondence between the genotype and the phenotype where any particular genotype usually corresponds to multiple phenotypes whereas many different genotypes can often correspond to a given phenotype. Several general principles have emerged as a result of the intensive study of causative variants in genetic disorders. The following discussion highlights some of these principles. The reader is encouraged to use the Online Mendelian Inheritance in Man (OMIM) at <http://www3.ncbi.nlm.nih.gov/Omim> for further information or for specific genes and clinical phenotypes. It is likely that the phenotypic consequences of a given variant will depend on other genetic variants present in the same gene or in the same genome [637]. The review of Wolf [638] provides an excellent guide to the complex issues inherent in the study of the relationship between the mutant genotype and the clinical phenotype.

Cooper et al. reviewed the different proposed molecular mechanisms of penetrance. This could be a function of the specific mutation(s) involved or of allele dosage, differential allelic expression, CNV or the modulating influence of additional genetic variants in *cis* or in *trans* [639].

6.5.1 Variants in the Same Gene May Be Responsible for More than One Disorder

There are many examples to illustrate the principle that variants in a single gene can cause different and distinct clinical phenotypes (“allelic heterogeneity”). Historically, the first example is that of the β -globin (*HBB*) gene on 11pter. Mutations of this gene cause β -thalassemia, sickle

cell disease, and methemoglobinemia. The *L1CAM* gene on Xq28 has been shown to be mutated in hydrocephalus and stenosis of aqueduct of Sylvius, MASA syndrome (mental retardation, aphasia, shuffling gait, and adducted thumbs), and spastic paraplegia 1. The *COL1A2* gene on 7q21–q22 is involved in four different clinical forms of osteogenesis imperfecta (types II, III, IV, and atypical) as well as Ehlers–Danlos syndrome type VII B. The fibroblast growth factor receptor 2 (*FGFR2*) gene is mutated in three different craniosynostosis syndromes, namely Pfeiffer, Crouzon, and Jackson–Weiss. The *COL2A1* gene is implicated in Stickler syndrome type 1, SED congenita, Kniest dysplasia, achondrogenesis-hypochondrogenesis type 2, precocious osteoarthritis, Wagner syndrome type 2, and SMED Strudwick type. In a survey of 1014 genes causing disorders in OMIM, 165 genes were associated with two disorders, 52 genes with three disorders, 24 genes with four disorders, and 19 genes with five or more disorders [80].

6.5.2 One Disorder May Be Caused by Variants in More than One Gene

There are a plethora of similar clinical phenotypes due to mutations in different genes. This observation, also known as “nonallelic” or “locus” heterogeneity, is well understood thanks to linkage analyses for genetic disorders and the search for mutations in different genes. Thus, tuberous sclerosis, a relatively common autosomal dominant disorder, is caused by lesions in at least two different loci: *TSC1* on 9q34 and *TSC2* on 16p13.3. Approximately 60% of TSC families show linkage to the *TSC2* locus and 40% to the *TSC1* locus. HNPCC has been associated with pathogenic variants in five different genes: *MLH1* on 3p, *MSH2* on 2p16, *PMS1* on 2q31–q33, *PMS2* on 7p22, and *MSH6* on 2p16. Retinitis pigmentosa has so far been associated with a total of 23 different genes and the list is still growing. We expect that disorders of complex or polygenic phenotypes, such as hypertension, atherosclerosis, diabetes, schizophrenia, and manic-depressive illness, will be associated with a considerable number of genes scattered throughout the genome.

6.5.3 One and the Same Variant May Give Rise to Different Clinical Phenotypes (“Polypheny”)

The clinical phenotype does not only depend on the one variant in the responsible gene; it can be modified by the

action of any of the other ~20,000 protein coding genes and other functional genomic elements in the genome. The environment can also play an important role in the full development of the clinical phenotype. The classic sickle cell disease variant in the β -globin (*HBB*) gene (Glu6Val) may be associated with severe or mild sickle cell disease. The amelioration of the severe clinical phenotype in this case can be attributed to the increased expression of γ -globin genes and the presence of high levels of HbF. The genomic environment of the β -globin gene cluster may, therefore, modify the severity of sickle cell disease as may genetic variation originating from other loci, for example, the α -globin genes [112]. Another example of this phenomenon has recently been provided by studies of certain craniosynostoses. Both Pfeiffer and Crouzon syndromes can be associated with the same C342Y or C342R variants in the *FGFR2* gene.

The clinical phenotype associated with the D178N missense variant in the prion protein (*PRNP*) gene is critically dependent on the presence of the Met or Val129 polymorphic allele to which it is coupled. When D178N lies in *cis* to the Met129 allele, fatal familial insomnia (FFI) results, whereas D178N coupled to the Val129 allele is associated with Creutzfeldt–Jakob disease [640]. The Met/Val129 polymorphism also exerts an effect in *trans* through the normal allele because FFI is more severe and of longer duration in patients homozygous for either the Met or the Val allele.

One of the best examples of the contribution of the environment to the clinical phenotype of single gene disorders is that of phenylketonuria due to PAH deficiency. Individual homozygous or compound heterozygous for pathogenic variants in the *PAH* gene develop severe mental handicap if fed a normal diet. However, the cognitive status remains normal if these individuals are fed with a special, “phenylalanine-free” diet.

6.5.4 Variants in More than One Gene May Be Required for a Given Clinical Phenotype (Digenic Inheritance; Triallelic Inheritance)

Digenic inheritance refers to clinical phenotypes caused by the coinheritance of variants in two unlinked genes. Thus, one form of retinitis pigmentosa is due to the coinheritance of variants in the *RDS* gene on 6p and the *ROM* gene on 11q [641]. Individuals with either one or the other variant do not suffer from the disease. In similar vein, digenic inheritance of variants in the *MITF* and *TYR* genes has been reported as a cause of Waardenburg

syndrome type 2 in conjunction with ocular albinism [642]. This phenomenon may be common in polygenic disorders and in disorders with “low penetrance.”

Triallelic inheritance refers to clinical phenotypes, with apparent recessive mode of inheritance, caused by the coinheritance of three mutant alleles: two in one gene and one in another gene. An example of triallelic inheritance is provided by the Bardet–Biedl syndrome. There are pedigrees in which affected individuals have two mutant alleles in the *BBS6* gene and one mutant allele in the *BBS2* gene. Other pedigrees have two mutant alleles in the *BBS2* gene and one mutant allele in *BBS6* [643]. This type of inheritance indicates that some forms of BBS have a complex pattern of inheritance. As above, this phenomenon may be relevant in polygenic disorders and in disorders with “low penetrance.”

6.5.5 Different Variants in the Same Gene May Give Rise to Distinct Dominant and Recessive Forms of the Same Disease

vWF deficiency is a relatively common monogenic disease of blood coagulation. Many pathogenic variants have been studied in the *VWF* gene on chromosome 12p. A fraction of pathogenic variants (usually deletions, nonsense codons, or frameshift variants) cause vWF deficiency with a recessive mode of inheritance; other variants (mostly missense substitutions) are, however, associated with a dominant mode of inheritance of the vWF deficiency [644].

Although the majority of hitherto characterized growth hormone (*GH1*) gene lesions (including gross deletions and missense/nonsense variants) that underlie familial short stature are inherited in autosomal recessive fashion, there is a group of intron 3 splicing variants that are characterized by a dominant mode of inheritance [429]. These lesions result in the in-frame skipping of exon 3 encoding 40 amino acids including a Cys residue. The dominant negative nature of this type of mutation is thought to be explicable in terms of the participation of the resulting free unpaired cysteine residue in an illegitimate intermolecular disulfide linkage leading to dimerization of the mutant molecule with a normal GH molecule and inhibition of GH secretion.

6.6 WHY STUDY MUTATION?

Although the sequencing of the human genome was finished some time ago, its annotation is still far from complete [645–647]. Full exploitation of the emerging data,

specifically in relation to understanding the etiology of inherited disease and disease predisposition, is however, likely to be hampered by our ignorance of the basic processes underlying interindividual, interpopulation, and interspecies genetic diversity. At the population level, such an understanding is seen as essential for any meaningful interpretation of the prevalence/incidence patterns observed for diseases with a genetic basis. Within families, it is a prerequisite for being able to explain how interindividual variation arises and how variable phenotypic expression can be associated with identical gene lesions. Thus, for human genome sequence data to be useful in the context of molecular medicine, they must eventually be related to the genetic variation underlying human inherited disease. To this end, the meta-analysis of pathological germline variants in human genes should facilitate

1. the assessment of the spectrum of known genetic variation underlying human inherited disease [5],
2. the identification of factors determining the propensity of DNA sequences to undergo germline mutation [160],
3. the optimization of mutational screening strategies [648],
4. improvements in our ability to predict the clinical phenotype from knowledge of the mutant genotype [9,387],
5. the identification of disease states that exhibit incomplete mutational spectra, prompting the search for, and detection of, novel gene lesions associated with different clinical phenotypes [649],
6. extrapolation toward the genetic basis of other, more complex traits and diseases [650],
7. improvements in our understanding of the biological function(s) of a given protein [388],
8. meaningful comparison between the mechanisms of mutagenesis underlying both inherited and somatic diseases [599],
9. studies of human genetic diseases in their evolutionary context [201].

Our genes have evolved slowly, probably via a myriad of meandering and circuitous pathways, escorted through the millennia of erratic environmental influences by the molding force of natural selection. Perhaps this hesitant evolutionary past accounts for present day genes containing, encoded within their nucleotide sequences, the potential seeds of their own destruction. How apt in this context is the poet's description of nature: “so careful of the type she seems, so careless of

the single life” (Alfred Lord Tennyson, “In memoriam A.H.H.”, 1850).

ACKNOWLEDGMENTS

The authors wish to thank Professor Michael Krawczak (Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität Kiel, Germany) for his contributions to earlier published versions of this chapter and Peter Stenson and Eddy Ball for their provision of HGMD data. Earlier versions of this chapter have been published in other textbooks. We thank the editor, professor Bruce Korf for the expert editorial comments and suggestions for the improvement of this chapter.

REFERENCES

- [1] Antonarakis SE. Genomic databases: a WHO affair. *Science* 2017;356:812–3.
- [2] Ingram VM. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature* 1956;178:792–4.
- [3] Orkin SH, Alter BP, Altay C, Mahoney MJ, Lazarus H, Hobbins JC, Nathan DG. Application of endonuclease mapping to the analysis and prenatal diagnosis of thalassemias caused by globin-gene deletion. *N Engl J Med* 1978;299:166–72.
- [4] Chang JC, Kan YW. Beta 0 thalassemia, a nonsense mutation in man. *Proc Natl Acad Sci U S A* 1979;76:2886–9.
- [5] Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* 2010;31:631–55.
- [6] Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J, Dopazo H. Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J Mol Biol* 2006;358:1390–404.
- [7] Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Hum Mutat* 2008;29:198–204.
- [8] Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* 2002;315:771–86.
- [9] Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res* 2009;19:1562–9.
- [10] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009;25:2744–50.
- [11] Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* 2001;10:2319–28.
- [12] Nakken S, Rodland EA, Hovig E. Impact of DNA physical properties on local sequence bias of human mutation. *Hum Mutat* 2010;31:1316–25.
- [13] Alkuraya FS. Discovery of mutations for Mendelian disorders. *Hum Genet* 2016;135:615–23.
- [14] Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. *Nat Rev Genet* 2006;7:277–82.
- [15] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, DeFlaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation C. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–91.
- [16] MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoiyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;508:469–76.
- [17] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, Committee ALQA. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–24.
- [18] Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat* 2015;36:928–30.

- [19] Shendure J, Akey JM. The origins, determinants, and consequences of human mutations. *Science* 2015;349:1478–83.
- [20] Vogel F, Motulsky A. *Human genetics*. Verlag Berlin: Springer; 1986.
- [21] Antonarakis SE, Kazazian Jr HH, Orkin SH. DNA polymorphism and molecular pathology of the human globin gene clusters. *Hum Genet* 1985;69:1–14.
- [22] Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum Genet* 1985;69:201–5.
- [23] Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet* 1998;19:233–40.
- [24] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–33.
- [25] Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077–82.
- [26] Orkin SH, Kazazian Jr HH, Antonarakis SE, Ostrer H, Goff SC, Sexton JP. Abnormal RNA processing due to the exon mutation of beta E-globin gene. *Nature* 1982;300:768–9.
- [27] Crawford DC, Akey DT, Nickerson DA. The patterns of natural variation in human genes. *Annu Rev Genomics Hum Genet* 2005;6:287–312.
- [28] Nishihara S, Narimatsu H, Iwasaki H, Yazawa S, Akamatsu S, Ando T, Seno T, Narimatsu I. Molecular genetic analysis of the human Lewis histo-blood group system. *J Biol Chem* 1994;269:29271–8.
- [29] Kelly RJ, Rouquier S, Giorgi D, Lennon GG, Lowe JB. Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating non-sense mutation commonly correlates with the non-secretor phenotype. *J Biol Chem* 1995;270:4640–9.
- [30] Rebbeck TR. Molecular epidemiology of the human glutathione S-transferase genotypes GSTM1 and GSTT1 in cancer susceptibility. *Cancer Epidemiol Biomark Prev* 1997;6:733–43.
- [31] Thein SL. Genetic insights into the clinical diversity of beta thalassaemia. *Br J Haematol* 2004;124:264–74.
- [32] Green H, Djian P. Consecutive actions of different gene-altering mechanisms in the evolution of involucre. *Mol Biol Evol* 1992;9:977–1017.
- [33] Dawson SJ, Wiman B, Hamsten A, Green F, Humphries S, Henney AM. The two allele sequences of a common polymorphism in the promoter of the plasminogen activator inhibitor-1 (PAI-1) gene respond differently to interleukin-1 in HepG2 cells. *J Biol Chem* 1993;268:10739–45.
- [34] Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* 2000;67:881–900.
- [35] Rigat B, Hubert C, Alhenc-Gelas F, Cambien F, Corvol P, Soubrier F. An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J Clin Invest* 1990;86:1343–6.
- [36] Small K, Iber J, Warren ST. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat Genet* 1997;16:96–9.
- [37] Neitz M, Neitz J, Grishok A. Polymorphism in the number of genes encoding long-wavelength-sensitive cone pigments among males with normal color vision. *Vision Res* 1995;35:2395–407.
- [38] Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 2006;7:61–80.
- [39] Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. *Hum Mol Genet* 2006;15(Spec No 1):R9–16.
- [40] ElSharawy A, Hundrieser B, Brosch M, Wittig M, Huse K, Platzer M, Becker A, Simon M, Rosenstiel P, Schreiber S, Krawczak M, Hampe J. Systematic evaluation of the effect of common SNPs on pre-mRNA splicing. *Hum Mutat* 2009;30:625–32.
- [41] Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. *Nature* 2005;437:1299–320.
- [42] Consortium TIH. The International HapMap project. *Nature* 2003;426:789–96.
- [43] Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. *Science* 2005;307:1072–9.

- [44] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
- [45] Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;12:745–55.
- [46] Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, Barnett AH, Bates C, Bellary S, Bockett NA, Giorda K, Griffiths CJ, Hemingway H, Jia Z, Kelly MA, Khawaja HA, Lek M, McCarthy S, McEachan R, O'Donnell-Luria A, Paigen K, Parisinos CA, Sheridan E, Southgate L, Tee L, Thomas M, Xue Y, Schnall-Levin M, Petkov PM, Tyler-Smith C, Maher ER, Trembath RC, MacArthur DG, Wright J, Durbin R, van Heel DA. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* 2016;352:474–7.
- [47] Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C, Genomes Project C. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet* 2012;91:1022–32.
- [48] Jeffreys AJ, Wilson V, Thein SL. Hypervariable 'minisatellite' regions in human DNA. *Nature* 1985;314:67–73.
- [49] Wyman AR, White R. A highly polymorphic locus in human DNA. *Proc Natl Acad Sci U S A* 1980;77:6754–8.
- [50] Jeffreys AJ, Neumann R, Wilson V. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 1990;60:473–85.
- [51] Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 1985;230:1350–4.
- [52] Litt M, Luty JA. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 1989;44:397–401.
- [53] Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 1989;44:388–96.
- [54] Economou EP, Bergen AW, Warren AC, Antonarakis SE. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc Natl Acad Sci U S A* 1990;87:2951–4.
- [55] Anagnou NP, O'Brien SJ, Shimada T, Nash WG, Chen MJ, Nienhuis AW. Chromosomal organization of the human dihydrofolate reductase genes: dispersion, selective amplification, and a novel form of polymorphism. *Proc Natl Acad Sci U S A* 1984;81:5170–4.
- [56] Cooper DN. Human gene evolution. Oxford: Bios Scientific; 1999.
- [57] Buckland PR. Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann Med* 2003;35:308–15.
- [58] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–51.
- [59] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. *Science* 2004;305:525–8.
- [60] Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 2005;77:78–88.
- [61] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwork C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. *Nature* 2006;444:444–54.
- [62] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, MacArthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;464:704–12.
- [63] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stutz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470:59–65.

- [64] Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 2010;143:837–47.
- [65] Balasubramanian S, Habegger L, Frankish A, MacArthur DG, Harte R, Tyler-Smith C, Harrow J, Gerstein M. Gene inactivation and its implications for annotation in the era of personal genomics. *Genes Dev* 2011;25:1–10.
- [66] Dear PH. Copy-number variation: the end of the human genome? *Trends Biotechnol* 2009;27:448–54.
- [67] Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006;38:75–81.
- [68] McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM. Common deletion polymorphisms in the human genome. *Nat Genet* 2006;38:86–92.
- [69] Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 2006;38:82–5.
- [70] MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet* 2010;19:R125–30.
- [71] Waalen J, Beutler E. Genetic screening for low-penetrance variants in protein-coding genes. *Annu Rev Genomics Hum Genet* 2009;10:431–50.
- [72] Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. Genetic variation in an individual human exome. *PLoS Genet* 2008;4:e1000160.
- [73] Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *Am J Hum Genet* 2009;84:224–34.
- [74] Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, Imanishi T. Distribution and effects of nonsense polymorphisms in human genes. *PLoS One* 2008;3:e3393.
- [75] Hsiao TL, Vitkup D. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet* 2008;4:e1000014.
- [76] Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Hum Genomics* 2010;4:406–10.
- [77] Krawczak M, Chuzhanova NA, Stenson PD, Johansen BN, Ball EV, Cooper DN. Changes in primary DNA sequence complexity influence the phenotypic consequences of mutations in human gene regulatory regions. *Hum Genet* 2000;107:362–5.
- [78] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M, Cooper DN. Human gene mutation database (HGMD): 2003 update. *Hum Mutat* 2003;21:577–81.
- [79] Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;136:665–77.
- [80] Antonarakis SE, McKusick VA. OMIM passes the 1,000-disease-gene mark. *Nat Genet* 2000;25:11.
- [81] Loeb LA, Kunkel TA. Fidelity of DNA synthesis. *Annu Rev Biochem* 1982;51:429–57.
- [82] Krawczak M, Ball EV, Cooper DN. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet* 1998;63:474–88.
- [83] Antonarakis SE, Krawczak M, Cooper DN. The nature and mechanisms of human gene mutation. In: Scriver CR, Beaudet AL, Valle D, et al., editors. *The metabolic and molecular bases of inherited disease*. New York: McGraw-Hill; 2001. p. 343–77.
- [84] Youssoufian H, Antonarakis SE, Bell W, Griffin AM, Kazazian Jr HH. Nonsense and missense mutations in hemophilia A: estimate of the relative mutation rate at CG dinucleotides. *Am J Hum Genet* 1988;42:718–25.
- [85] Youssoufian H, Kazazian Jr HH, Phillips DG, Aronis S, Tsiftis G, Brown VA, Antonarakis SE. Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots. *Nature* 1986;324:380–2.
- [86] Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. *Hum Genet* 1988;78:151–5.
- [87] Antonarakis SE, Kazazian HH, Tuddenham EG. Molecular etiology of factor VIII deficiency in hemophilia A. *Hum Mutat* 1995;5:1–22.
- [88] Cooper DN. Eukaryotic DNA methylation. *Hum Genet* 1983;64:315–33.
- [89] Driscoll DJ, Migeon BR. Sex difference in methylation of single-copy genes in human meiotic germ cells: implications for X chromosome inactivation, parental imprinting, and origin of CpG mutations. *Somat Cell Mol Genet* 1990;16:267–82.
- [90] Hurst LD, Ellegren H. Sex biases in the mutation rate. *Trends Genet* 1998;14:446–52.
- [91] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462:315–22.

- [92] Rodenhiser DI, Andrews JD, Mancini DN, Jung JH, Singh SM. Homonucleotide tracts, short repeats and CpG/CpNpG motifs are frequent sites for heterogeneous mutations in the neurofibromatosis type 1 (NF1) tumour-suppressor gene. *Mutat Res* 1997;373:185–95.
- [93] Cheung LW, Lee YF, Ng TW, Ching WK, Khoo US, Ng MK, Wong AS. CpG/CpNpG motifs in the coding region are preferred sites for mutagenesis in the breast cancer susceptibility genes. *FEBS Lett* 2007;581:4668–74.
- [94] Kondrashov AS. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 2003;21:12–27.
- [95] Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 2007;80:727–39.
- [96] Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 2008;4:e1000083.
- [97] Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 2006;173:891–900.
- [98] Bacolla A, Wang G, Jain A, Chuzhanova NA, Cer RZ, Collins JR, Cooper DN, Bohr VA, Vasquez KM. Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. *J Biol Chem* 2011.
- [99] Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002;3:285–98.
- [100] Gorlov IP, Kimmel M, Amos CI. Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Hum Mol Genet* 2006;15:1143–50.
- [101] Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C. Silent (synonymous) SNPs: should we care about them? *Methods Mol Biol* 2009;578:23–39.
- [102] Sanford JR, Wang X, Mort M, Vanduy N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* 2009;19:381–94.
- [103] Sauna ZE, Okunji C, Hunt RC, Gupta T, Allen CE, Plum E, Blaisdell A, Grigoryan V, Geetha S, Fathke R, Soejima K, Kimchi-Sarfaty C. Characterization of conformation-sensitive antibodies to ADAMTS13, the von Willebrand cleavage protease. *PLoS One* 2009;4:e6506.
- [104] Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 2007;8:749–61.
- [105] Nackley AG, Shabalina SA, Lambert JE, Conrad MS, Gibson DG, Spiridonov AN, Satterfield SK, Diatchenko L. Low enzymatic activity haplotypes of the human catechol-O-methyltransferase gene: enrichment for marker SNPs. *PLoS One* 2009;4:e5237.
- [106] Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskiy O, Makarov SS, Maixner W, Diatchenko L. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 2006;314:1930–3.
- [107] Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcano AM, Ambudkar SV, Gottesman MM. A “silent” polymorphism in the MDR1 gene changes substrate specificity. *Science* 2007;315:525–8.
- [108] Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol* 2008;383:281–91.
- [109] Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, Karra K, Davydov E, Batzoglu S, Myers RM, Sidow A. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res* 2010;20:301–10.
- [110] Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O’Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Sholders CC, Proudfoot NJ. The structure and evolution of the human beta-globin gene family. *Cell* 1980;21:653–68.
- [111] Kunkel TA. The mutational specificity of DNA polymerases-alpha and -gamma during in vitro DNA synthesis. *J Biol Chem* 1985;260:12866–74.
- [112] Cooper DN. *Human gene mutation*. Oxford: Bios Scientific; 1993.
- [113] Schmucker B, Krawczak M. Meiotic microdeletion breakpoints in the BRCA1 gene are significantly associated with symmetric DNA-sequence elements. *Am J Hum Genet* 1997;61:1454–6.
- [114] Krawczak M, Cooper DN. Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. *Hum Genet* 1991;86:425–41.
- [115] Cooper DN, Krawczak M. Mechanisms of insertional mutagenesis in human genes causing genetic disease. *Hum Genet* 1991;87:409–15.
- [116] Ball EV, Stenson PD, Abeyasinghe SS, Krawczak M, Cooper DN, Chuzhanova NA. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 2005;26:205–13.
- [117] Kondrashov AS, Rogozin IB. Context of deletions and insertions in human coding sequences. *Hum Mutat* 2004;23:177–85.

- [118] Kamat MA, Bacolla A, Cooper DN, Chuzhanova N. A role for non-B DNA forming sequences in mediating microlesions causing human inherited disease. *Hum Mutat* 2016;37:65–73.
- [119] Caskey CT, Pizzuti A, Fu YH, Fenwick Jr RG, Nelson DL. Triplet repeat mutations in human disease. *Science* 1992;256:784–9.
- [120] Mandel JL. Questions of expansion. *Nat Genet* 1993;4:8–9.
- [121] Rousseau F, Heitz D, Mandel J.L.. The unstable and methylatable mutations causing the fragile X syndrome. *Hum Mutat* 1992;1:91–6.
- [122] Harper PS, Harley HG, Reardon W, Shaw DJ. Anticipation in myotonic dystrophy: new light on an old problem. *Am J Hum Genet* 1992;51:10–6.
- [123] Van Esch H. The Fragile X premutation: new insights and clinical consequences. *Eur J Med Genet* 2006;49:1–8.
- [124] Housman D. Gain of glutamines, gain of function? *Nat Genet* 1995;10:3–4.
- [125] Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, Day JW, Ranum LP. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* 2001;293:864–7.
- [126] Savkur RS, Philips AV, Cooper TA, Dalton JC, Moseley ML, Ranum LP, Day JW. Insulin receptor splicing alteration in myotonic dystrophy type 2. *Am J Hum Genet* 2004;74:1309–13.
- [127] Fu Y-H, Kuhl D, Pizzuti A, Pieretti M, Sutcliffe JS, Richards CS, Verkerk AJMH, Holden J, Fenwick RJ, Warren ST, Oostra BA, Nelson DL, Caskey CT. Variation of the CGG repeat at the Fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* 1991;67:1047–58.
- [128] Conway GS, Hettiarachchi S, Murray A, Jacobs PA. Fragile X premutations in familial premature ovarian failure. *Lancet* 1995;346:309–10.
- [129] Jacquemont S, Hagerman RJ, Leehey M, Grigsby J, Zhang L, Brunberg JA, Greco C, Des Portes V, Jardini T, Levine R, Berry-Kravis E, Brown WT, Schaeffer S, Kissel J, Tassone F, Hagerman PJ. Fragile X premutation tremor/ataxia syndrome: molecular, clinical, and neuroimaging correlates. *Am J Hum Genet* 2003;72:869–78.
- [130] Kang S, Ohshima K, Jaworski A, Wells R.D.. CTG triplet repeats from the myotonic dystrophy gene are expanded in *Escherichia coli* distal to the replication origin as a single large event. *J Mol Biol* 1996;258:543–7.
- [131] Chung MY, Ranum LP, Duvick LA, Servadio A, Zoghbi HY, Orr HT. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat Genet* 1993;5:254–8.
- [132] Wells RD, Warren AC. Genetic instabilities and hereditary neurological disorders. San Diego: Academic Press; 1998.
- [133] Brais B, Bouchard JP, Xie YG, Rochefort DL, Chretien N, Tome FM, Lafreniere RG, Rommens JM, Uyama E, Nohira O, Blumen S, Korczyn AD, Heutink P, Mathieu J, Duranceau A, Codere F, Fardeau M, Rouleau GA. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat Genet* 1998;18:164–7.
- [134] Muragaki Y, Mundlos S, Upton J, Olsen BR. Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. *Science* 1996;272:548–51.
- [135] Pearson C.E., Nichol Edamura K., Cleary J.D.. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* 2005;6:729–42.
- [136] Lalioti MD, Scott HS, Buresi C, Rossier C, Bottani A, Morris MA, Malafosse A, Antonarakis SE. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* 1997;386:847–51.
- [137] Lalioti MD, Scott HS, Antonarakis SE. Altered spacing of promoter elements due to the dodecamer repeat expansion contributes to reduced expression of the cystatin B gene in EPM1. *Hum Mol Genet* 1999;8:1791–8.
- [138] Matsuura T, Yamagata T, Burgess DL, Rasmussen A, Grewal RP, Watase K, Khajavi M, McCall AE, Davis CF, Zu L, Achari M, Pulst SM, Alonso E, Noebels JL, Nelson DL, Zoghbi HY, Ashizawa T. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat Genet* 2000;26:191–4.
- [139] Verpy E., Leibovici M., Zwaenepoel I., Liu X.Z., Gal A., Salem N, Mansour A, Blanchard S, Kobayashi I, Keats BJ, Slim R, Petit C. A defect in harmonin, a PDZ domain-containing protein expressed in the inner ear sensory hair cells, underlies Usher syndrome type 1C. *Nat Genet* 2000;26:51–5.
- [140] Yu S, Mangelsdorf M, Hewett D, Hobson L, Baker E, Eyre HJ, Lapsys N, Le Paslier D, Doggett NA, Sutherland GR, Richards RI. Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* 1997;88:367–74.
- [141] Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, Rosenfeld JA, Ballif BC, Shaffer LG, Graves TA, Wilson RK, Schwartz DC, Eichler EE. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* 2010;42:745–50.
- [142] Ciccone R, Mattina T, Giorda R, Bonaglia MC, Rocchi M, Pramparo T, Zuffardi O. Inversion polymorphisms

- and non-contiguous terminal deletions: the cause and the (unpredicted) effect of our genome architecture. *J Med Genet* 2006;43:e19.
- [143] Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol I, Schinzel A, Estivill X, Zuffardi O. Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum Mol Genet* 2003;12:849–58.
- [144] Hobart HH, Morris CA, Mervis CB, Pani AM, Kistler DJ, Rios CM, Kimberley KW, Gregg RG, Bray-Ward P. Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. *Am J Med Genet C Semin Med Genet* 2010;154C:220–8.
- [145] Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, Matsumoto N. Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdeletion. *Am J Hum Genet* 2005;76:52–67.
- [146] Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Mehdi SQ, Kajuna SL, Barta C, Kungulilo S, Karoma NJ, Lu RB, Zhukova OV, Kim JJ, Comas D, Siniscalco M, New M, Li P, Li H, Manolopoulos VG, Speed WC, Rajeevan H, Pakstis AJ, Kidd JR, Kidd KK. The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am J Hum Genet* 2010;86:161–71.
- [147] Rao PN, Li W, Vissers LE, Veltman JA, Ophoff RA. Recurrent inversion events at 17q21.31 microdeletion locus are linked to the MAPT H2 haplotype. *Cytogenet Genome Res* 2010;129:275–9.
- [148] Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ, Lupski JR. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *Am J Hum Genet* 2011;89:580–8.
- [149] Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L, South ST, Coleman K, Fernhoff PM, Deray MJ, Rosengren S, Roeder ER, Enciso VB, Chinault AC, Patel A, Kang SH, Shaw CA, Lupski JR, Cheung SW. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res* 2011;21:33–46.
- [150] Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G, Kallicki J, Anderson P, Tsalenko A, Yamada NA, Tsang P, Kaul R, Wilson RK, Bruhn L, Eichler EE. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat Methods* 2010;7:365–71.
- [151] Ashley T, Gaeth AP, Inagaki H, Seftel A, Cohen MM, Anderson LK, Kurahashi H, Emanuel BS. Meiotic recombination and spatial proximity in the etiology of the recurrent t(11;22). *Am J Hum Genet* 2006;79:524–38.
- [152] Edelmann L, Spiteri E, Koren K, Pulijaal V, Bialer MG, Shanske A, Goldberg R, Morrow BE. AT-rich palindromes mediate the constitutional t(11;22) translocation. *Am J Hum Genet* 2001;68:1–13.
- [153] Kurahashi H, Inagaki H, Ohye T, Kogo H, Tsutsumi M, Kato T, Tong M, Emanuel BS. The constitutional t(11;22): implications for a novel mechanism responsible for gross chromosomal rearrangements. *Clin Genet* 2010;78:299–309.
- [154] Soutoglou E, Dorn JF, Sengupta K, Jasin M, Nussenzweig A, Ried T, Danuser G, Misteli T. Positional stability of single double-strand breaks in mammalian cells. *Nat Cell Biol* 2007;9:675–82.
- [155] Nikiforova MN, Stringer JR, Blough R, Medvedovic M, Fagin JA, Nikiforov YE. Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. *Science* 2000;290:138–41.
- [156] Meaburn KJ, Misteli T, Soutoglou E. Spatial genome organization in the formation of chromosomal translocations. *Semin Cancer Biol* 2007;17:80–90.
- [157] Wijchers PJ, de Laat W. Genome organization influences partner selection for chromosomal rearrangements. *Trends Genet* 2011;27:63–71.
- [158] Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. *Hum Mutat* 2003;22:229–44.
- [159] Chuzhanova N, Abeysinghe SS, Krawczak M, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer II: potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. *Hum Mutat* 2003;22:245–51.
- [160] Cooper DN, Bacolla A, Ferec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM. On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Hum Mutat* 2011;32:1075–99.
- [161] Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Complex gene rearrangements caused by serial replication slippage. *Hum Mutat* 2005a;26:125–34.
- [162] Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. *Hum Mutat* 2005b;26:362–73.

- [163] Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Hum Mutat* 2005c;25:207–21.
- [164] Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 2007;131:1235–47.
- [165] Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* 2009;5:e1000327.
- [166] Chen JM, Cooper DN, Ferec C, Kehrer-Sawatzki H, Patrinos GP. Genomic rearrangements in inherited disease and cancer. *Semin Cancer Biol* 2010;20:222–33.
- [167] Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, Rosenberg C, Ignatius J, Raynaud M, Hollanders K, Govaerts K, Vandenreijt K, Niel F, Blanc P, Stevenson RE, Fryns JP, Marynen P, Schwartz CE, Froyen G. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res* 2008;18:847–58.
- [168] Kulikowski LD, Yoshimoto M, da Silva Bellucco FT, Belangero SI, Christofolini DM, Pacanaro AN, Bortolai A, Smith Mde A, Squire JA, Melaragno MI. Cytogenetic molecular delineation of a terminal 18q deletion suggesting neo-telomere formation. *Eur J Med Genet* 2010;53:404–7.
- [169] Zuffardi O, Bonaglia M, Ciccone R, Giorda R. Inverted duplications deletions: underdiagnosed rearrangements? *Clin Genet* 2009;75:505–13.
- [170] Hannes F, Van Houdt J, Quarrell OW, Poot M, Hochstenbach R, Fryns JP, Vermeesch JR. Telomere healing following DNA polymerase arrest-induced breakages is likely the main mechanism generating chromosome 4p terminal deletions. *Hum Mutat* 2010;31:1343–51.
- [171] Cha RS, Kleckner N. ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. *Science* 2002;297:602–6.
- [172] Pelletier R, Krasilnikova MM, Samadashwily GM, Lahue R, Mirkin SM. Replication and expansion of trinucleotide repeats in yeast. *Mol Cell Biol* 2003;23:1349–57.
- [173] Koszul R, Caburet S, Dujon B, Fischer G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* 2004;23:234–43.
- [174] Choi BO, Kim NK, Park SW, Hyun YS, Jeon HJ, Hwang JH, Chung KW. Inheritance of Charcot-Marie-Tooth disease 1A with rare nonrecurrent genomic rearrangement. *Neurogenetics* 2010.
- [175] Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* 2009;41:849–53.
- [176] Ankala A, Kohn JN, Hegde A, Meka A, Ephrem CL, Askree SH, Bhide S, Hegde MR. Aberrant firing of replication origins potentially explains intragenic non-recurrent rearrangements within genes, including the human DMD gene. *Genome Res* 2011.
- [177] Doksan Y, Bermejo R, Fiorani S, Haber JE, Foiani M. Replicon dynamics, dormant origin firing, and terminal fork integrity after double-strand break formation. *Cell* 2009;137:247–58.
- [178] Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 2016;17:224–38.
- [179] Ballabio A, Carrozzo R, Parenti G, Gil A, Zollo M, Persico MG, Gillard E, Affara N, Yates J, Ferguson-Smith MA, et al. Molecular heterogeneity of steroid sulfatase deficiency: a multicenter study on 57 unrelated patients, at DNA and protein levels. *Genomics* 1989;4:36–40.
- [180] den Dunnen JT, Bakker E, Breteler EG, Pearson PL, van Ommen GJ. Direct detection of more than 50% of the Duchenne muscular dystrophy mutations by field inversion gels. *Nature* 1987;329:640–2.
- [181] Nicholls RD, Fischel-Ghodsian N, Higgs DR. Recombination at the human alpha-globin gene cluster: sequence features and topological constraints. *Cell* 1987;49:369–78.
- [182] Vnencak-Jones CL, Phillips 3rd JA. Hot spots for growth hormone gene deletions in homologous regions outside of Alu repeats. *Science* 1990;250:1745–8.
- [183] Embury SH, Miller JA, Dozy AM, Kan YW, Chan V, Todd D. Two different molecular organizations account for the single alpha-globin gene of the alpha-thalassemia-2 genotype. *J Clin Invest* 1980;66:1319–25.
- [184] Goossens M, Dozy AM, Embury SH, Zachariades Z, Hadjiminis MG, Stamatoyannopoulos G, Kan YW. Triplicated alpha-globin loci in humans. *Proc Natl Acad Sci U S A* 1980;77:518–21.
- [185] Baglioni C. The fusion of two peptide chains in hemoglobin Lepore and its interpretation as a genetic deletion. *Proc Natl Acad Sci U S A* 1962;48:1880–6.
- [186] Shapiro LJ, Yen P, Pomerantz D, Martin E, Rolewicz L, Mohandas T. Molecular studies of deletions at the human steroid sulfatase locus. *Proc Natl Acad Sci U S A* 1989;86:8477–81.
- [187] Guioli S, Incerti B, Zanaria E, Bardoni B, Franco B, Taylor K, Ballabio A, Camerino G. Kallmann syndrome due to a translocation resulting in an X/Y fusion gene. *Nat Genet* 1992;1:337–40.

- [188] Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *Am J Hum Genet* 1998;62:1023–33.
- [189] Dorschner MO, Sybert VP, Weaver M, Pletcher BA, Stephens K. NF1 microdeletion breakpoints are clustered at flanking repetitive sequences. *Hum Mol Genet* 2000;9:35–46.
- [190] Francke U. Williams-Beuren syndrome: genes and mechanisms. *Hum Mol Genet* 1999;8:1947–54.
- [191] Juyal RC, Figuera LE, Hauge X, Elsea SH, Lupski JR, Greenberg F, Baldini A, Patel PI. Molecular analyses of 17p11.2 deletions in 62 Smith-Magenis syndrome patients. *Am J Hum Genet* 1996;58:998–1007.
- [192] Edelmann L, Pandita RK, Morrow BE. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am J Hum Genet* 1999;64:1076–86.
- [193] Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA, Driscoll DA, McDonald-McGinn DM, Zackai EH, Budarf ML, Emanuel BS. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* 2000;9:489–501.
- [194] Christian SL, Fantes JA, Mewborn SK, Huang B, Ledbetter DH. Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). *Hum Mol Genet* 1999;8:1025–37.
- [195] Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid BM, Schoumans J, Knoers NV, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BB. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 2006;38:999–1001.
- [196] Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 2006;38:1038–42.
- [197] Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, Krepisch-Santos AC, Varela MC, Koiffmann CP, Lees AJ, Rosenberg C, Firth HV, de Silva R, Carter NP. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 2006;38:1032–7.
- [198] Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K. A common inversion under selection in Europeans. *Nat Genet* 2005;37:129–37.
- [199] Ji Y, Eichler EE, Schwartz S, Nicholls RD. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* 2000;10:597–610.
- [200] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. *Science* 2002;297:1003–7.
- [201] Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* 2004;13(Spec No 1):R57–64.
- [202] Deininger PL, Batzer MA. Alu repeats and human disease. *Mol Genet Metab* 1999;67:183–93.
- [203] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissen-

- bach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de Ja-Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
- [204] Lehrman MA, Goldstein JL, Russell DW, Brown MS. Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholesterolemia. *Cell* 1987;48:827–35.
- [205] Stoppa-Lyonnet D, Duponchel C, Meo T, Laurent J, Carter PE, Arala-Chaves M, Cohen JH, Dewald G, Goetz J, Hauptmann G, et al. Recombinational biases in the rearranged C1-inhibitor genes of hereditary angioedema patients. *Am J Hum Genet* 1991;49:1055–62.
- [206] Kornreich R, Bishop DF, Desnick RJ. α -galactosidase A gene rearrangements causing Fabry disease. *J Biol Chem* 1990;265:9319–26.
- [207] Abo-Dalo B, Kutsche K, Mautner V, Kluwe L. Large intragenic deletions of the NF2 gene: breakpoints and associated phenotypes. *Genes Chromosomes Cancer* 2010;49:171–5.
- [208] Champion KJ, Basehore MJ, Wood T, Destree A, Van-nuffel P, Maystadt I. Identification and characterization of a novel homozygous deletion in the α -N-acetylglucosaminidase gene in a patient with Sanfilippo type B syndrome (mucopolysaccharidosis IIIB). *Mol Genet Metab* 2010;100:51–6.
- [209] Cozar M, Bembi B, Dominissini S, Zampieri S, Vilageliu L, Grinberg D, Dardis A. Molecular characterization of a new deletion of the GBA1 gene due to an inter Alu recombination event. *Mol Genet Metab* 2010.
- [210] Gentsch M, Kaczmarczyk A, van Leeuwen K, de Boer M, Kaus-Drobek M, Dagher MC, Kaiser P, Arkwright PD, Gahr M, Rosen-Wolff A, Bochtler M, Secord E, Britto-Williams P, Saifi GM, Maddalena A, Dbaibo G, Bustamante J, Casanova JL, Roos D, Roesler J. Alu-repeat-induced deletions within the NCF2 gene causing p67-phox-deficient chronic granulomatous disease (CGD). *Hum Mutat* 2010;31:151–8.
- [211] Goldmann R, Tichy L, Freiburger T, Zapletalova P, Letocha O, Soska V, Fajkus J, Fajkusova L. Genomic characterization of large rearrangements of the LDLR gene in Czech patients with familial hypercholesterolemia. *BMC Med Genet* 2010;11:115.
- [212] Resta N, Giorda R, Bagnulo R, Beri S, Della Mina E, Stella A, Piglionica M, Susca FC, Guanti G, Zuffardi O, Ciccone R. Breakpoint determination of 15 large deletions in Peutz-Jeghers subjects. *Hum Genet* 2010;128:373–82.
- [213] Shlien A, Baskin B, Achatz MI, Stavropoulos DJ, Nichols KE, Hudgins L, Morel CF, Adam MP, Zhukova N, Rotin L, Novokmet A, Druker H, Shago M, Ray PN, Hainaut P, Malkin D. A common molecular mechanism underlies two phenotypically distinct 17p13.1 microdeletion syndromes. *Am J Hum Genet* 2010;87:631–42.
- [214] Tuohy TM, Done MW, Lewandowski MS, Shires PM, Saraiya DS, Huang SC, Neklason DW, Burt RW. Large intron 14 rearrangement in APC results in splice defect and attenuated FAP. *Hum Genet* 2010;127:359–69.
- [215] Yang Z, Funke BH, Cripe LH, Vick 3rd GW, Mancini-Dinardo D, Pena LS, Kanter RJ, Wong B, Westerfield BH, Varela JJ, Fan Y, Towbin JA, Vatta M. LAMP2 microdeletions in patients with Danon disease. *Circ Cardiovasc Genet* 2010;3:129–37.
- [216] Zhang F, Seeman P, Liu P, Weterman MA, Gonzaga-Jauregui C, Towne CF, Batish SD, De Vriendt E, De Jonghe P, Rautenstrauss B, Krause KH, Khajavi M, Posadka J, Vandenberghe A, Palau F, Van Maldergem L, Baas F, Timmerman V, Lupski JR. Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. *Am J Hum Genet* 2010;86:892–903.
- [217] Rudiger NS, Gregersen N, Kielland-Brandt MC. One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Res* 1995;23:256–60.
- [218] Roth DB, Wilson JH. Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. *Mol Cell Biol* 1986;6:4295–304.
- [219] Woods-Samuels P, Kazazian Jr HH, Antonarakis SE. Nonhomologous recombination in the human genome: deletions in the human factor VIII gene. *Genomics* 1991;10:94–101.
- [220] McNaughton JC, Cockburn DJ, Hughes G, Jones WA, Laing NG, Ray PN, Stockwell PA, Petersen GB. Is gene deletion in eukaryotes sequence-dependent? A study of nine deletion junctions and nineteen other deletion breakpoints in intron 7 of the human dystrophin gene. *Gene* 1998;222:41–51.

- [221] Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, O'Connell CD, Cooper DN, Wells RD. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci U S A* 2004;101:14162–7.
- [222] Abeysinghe SS, Stenson PD, Krawczak M, Cooper DN. Gross rearrangement breakpoint database (GRaBD). *Hum Mutat* 2004;23:219–21.
- [223] Mine M, Chen JM, Brivet M, Desguerre I, Marchant D, de Lonlay P, Bernard A, Ferec C, Abitbol M, Ricquier D, Marsac C. A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. *Hum Mutat* 2007;28:137–42.
- [224] Morisada N, Rendtorff ND, Nozu K, Morishita T, Miyakawa T, Matsumoto T, Hisano S, Iijima K, Tranebjaerg L, Shirahata A, Matsuo M, Kusuvara K. Branchio-oto-renal syndrome caused by partial EYA1 deletion due to LINE-1 insertion. *Pediatr Nephrol* 2010;25:1343–8.
- [225] Okubo M, Horinishi A, Saito M, Ebara T, Endo Y, Kaku K, Murase T, Eto M. A novel complex deletion-insertion mutation mediated by Alu repetitive elements leads to lipoprotein lipase deficiency. *Mol Genet Metab* 2007;92:229–33.
- [226] Schollen E, Keldermans L, Foulquier F, Briones P, Chabas A, Sanchez-Valverde F, Adamowicz M, Pronicka E, Wevers R, Matthijs G. Characterization of two unusual truncating PMM2 mutations in two CDG-Ia patients. *Mol Genet Metab* 2007;90:408–13.
- [227] Takasu M, Hayashi R, Maruya E, Ota M, Imura K, Kougo K, Kobayashi C, Saji H, Ishikawa Y, Asai T, Tokunaga K. Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon. *Tissue Antigens* 2007;70:144–50.
- [228] Awano H, Malueka RG, Yagi M, Okizuka Y, Takeshima Y, Matsuo M. Contemporary retrotransposition of a novel non-coding gene induces exon-skipping in dystrophin mRNA. *J Hum Genet* 2010.
- [229] Tabata A, Sheng JS, Ushikai M, Song YZ, Gao HZ, Lu YB, Okumura F, Iijima M, Mutoh K, Kishida S, Saheki T, Kobayashi K. Identification of 13 novel mutations including a retrotransposal insertion in SLC25A13 gene and frequency of 30 mutations found in patients with citrin deficiency. *J Hum Genet* 2008;53:534–45.
- [230] Kazazian Jr HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* 1988;332:164–6.
- [231] Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian Jr HH. Isolation of an active human transposable element. *Science* 1991;254:1805–8.
- [232] Ostertag EM, Kazazian Jr HH. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* 2001;35:501–38.
- [233] Kazazian Jr HH. Mobile elements and disease. *Curr Opin Genet Dev* 1998;8:343–50.
- [234] Woods-Samuels P, Wong C, Mathias SL, Scott AF, Kazazian Jr HH, Antonarakis SE. Characterization of a non-deleterious L1 insertion in an intron of the human factor VIII gene and further evidence of open reading frames in functional L1 elements. *Genomics* 1989;4:290–6.
- [235] Muratani K, Hada T, Yamamoto Y, Kaneko T, Shigeto Y, Ohue T, Furuyama J, Higashino K. Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proc Natl Acad Sci U S A* 1991;88:11315–9.
- [236] Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. Haemophilia B due to a *de novo* insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *Eur J Hum Genet* 1993;1:30–6.
- [237] Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. A *de novo* Alu insertion results in neurofibromatosis type 1. *Nature* 1991;353:864–6.
- [238] Li X, Scaringe WA, Hill KA, Roberts S, Mengos A, Careri D, Pinto MT, Kasper CK, Sommer SS. Frequency of recent retrotransposition events in the human factor IX gene. *Hum Mutat* 2001;17:511–9.
- [239] Chen JM, Stenson PD, Cooper DN, Ferec C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* 2005;117:411–27.
- [240] Audrezet MP, Chen JM, Raguene O, Chuzhanova N, Giteau K, Le Marechal C, Quere I, Cooper DN, Ferec C. Genomic rearrangements in the CFTR gene: extensive allelic heterogeneity and diverse mutational mechanisms. *Hum Mutat* 2004;23:343–57.
- [241] Bochukova EG, Roscioli T, Hedges DJ, Taylor IB, Johnson D, David DJ, Deininger PL, Wilkie AO. Rare mutations of FGFR2 causing apert syndrome: identification of the first partial gene deletion, and an Alu element insertion from a new subfamily. *Hum Mutat* 2009;30:204–11.
- [242] Oldridge M, Zackai EH, McDonald-McGinn DM, Iseki S, Morriss-Kay GM, Twigg SR, Johnson D, Wall SA, Jiang W, Theda C, Jabs EW, Wilkie AO. *De novo* alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *Am J Hum Genet* 1999;64:446–61.
- [243] Musova Z, Hedvicakova P, Mohrmann M, Tesarova M, Krepelova A, Zeman J, Sedlacek Z. A novel insertion of a rearranged L1 element in exon 44 of the dystrophin

- gene: further evidence for possible bias in retroposon integration. *Biochem Biophys Res Commun* 2006;347:145–9.
- [244] Narita N, Nishio H, Kitoh Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M. Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *J Clin Invest* 1993;91:1862–7.
- [245] Wimmer K, Callens T, Wernstedt A, Messiaen L. The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet* 2011;7:e1002371.
- [246] Wulff K, Gazda H, Schroder W, Robicka-Milewska R, Herrmann FH. Identification of a novel large F9 gene mutation—an insertion of an Alu repeated DNA element in exon e of the factor 9 gene. *Hum Mutat* 2000;15:299.
- [247] Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian Jr HH. Two independent retrotransposon insertions at the same site within the coding region of BTK. *Hum Mutat* 2005;25:324–5.
- [248] Scott HS, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S, Minoshima S, Younus F, Mehdi SQ, Radhakrishna U, Papasavvas MP, Gehrig C, Rossier C, Korostishevsky M, Gal A, Shimizu N, Bonne-Tamir B, Antonarakis SE. Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. *Nat Genet* 2001;27:59–63.
- [249] Turner C, Killoran C, Thomas NS, Rosenberg M, Chuzhanova NA, Johnston J, Kemel Y, Cooper DN, Biesecker LG. Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum Genet* 2003;112:303–9.
- [250] Millar DS, Tysoe C, Lazarou LP, Pilz DT, Mohammed S, Anderson K, Chuzhanova N, Cooper DN, Butler R. An isolated case of lissencephaly caused by the insertion of a mitochondrial genome-derived DNA sequence into the 5' untranslated region of the PAFAH1B1 (LIS1) gene. *Hum Genomics* 2010;4:384–93.
- [251] Lakich D, Kazazian Jr HH, Antonarakis SE, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* 1993;5:236–41.
- [252] Naylor JA, Green PM, Rizza CR, Giannelli F. Analysis of factor VIII mRNA reveals defects in everyone of 28 haemophilia A patients. *Hum Mol Genet* 1993;2:11–7.
- [253] Rossiter JP, Young M, Kimberland ML, Hutter P, Ketterling RP, Gitschier J, Horst J, Morris MA, Schaid DJ, de Moerloose P, Sommer SS, Kazazian HH, Antonarakis SE. Factor VIII gene inversions causing severe haemophilia A originate almost exclusively in male germ cells. *Hum Mol Genet* 1994;3:1035–9.
- [254] Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tonnesen T, Carlberg BM, Pettersson U. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet* 1995;4:615–21.
- [255] Jennings MW, Jones RW, Wood WG, Weatherall DJ. Analysis of an inversion within the human beta globin gene cluster. *Nucleic Acids Res* 1985;13:2897–907.
- [256] Karathanasis SK, Ferris E, Haddad IA. DNA inversion within the apolipoproteins AI/CIII/AIV-encoding gene cluster of certain patients with premature atherosclerosis. *Proc Natl Acad Sci U S A* 1987;84:7198–202.
- [257] Hu XY, Ray PN, Murphy EG, Thompson MW, Worton RG. Duplicational mutation at the Duchenne muscular dystrophy locus: its frequency, distribution, origin, and phenotype-genotype correlation. *Am J Hum Genet* 1990;46:682–95.
- [258] Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat Genet* 1992;2:292–300.
- [259] Woodward K, Kendall E, Vetrie D, Malcolm S. Pelizaeus-Merzbacher disease: identification of Xq22 proteolipid-protein duplications and characterization of breakpoints by interphase FISH. *Am J Hum Genet* 1998;63:207–17.
- [260] de Mollerat XJ, Gurrieri F, Morgan CT, Sangiorgi E, Everman DB, Gaspari P, Amiel J, Bamshad MJ, Lyle R, Blouin JL, Allanson JE, Le Marec B, Wilson M, Braverman NE, Radhakrishna U, Delozier-Blanchet C, Abbott A, Elghouzzi V, Antonarakis S, Stevenson RE, Munnich A, Neri G, Schwartz CE. A genomic rearrangement resulting in a tandem duplication is associated with split hand-split foot malformation 3 (SHFM3) at 10q24. *Hum Mol Genet* 2003;12:1959–71.
- [261] Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, Dubas F, Frebourg T, Campion D. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 2006;38:24–6.
- [262] Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, Hulihan M, Peuralinna T, Dutra A, Nussbaum R, Lincoln S, Crawley A, Hanson M, Maraganore D, Adler C, Cookson MR, Muenter M, Baptista M, Miller D, Blacato J, Hardy J, Gwinn-Hardy K. alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 2003;302:841.

- [263] Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferrec C. Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet* 2006;38:1372–4.
- [264] Fu W, Zhang F, Wang Y, Gu X, Jin L. Identification of copy number variation hotspots in human populations. *Am J Hum Genet* 2010;87:494–504.
- [265] Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 2009;41:1061–7.
- [266] Wang RT, Ahn S, Park CC, Khan AH, Lange K, Smith DJ. Effects of genome-wide copy number variation on expression in mammalian cells. *BMC Genomics* 2011;12:562.
- [267] Beckmann JS, Sharp AJ, Antonarakis SE. CNVs and genetic medicine (excitement and consequences of a rediscovery). *Cytogenet Genome Res* 2008;123:7–16.
- [268] de Smith AJ, Walters RG, Coin LJ, Steinfeld I, Yakhini Z, Sladek R, Froguel P, Blakemore AI. Small deletion variants have stable breakpoints commonly associated with alu elements. *PLoS One* 2008;3:e3104.
- [269] Henriksen CN, Chaigat E, Reymond A. Copy number variants, diseases and gene expression. *Hum Mol Genet* 2009;18:R1–8.
- [270] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010;61:437–55.
- [271] Lower KM, Hughes JR, De Gobbi M, Henderson S, Viprakasit V, Fisher C, Goriely A, Ayyub H, Sloane-Stanley J, Vernimmen D, Langford C, Garrick D, Gibbons RJ, Higgs DR. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc Natl Acad Sci U S A* 2009;106:21771–6.
- [272] Dathe K, Kjaer KW, Brehm A, Meinecke P, Nurnberg P, Neto JC, Brunoni D, Tommerup N, Ott CE, Klopocki E, Seemann P, Mundlos S. Duplications involving a conserved regulatory element downstream of BMP2 are associated with brachydactyly type A2. *Am J Hum Genet* 2009;84:483–92.
- [273] Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Robertson-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT. Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* 2006;439:851–5.
- [274] Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* 2006;79:439–48.
- [275] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 2005;307:1434–40.
- [276] Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Rocanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008;320:539–43.
- [277] Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, Saeed S, Hamilton-Shield J, Clayton-Smith J, O'Rahilly S, Hurles ME, Farooqi IS. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 2010;463:666–70.
- [278] McCarroll SA. Extending genome-wide association studies to copy-number variation. *Hum Mol Genet* 2008;17:R135–42.
- [279] Merikangas AK, Corvin AP, Gallagher L. Copy-number variants in neurodevelopmental disorders: promises and challenges. *Trends Genet* 2009;25:536–44.
- [280] Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 2010;12:e8.
- [281] Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'Arcy M, deBerardinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Takeda T, Rappaport EF, Grant SF, Berrettini W, Devoto M, Shaikh TH, Hakonarson H, White PS. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry* 2010;15:637–46.
- [282] Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garriss M, Soorya L, Klei L, Piven J, Meyer

- KJ, Anagnostou E, Sakurai T, Game RM, Rudd DS, Zurawiecki D, McDougle CJ, Davis LK, Miller J, Posey DJ, Michaels S, Kolevzon A, Silverman JM, Bernier R, Levy SE, Schultz RT, Dawson G, Owley T, McMahon WM, Wassink TH, Sweeney JA, Nurnberger JI, Coon H, Sutcliffe JS, Minshew NJ, Grant SF, Bucan M, Cook EH, Buxbaum JD, Devlin B, Schellenberg GD, Hakonarson H. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009;459:569–73.
- [283] Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Moller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Muhleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeny LA, Franke B, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreasen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nothen MM, Peltonen L, Collier DA, St Clair D, Stefansson K. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008;455:232–6.
- [284] Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chevre JC, Lecoecur C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP, Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrarini A, Hadjikhani N, Campion D, Guilmatre A, Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Beri-Dexheimer M, Jonveaux P, Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M, Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S, Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T, Metspalu A, Nelis M, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson P, Sjostrom L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Mannik K, Jarvelin MR, Pattou F, Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* 2010;463:671–5.
- [285] Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 2011;43:864–8.
- [286] Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, Nanda S, Druker H, Scherer SW, Malkin D. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci U S A* 2008;105:11264–9.
- [287] Fanciulli M, Petretto E, Aitman TJ. Gene copy number variation and common human disease. *Clin Genet* 2010;77:201–13.
- [288] Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L, Walsh T, McCarthy SE, Baker C, Mefford HC, Kidd JM, Browning SR, Browning BL, Dickel DE, Levy DL, Ballif BC, Platky K, Farber DM, Gowans GC, Wetherbee JJ, Asamoah A, Weaver DD, Mark PR, Dickerson J, Garg BP, Ellingwood SA, Smith R, Banks VC, Smith W, McDonald MT, Hoo JJ, French BN, Hudson C, Johnson JP, Ozmore JR, Moeschler JB, Surti U, Escobar LF, El-Khechen D, Gorski JL, Kussmann J, Salbert B, Lacassie Y, Biser A, McDonald-McGinn DM, Zackai EH, Deardorff MA, Shaikh TH, Haan E, Friend KL, Fichera M, Romano C, Gecz J, DeLisi LE, Sebat J, King MC, Shaffer LG, Eichler EE. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* 2010;42:203–9.
- [289] Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ, Papasian CJ, Recker RR, Song XP, Cheng J, Deng HW. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am J Hum Genet* 2008;83:663–74.
- [290] Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Wiklund F, Wiley K, Isaacs SD, Stattin P, Xu J, Duggan D, Carpten JD, Isaacs WB, Gronberg H, Zheng SL, Chang BL. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. *Cancer Res* 2009;69:2176–9.
- [291] Thean LF, Loi C, Ho KS, Koh PK, Eu KW, Cheah PY. Genome-wide scan identifies a copy number variable region at 3q26 that regulates PPM1L in APC mutation-negative familial colorectal cancer patients. *Genes Chromosomes Cancer* 2010;49:99–106.

- [292] Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 2007;8:762–75.
- [293] Tusie-Luna MT, White PC. Gene conversions and unequal crossovers between CYP21 (steroid 21-hydroxylase gene) and CYP21P involve different mechanisms. *Proc Natl Acad Sci U S A* 1995;92:10796–800.
- [294] Watnick TJ, Gandolph MA, Weber H, Neumann HP, Germino GG. Gene conversion is a likely cause of mutation in PKD1. *Hum Mol Genet* 1998;7:1239–43.
- [295] Gorlach A, Lee PL, Roesler J, Hopkins PJ, Christensen B, Green ED, Chanock SJ, Curnutte JT. A p47-phox pseudogene carries the most common mutation causing p47-phox- deficient chronic granulomatous disease. *J Clin Invest* 1997;100:1907–18.
- [296] Minegishi Y, Coustan-Smith E, Wang YH, Cooper MD, Campana D, Conley ME. Mutations in the human λ 5/14.1 gene result in B cell deficiency and agammaglobulinemia. *J Exp Med* 1998;187:71–7.
- [297] Eyal N, Wilder S, Horowitz M. Prevalent and rare mutations among Gaucher patients. *Gene* 1990;96:277–83.
- [298] Eikenboom JC, Vink T, Briet E, Sixma JJ, Reitsma PH. Multiple substitutions in the von Willebrand factor gene that mimic the pseudogene sequence. *Proc Natl Acad Sci U S A* 1994;91:2221–4.
- [299] Schollen E, Pardon E, Heykants L, Renard J, Doggett NA, Callen DF, Cassiman JJ, Matthijs G. Comparative analysis of the phosphomannomutase genes PMM1, PMM2 and PMM2psi: the sequence variation in the processed pseudogene is a reflection of the mutations found in the functional gene. *Hum Mol Genet* 1998;7:157–64.
- [300] Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Ferec C, Wells RD, Cooper DN. Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. *Hum Mutat* 2009;30:1189–98.
- [301] Casola C, Zekonyte U, Phillips AD, Cooper DN, Hahn MW. Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. *Genome Res* 2011.
- [302] Necsulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat* 2011;32:198–206.
- [303] Oron-Karni V, Filon D, Rund D, Oppenheim A. A novel mechanism generating short deletion/insertions following slippage is suggested by a mutation in the human α 2-globin gene. *Hum Mol Genet* 1997;6:881–5.
- [304] Chuzhanova NA, Anassis EJ, Ball EV, Krawczak M, Cooper DN. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 2003;21:28–44.
- [305] Ketterling RP, Ricke DO, Wurster MW, Sommer SS. Deletions with inversions: report of a mutation and review of the literature. *Hum Mutat* 1993;2:53–7.
- [306] Carvalho CM, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH, Friehling L, Lee S, Smith R, Del Gaudio D, Withers M, Liu P, Cheung SW, Belmont JW, Zoghbi HY, Hastings PJ, Lupski JR. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet* 2011;43:1074–81.
- [307] Chen JM, Ferec C, Cooper DN. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. *Hum Mutat* 2009;30:1435–48.
- [308] Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, Futreal PA, Campbell PJ. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011;144:27–40.
- [309] Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M, Cuppen E. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* 2011;20:1916–24.
- [310] Chen JM, Ferec C, Cooper DN. Transient hypermutability, chromothripsis and replication-based mechanisms in the generation of concurrent clustered mutations. *Mutat Res* 2011.
- [311] Shim SH, Wyandt HE, McDonald-McGinn DM, Zackai EZ, Milunsky A. Molecular cytogenetic characterization of multiple intrachromosomal rearrangements of chromosome 2q in a patient with Waardenburg's syndrome and other congenital defects. *Clin Genet* 2004;66:46–52.
- [312] Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, Bacino CA, Campos-Acevedo LD, Delgado MR, Freedenberg D, Garnica A, Grebe TA, Hernandez-Almaguer D, Immken L, Lalani SR, McLean SD, Northrup H, Scaglia F, Strathearn L, Trapani P, Kang SH, Patel A, Cheung SW, Hastings PJ, Stankiewicz P, Lupski JR, Bi W. Chromosome catastrophes involve

- replication mechanisms generating complex genomic rearrangements. *Cell* 2011;146:889–903.
- [313] Fukami M, Shima H, Suzuki E, Ogata T, Matsubara K, Kamimaki T. Catastrophic cellular events leading to complex chromosomal rearrangements in the germline. *Clin Genet* 2017;91:653–60.
- [314] van Leeuwen FW, Kros JM, Kamphorst W, van Schravendijk C, de Vos RA. Molecular misreading: the occurrence of frameshift proteins in different diseases. *Biochem Soc Trans* 2006;34:738–42.
- [315] Linton MF, Pierotti V, Young SG. Reading-frame restoration with an apolipoprotein B gene frameshift mutation. *Proc Natl Acad Sci U S A* 1992;89:11431–5.
- [316] Young M, Inaba H, Hoyer LW, Higuchi M, Kazazian Jr HH, Antonarakis SE. Partial correction of a severe molecular defect in hemophilia A, because of errors during expression of the factor VIII gene. *Am J Hum Genet* 1997;60:565–73.
- [317] Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, Jhanwar S, Winawer S, Offit K, Luce MC, Kinzler KW, Vogelstein B. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* 1997;17:79–83.
- [318] van Leeuwen FW, de Kleijn DP, van den Hurk HH, Neubauer A, Sonnemans MA, Sluijs JA, Koycu S, Ramdijlal RD, Salehi A, Martens GJ, Grosveld FG, Peter J, Burbach H, Hol EM. Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. *Science* 1998;279:242–7.
- [319] Paoloni-Giacobino A, Rossier C, Papasavvas MP, Antonarakis SE. Frequency of replication/transcription errors in (A)/(T) runs of human genes. *Hum Genet* 2001;109:40–7.
- [320] Suter CM, Martin DI, Ward RL. Germline epimutation of MLH1 in individuals with multiple cancers. *Nat Genet* 2004;36:497–501.
- [321] Hitchins M, Williams R, Cheong K, Halani N, Lin VA, Packham D, Ku S, Buckle A, Hawkins N, Burn J, Gallinger S, Goldblatt J, Kirk J, Tomlinson I, Scott R, Spigelman A, Suter C, Martin D, Suthers G, Ward R. MLH1 germline epimutations as a factor in hereditary nonpolyposis colorectal cancer. *Gastroenterology* 2005;129:1392–9.
- [322] Mariot V, Maupetit-Mehouas S, Sinding C, Kottler ML, Linglart A. A maternal epimutation of GNAS leads to Albright osteodystrophy and parathyroid hormone resistance. *J Clin Endocrinol Metab* 2008;93:661–5.
- [323] Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, Plomin R, Mill J. Allelic skewing of DNA methylation is widespread across the genome. *Am J Hum Genet* 2010;86:196–212.
- [324] Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. *Science* 2010;327:996–1000.
- [325] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008;40:897–903.
- [326] Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 2011;333:53–8.
- [327] Lualdi S, Tappino B, Di Duca M, Dardis A, Anderson CJ, Biassoni R, Thompson PW, Corsolini F, Di Rocco M, Bembi B, Regis S, Cooper DN, Filocamo M. Enigmatic in vivo iduronate-2-sulfatase (IDS) mutant transcript correction to wild-type in Hunter syndrome. *Hum Mutat* 2010;31:E1261–85.
- [328] Reich DE, Cargill M, Bolck S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. *Nature* 2001;411:199–204.
- [329] Subramanian S, Kumar S. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* 2006;7:306.
- [330] Miller MP, Parker JD, Rissing SW, Kumar S. Quantifying the intragenic distribution of human disease mutations. *Ann Hum Genet* 2003;67:567–79.
- [331] Flint J, Harding RM, Clegg JB, Boyce AJ. Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. *Hum Genet* 1993;91:91–117.
- [332] Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* 2003;4:293–340.
- [333] Zlotogora J. High frequencies of human genetic diseases: founder effect with genetic drift or selection? *Am J Med Genet* 1994;49:10–3.
- [334] Zschocke J. Phenylketonuria mutations in Europe. *Hum Mutat* 2003;21:345–56.
- [335] Mockenhaupt FP, Mandelkow J, Till H, Ehrhardt S, Eggelte TA, Bienzle U. Reduced prevalence of *Plasmodium falciparum* infection and of concomitant anaemia in pregnant women with heterozygous G6PD deficiency. *Trop Med Int Health* 2003;8:118–24.
- [336] Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, Gupta S, Warn P, Allsopp CE, Gilbert SC, Peschu N, et al. Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* 1995;376:246–9.

- [337] Aidoo M, Terlouw DJ, Kolczak MS, McElroy PD, ter Kuile FO, Kariuki S, Nahlen BL, Lal AA, Udhayakumar V. Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* 2002;359:1311–2.
- [338] Williams TN, Wambua S, Uyoga S, Macharia A, Mwacharo JK, Newton CR, Maitland K. Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal *Plasmodium falciparum* malaria on the coast of Kenya. *Blood* 2005;106:368–71.
- [339] Williams TN, Mwangi TW, Wambua S, Peto TE, Weatherall DJ, Gupta S, Recker M, Penman BS, Uyoga S, Macharia A, Mwacharo JK, Snow RW, Marsh K. Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. *Nat Genet* 2005;37:1253–7.
- [340] Witchel SF, Lee PA, Suda-Hartman M, Trucco M, Hoffman EP. Evidence for a heterozygote advantage in congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *J Clin Endocrinol Metab* 1997;82:2097–101.
- [341] Datz C, Haas T, Rinner H, Sandhofer F, Patsch W, Paulweber B. Heterozygosity for the C282Y mutation in the hemochromatosis gene is associated with increased serum iron, transferrin saturation, and hemoglobin in young women: a protective role against iron deficiency? *Clin Chem* 1998;44:2429–32.
- [342] Mead S, Stumpf MP, Whitfield J, Beck JA, Poulter M, Campbell T, Uphill JB, Goldstein D, Alpers M, Fisher EM, Collinge J. Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science* 2003;300:640–3.
- [343] Kerlin BA, Yan SB, Isermann BH, Brandt JT, Sood R, Basson BR, Joyce DE, Weiler H, Dhainaut JF. Survival advantage associated with heterozygous factor V Leiden mutation in patients with severe sepsis and in mouse endotoxemia. *Blood* 2003;102:3085–92.
- [344] Common JE, Di WL, Davies D, Kelsell DP. Further evidence for heterozygote advantage of GJB2 deafness mutations: a link with cell survival. *J Med Genet* 2004;41:573–5.
- [345] Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ. Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* 1994;266:107–9.
- [346] Schroeder SA, Gaughan DM, Swift M. Protection against bronchial asthma by CFTR delta F508 mutation: a heterozygote advantage in cystic fibrosis. *Nat Med* 1995;1:703–5.
- [347] Pier GB. Role of the cystic fibrosis transmembrane conductance regulator in innate immunity to *Pseudomonas aeruginosa* infections. *Proc Natl Acad Sci U S A* 2000;97:8822–8.
- [348] Hogenauer C, Santa Ana CA, Porter JL, Millard M, Gelfand A, Rosenblatt RL, Prestidge CB, Fordtran JS. Active intestinal chloride secretion in human carriers of cystic fibrosis mutations: an evaluation of the hypothesis that heterozygotes have subnormal active intestinal chloride secretion. *Am J Hum Genet* 2000;67:1422–7.
- [349] Motulsky AG. Jewish diseases and origins. *Nat Genet* 1995;9:99–101.
- [350] Ostrer H. A genetic profile of contemporary Jewish populations. *Nat Rev Genet* 2001;2:891–8.
- [351] Frisch A, Colombo R, Michaelovsky E, Karpati M, Goldman B, Peleg L. Origin and spread of the 1278insTATC mutation causing Tay-Sachs disease in Ashkenazi Jews: genetic drift as a robust and parsimonious hypothesis. *Hum Genet* 2004;114:366–76.
- [352] Risch N, Tang H, Katzenstein H, Ekstein J. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am J Hum Genet* 2003;72:812–22.
- [353] Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* 2003;301:643–6.
- [354] Goriely A, McVean GA, van Pelt AM, O'Rourke AW, Wall SA, de Rooij DG, Wilkie AO. Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia. *Proc Natl Acad Sci U S A* 2005;102:6051–6.
- [355] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;24:537–44.
- [356] Cai JJ, Borenstein E, Chen R, Petrov DA. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol Evol* 2009;1:131–44.
- [357] Domazet-Lošo T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol* 2008;25:2699–707.
- [358] Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001;409:853–5.
- [359] Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* 2008;105:20870–5.
- [360] Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 2004;32:3108–14.

- [361] Kondrashov FA, Ogurtsov AY, Kondrashov AS. Bio-informatical assay of human gene morbidity. *Nucleic Acids Res* 2004;32:1731–7.
- [362] Chelala C, Auffray C. Sex-linked recombination variation and distribution of disease-related genes. *Gene* 2005;346:29–39.
- [363] Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Alba MM, Ponting CP, Fechtel K. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 2004;5:R47.
- [364] Chuang JH, Li H. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol* 2004;2:E29.
- [365] Green P, Ewing B, Miller W, Thomas PJ, Green ED. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* 2003;33:514–7.
- [366] Majewski J. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet* 2003;73:688–92.
- [367] Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci U S A* 2005;102:9836–41.
- [368] Hurles M. How homologous recombination generates a mutable genome. *Hum Genomics* 2005;2:179–86.
- [369] Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 2002;32:135–42.
- [370] Chen CL, Duquenne L, Audit B, Guilbaud G, Rapailles A, Baker A, Huvet M, d'Aubenton-Carafa Y, Hyrien O, Arneodo A, Thermes C. Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol* 2011;28:2327–37.
- [371] den Dunnen JT, Antonarakis SE. Nomenclature for the description of human sequence variations. *Hum Genet* 2001;109:121–4.
- [372] den Dunnen JT, Dagleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat* 2016;37:564–9.
- [373] Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* 2008;29:6–13.
- [374] Gao L, Zhang J. Why are some human disease-associated mutations fixed in mice? *Trends Genet* 2003;19:678–81.
- [375] Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A. Epistatic interactions: how strong in disease and evolution? *Trends Genet* 2006;22:581–5.
- [376] Corona E, Dudley JT, Butte AJ. Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PLoS One* 2010;5:e12236.
- [377] Di Rienzo A, Hudson RR. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet* 2005;21:596–601.
- [378] Cotton RG, Scriver CR. Proof of “disease causing” mutation. *Hum Mutat* 1998;12:1–3.
- [379] Arai M, Inaba H, Higuchi M, Antonarakis SE, Kazazian Jr HH, Fujimaki M, Hoyer LW. Direct characterization of factor VIII in plasma: detection of a mutation altering a thrombin cleavage site (arginine-372---histidine). *Proc Natl Acad Sci U S A* 1989;86:4277–81.
- [380] Higuchi M, Wong C, Kochhan L, Olek K, Aronis S, Kasper CK, Kazazian Jr HH, Antonarakis SE. Characterization of mutations in the factor VIII gene by direct sequencing of amplified genomic DNA. *Genomics* 1990;6:65–71.
- [381] Aly AM, Higuchi M, Kasper CK, Kazazian Jr HH, Antonarakis SE, Hoyer LW. Hemophilia A due to mutations that create new N-glycosylation sites. *Proc Natl Acad Sci U S A* 1992;89:4933–7.
- [382] Owen MC, Brennan SO, Lewis JH, Carrell RW. Mutation of antitrypsin to antithrombin. alpha 1-antitrypsin Pittsburgh (358 Met leads to Arg), a fatal bleeding disorder. *N Engl J Med* 1983;309:694–8.
- [383] Vogt G, Chapgier A, Yang K, Chuzhanova N, Feinberg J, Fieschi C, Boisson-Dupuis S, Alcais A, Filipe-Santos O, Bustamante J, de Beaucoudrey L, Al-Mohsen I, Al-Hajjar S, Al-Ghonaïum A, Adimi P, Mirsaeidi M, Khalilzadeh S, Rosenzweig S, de la Calle Martin O, Bauer TR, Puck JM, Ochs HD, Furthner D, Engelhorn C, Belohradsky B, Mansouri D, Holland SM, Schreiber RD, Abel L, Cooper DN, Soudais C, Casanova JL. Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. *Nat Genet* 2005;37:692–700.
- [384] Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, Reitsma PH. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 1994;369:64–7.
- [385] Challis BG, Pritchard LE, Creemers JW, Delpianque J, Keogh JM, Luan J, Wareham NJ, Yeo GS, Bhattacharyya S, Froguel P, White A, Farooqi IS, O'Rahilly S. A missense mutation disrupting a dibasic prohormone processing site in pro-opiomelanocortin (POMC) increases susceptibility to early-onset obesity through a novel molecular mechanism. *Hum Mol Genet* 2002;11:1997–2004.

- [386] Byers P. Disorders of collagen biosynthesis and structure. In: Scriver CR, Beaudet AL, Valle D, et al., editors. *The metabolic and molecular bases of inherited disease*. New York: McGraw-Hill; 2001. p. 5241–86.
- [387] Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P, Mooney SD. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum Mutat* 2010;31:335–46.
- [388] Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. *Trends Genet* 2003;19:505–13.
- [389] Gong S, Blundell TL. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PLoS One* 2010;5:e9186.
- [390] Schuster-Bockler B, Bateman A. Protein interactions in human genetic diseases. *Genome Biol* 2008;9:R9.
- [391] Xin F, Myers S, Li YF, Cooper DN, Mooney SD, Radivojac P. Structure-based kernels for the prediction of catalytic residues and their involvement in human inherited disease. *Bioinformatics* 2010;26:1975–82.
- [392] Li S, Iakoucheva LM, Mooney SD, Radivojac P. Loss of post-translational modification sites in disease. *Pac Symp Biocomput* 2010:337–47.
- [393] Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Protein disorder in the human diseasome: unfoldomics of human genetic diseases. *BMC Genomics* 2009;10(Suppl. 1):S12.
- [394] Wang Z, Moulton J. SNPs, protein structure, and disease. *Hum Mutat* 2001;17:263–70.
- [395] Yue P, Li Z, Moulton J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005;353:459–73.
- [396] Laskowski RA, Thornton JM. Understanding the molecular machinery of genetics through 3D structures. *Nat Rev Genet* 2008;9:141–51.
- [397] Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum Mutat* 2009;30:703–14.
- [398] Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;12:628–40.
- [399] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;4:45–61.
- [400] Sunyaev S, Ramensky V, Koch I, Lathe 3rd W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet* 2001;10:591–7.
- [401] Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Hum Mutat* 2002;20:98–109.
- [402] Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 2003;4:R72.
- [403] Youssoufian H, Antonarakis SE, Aronis S, Tsiftis G, Phillips D.G., Kazazian Jr. H.H.. Characterization of five partial deletions of the factor VIII gene. *Proc Natl Acad Sci U S A* 1987;84:3772–6.
- [404] Cecchini KR, Raja Banerjee A, Kim TH. Towards a genome-wide reconstruction of cis-regulatory networks in the human genome. *Semin Cell Dev Biol* 2009;20:842–8.
- [405] Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* 2009;106:7507–12.
- [406] Wen Y, Liu Y, Xu Y, Zhao Y, Hua R, Wang K, Sun M, Li Y, Yang S, Zhang XJ, Kruse R, Cichon S, Betz RC, Nothen MM, van Steensel MA, van Geel M, Steijlen PM, Hohl D, Huber M, Dunnill GS, Kennedy C, Messenger A, Munro CS, Terrinoni A, Hovnanian A, Bodemer C, de Prost Y, Paller AS, Irvine AD, Sinclair R, Green J, Shang D, Liu Q, Luo Y, Jiang L, Chen HD, Lo WH, McLean WH, He CD, Zhang X. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat Genet* 2009;41:228–33.
- [407] Antonarakis SE, Irkin SH, Cheng TC, Scott AF, Sexton JP, Trusko SP, Charache S, Kazazian Jr HH. Beta-thalassemia in American Blacks: novel mutations in the “TATA” box and an acceptor splice site. *Proc Natl Acad Sci U S A* 1984;81:1154–8.
- [408] Orkin SH, Antonarakis SE, Kazazian Jr HH. Base substitution at position -88 in a beta-thalassemic globin gene. Further evidence for the role of distal promoter element ACACCC. *J Biol Chem* 1984;259:8679–81.
- [409] Perkins AC, Sharpe AH, Orkin SH. Lethal beta-thalassemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature* 1995;375:318–22.
- [410] Collins FS, Stoeckert Jr CJ, Serjeant GR, Forget BG, Weissman SM. G gamma beta+ hereditary persistence of fetal hemoglobin: cosmid cloning and identification of a specific mutation 5' to the G gamma gene. *Proc Natl Acad Sci U S A* 1984;81:4894–8.
- [411] Crossley M, Brownlee GG. Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. *Nature* 1990;345:444–6.
- [412] Koivisto UM, Palvimo JJ, Janne OA, Kontula K. A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a

- cause of heterozygous familial hypercholesterolemia. *Proc Natl Acad Sci U S A* 1994;91:10526–30.
- [413] Berg LP, Scopes DA, Alhaq A, Kakkar VV, Cooper DN. Disruption of a binding site for hepatocyte nuclear factor 1 in the protein C gene promoter is associated with hereditary thrombophilia. *Hum Mol Genet* 1994;3:2147–52.
- [414] Yang WS, Nevin DN, Peng R, Brunzell JD, Deeb SS. A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity. *Proc Natl Acad Sci U S A* 1995;92:4462–6.
- [415] Weatherall DJ, Clegg JB, Higgs DR, Wood WG. The hemoglobinopathies. In: Scriver CR, Beaudet AL, Valle D, et al., editors. *The metabolic and molecular bases of inherited disease*. New York: McGraw-Hill; 2001. p. 4571–636.
- [416] De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 2006;312:1215–7.
- [417] Semenza GL. Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Hum Mutat* 1994;3:180–99.
- [418] Inoue I, Nakajima T, Williams CS, Quackenbush J, Puryear R, Powers M, Cheng T, Ludwig EH, Sharma AM, Hata A, Jeunemaitre X, Lalouel JM. A nucleotide substitution in the promoter of human angiotensinogen is associated with essential hypertension and affects basal transcription in vitro. *J Clin Invest* 1997;99:1786–97.
- [419] Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* 1992;90:41–54.
- [420] Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. *Genes Dev* 2003;17:419–37.
- [421] Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat* 2006;28:150–8.
- [422] Treisman R, Orkin SH, Maniatis T. Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. *Nature* 1983;302:591–6.
- [423] Sharp PA. Splicing of messenger RNA precursors. *Science* 1987;235:766–71.
- [424] Rosenthal A, Joutet M, Kenwrick S. Aberrant splicing of neural cell adhesion molecule L1 mRNA in a family with X-linked hydrocephalus. *Nat Genet* 1992;2:107–12.
- [425] De Klein A, Riegman PH, Bijlsma EK, Helderdoorn A, Muijtjens M, den Bakker MA, Avezaat CJ, Zwarthoff EC. A G→A transition creates a branch point sequence and activation of a cryptic exon, resulting in the hereditary disorder neurofibromatosis 2. *Hum Mol Genet* 1998;7:393–8.
- [426] Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 2000;28:4364–75.
- [427] Shaw MA, Brunetti-Pierri N, Kadasi L, Kovacova V, Van Maldergem L, De Brasi D, Salerno M, Gecz J. Identification of three novel SEDL mutations, including mutation in the rare, non-canonical splice site of exon 4. *Clin Genet* 2003;64:235–42.
- [428] Bradley KJ, Cavaco BM, Bowl MR, Harding B, Young A, Thakker RV. Utilisation of a cryptic non-canonical donor splice site of the gene encoding PARAFIBROMIN is associated with familial isolated primary hyperparathyroidism. *J Med Genet* 2005;42:e51.
- [429] Cogan JD, Prince MA, Lekhakula S, Bunday S, Futrakul A, McCarthy EM, Phillips 3rd JA. A novel mechanism of aberrant pre-mRNA splicing in humans. *Hum Mol Genet* 1997;6:909–12.
- [430] Santisteban I, Arredondo-Vega FX, Kelly S, Loubser M, Meydan N, Roifman C, Howell PL, Bowen T, Weinberg KI, Schroeder ML, et al. Three new adenosine deaminase mutations that define a splicing enhancer and cause severe and partial phenotypes: implications for evolution of a CpG hotspot and expression of a transduced ADA cDNA. *Hum Mol Genet* 1995;4:2081–7.
- [431] Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A, Hackett J, Adamson J, Lincoln S, Dickson D, Davies P, Petersen RC, Stevens M, de Graaff E, Wauters E, van Baren J, Hillebrand M, Joosse M, Kwon JM, Nowotny P, Che LK, Norton J, Morris JC, Reed LA, Trojanowski J, Basun H, Lannfelt L, Neystat M, Fahn S, Dark F, Tannenberg T, Dodd PR, Hayward N, Kwok JB, Schofield PR, Andreadis A, Snowden J, Craufurd D, Neary D, Owen F, Oostra BA, Hardy J, Goate A, van Swieten J, Mann D, Lynch T, Heutink P. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* 1998;393:702–5.
- [432] Pagani F, Buratti E, Stuaní C, Bendix R, Dork T, Baralle FE. A new type of mutation causes a splicing defect in ATM. *Nat Genet* 2002;30:426–9.
- [433] von Ahsen N, Oellerich M. The intronic prothrombin 19911A>G polymorphism influences splicing efficiency and modulates effects of the 20210G>A polymorphism on mRNA amount and expression in a stable reporter gene assay system. *Blood* 2004;103:586–93.
- [434] Dietz HC, Valle D, Francomano CA, Kendzior Jr RJ, Pyeritz RE, Cutting GR. The skipping of constitutive exons in vivo induced by nonsense mutations. *Science* 1993;259:680–3.

- [435] Chao HK, Hsiao KJ, Su TS. A silent mutation induces exon skipping in the phenylalanine hydroxylase gene in phenylketonuria. *Hum Genet* 2001;108:14–9.
- [436] Liu HX, Cartegni L, Zhang MQ, Krainer AR. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* 2001;27:55–8.
- [437] Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 2000;25:106–10.
- [438] Pagani F, Raponi M, Baralle FE. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A* 2005;102:6368–72.
- [439] Pagani F, Baralle FE. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet* 2004;5:389–96.
- [440] Gorlov IP, Gorlova OY, Frazier ML, Amos CI. Missense mutations in hMLH1 and hMSH2 are associated with exonic splicing enhancers. *Am J Hum Genet* 2003;73:1157–61.
- [441] Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A* 2011;108:11093–8.
- [442] Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res* 2011;21:1563–71.
- [443] Mitchell GA, Labuda D, Fontaine G, Saudubray JM, Bonnefont JP, Lyonnet S, Brody LC, Steel G, Obie C, Valle D. Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. *Proc Natl Acad Sci U S A* 1991;88:815–9.
- [444] Knebelmann B, Forestier L, Drouot L, Quinones S, Chuet C, Benessy F, Saus J, Antignac C. Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. *Hum Mol Genet* 1995;4:675–9.
- [445] Coutinho G, Xie J, Du L, Brusco A, Krainer AR, Gatti RA. Functional significance of a deep intronic mutation in the ATM gene and evidence for an alternative exon 28a. *Hum Mutat* 2005;25:118–24.
- [446] Harland M, Mistry S, Bishop DT, Bishop JA. A deep intronic mutation in CDKN2A is associated with disease in a subset of melanoma pedigrees. *Hum Mol Genet* 2001;10:2679–86.
- [447] Tuffery-Giraud S, Saquet C, Chambert S, Claustres M. Pseudoexon activation in the DMD gene as a novel mechanism for Becker muscular dystrophy. *Hum Mutat* 2003;21:608–14.
- [448] Pros E, Gomez C, Martin T, Fabregas P, Serra E, Lázaro C. Nature and mRNA effect of 282 different NF1 point mutations: focus on splicing alterations. *Hum Mutat* 2008;29:E173–93.
- [449] Dhir A, Buratti E. Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *FEBS J* 2010;277:841–55.
- [450] Choi JW, Park CS, Hwang M, Nam HY, Chang HS, Park SG, Han BG, Kimm K, Kim HL, Oh B, Kim Y. A common intronic variant of CXCR3 is functionally associated with gene expression levels and the polymorphic immune cell responses to stimuli. *J Allergy Clin Immunol* 2008;122:1119–26. e1117.
- [451] Fraser HB, Xie X. Common polymorphic transcript variation in human disease. *Genome Res* 2009;19:567–75.
- [452] Susa S, Daimon M, Sakabe J, Sato H, Oizumi T, Karasawa S, Wada K, Jimbu Y, Kameda W, Emi M, Muramatsu M, Kato T. A functional polymorphism of the TNF-alpha gene that is associated with type 2 DM. *Biochem Biophys Res Commun* 2008;369:943–7.
- [453] Vaz-Drágo R, Custódio N, Carmo-Fonseca M, PMID:28497172, DOI:10.1007/s00439-017-1809-4.
- [454] Chen JM, Ferenc C, Cooper DN. A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3' UTR variants. *Hum Genet* 2006b;120:301–33.
- [455] Orkin SH, Cheng TC, Antonarakis SE, Kazazian Jr HH. Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. *EMBO J* 1985;4:453–6.
- [456] Cai SP, Eng B, Francombe WH, Olivieri NF, Kendall AG, Wayne JS, Chui DH. Two novel beta-thalassemia mutations in the 5' and 3' noncoding regions of the beta-globin gene. *Blood* 1992;79:1342–6.
- [457] Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, Kulozik AE. Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nat Genet* 2001;28:389–92.
- [458] Poort SR, Rosendaal FR, Reitsma PH, Bertina RM. A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. *Blood* 1996;88:3698–703.
- [459] Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LS, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW. Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* 2005;310:317–20.

- [460] Sethupathy P, Borel C, Gagnebin M, Grant G.R, Deutsch S., et al. *Am J Hum Genet* 2006;81:405–13.
- [461] Bandiera S, Hatem E, Lyonnet S, Henrion-Caude A. microRNAs in diseases: from candidate to modifier genes. *Clin Genet* 2010;77:306–13.
- [462] Martin MM, Buckenberger JA, Jiang J, Malana GE, Nuovo GJ, Chotani M, Feldman DS, Schmittgen TD, Elton TS. The human angiotensin II type 1 receptor +1166 A/C polymorphism attenuates microRNA-155 binding. *J Biol Chem* 2007;282:24262–9.
- [463] Rademakers R, Eriksen JL, Baker M, Robinson T, Ahmed Z, Lincoln SJ, Finch N, Rutherford NJ, Crook RJ, Josephs KA, Boeve BF, Knopman DS, Petersen RC, Parisi JE, Caselli RJ, Wszolek ZK, Uitti RJ, Feldman H, Hutton ML, Mackenzie IR, Graff-Radford NR, Dickson DW. Common variation in the miR-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. *Hum Mol Genet* 2008;17:3631–42.
- [464] Sethupathy P, Borel C, Gagnebin M, Grant GR, Deutsch S, Elton TS, Hatzigeorgiou AG, Antonarakis SE. Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3' untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. *Am J Hum Genet* 2007;81:405–13.
- [465] Simon D, Laloo B, Barillot M, Barnette T, Blanchard C, Rooryck C, Marche M, Burgelin I, Coupry I, Chassaing N, Gilbert-Dussardier B, Lacombe D, Grosset C, Arveiler B. A mutation in the 3'-UTR of the HDAC6 gene abolishing the post-transcriptional regulation mediated by hsa-miR-433 is linked to a new form of dominant X-linked chondrodysplasia. *Hum Mol Genet* 2010;19:2015–27.
- [466] Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011;12:861–74.
- [467] Dong XY, Rodriguez C, Guo P, Sun X, Talbot JT, Zhou W, Petros J, Li Q, Vessella RL, Kibel AS, Stevens VL, Calle EE, Dong JT. SnoRNA U50 is a candidate tumor-suppressor gene at 6q14.3 with a mutation associated with clinically significant prostate cancer. *Hum Mol Genet* 2008;17:1031–42.
- [468] Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, Garnica A, Cheung SW, Beaudet AL. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet* 2008;40:719–21.
- [469] Li H, Xie H, Liu W, Hu R, Huang B, Tan YF, Xu K, Sheng ZF, Zhou HD, Wu XP, Luo XH. A novel microRNA targeting HDAC5 regulates osteoblast differentiation in mice and contributes to primary osteoporosis in humans. *J Clin Invest* 2009;119:3666–77.
- [470] Mencia A, Modamio-Hoybjor S, Redshaw N, Morin M, Mayo-Merino F, Olavarrieta L, Aguirre LA, del Castillo I, Steel KP, Dalmay T, Moreno F, Moreno-Pelayo MA. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat Genet* 2009;41:609–13.
- [471] Saus E, Soria V, Escaramis G, Vivarelli F, Crespo JM, Kagerbauer B, Menchon JM, Urretavizcaya M, Gratacos M, Estivill X. Genetic variants and abnormal processing of pre-miR-182, a circadian clock modulator, in major depression patients with late insomnia. *Hum Mol Genet* 2010;19:4017–25.
- [472] Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, Sommer SS, Rossi JJ. SNPs in human miRNA genes affect biogenesis and function. *RNA* 2009;15:1640–51.
- [473] Ridanpaa M, van Eenennaam H, Pelin K, Chadwick R, Johnson C, Yuan B, vanVenrooij W, Pruijn G, Salmela R, Rockas S, Makitie O, Kaitila I, de la Chapelle A. Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell* 2001;104:195–203.
- [474] Bastepe M, Frohlich LF, Linglart A, Abu-Zahra HS, Tojo K, Ward LM, Juppner H. Deletion of the NESP55 differentially methylated region causes loss of maternal GNAS imprints and pseudohypoparathyroidism type 1b. *Nat Genet* 2005;37:25–7.
- [475] Jarinova O, Stewart AF, Roberts R, Wells G, Lau P, Naing T, Buerki C, McLean BW, Cook RC, Parker JS, McPherson R. Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. *Arterioscler Thromb Vasc Biol* 2009;29:1671–7.
- [476] Brakenhoff RH, Henskens HA, van Rossum MW, Lubsen NH, Schoenmakers JG. Activation of the gamma E-crystallin pseudogene in the human hereditary Coppock-like cataract. *Hum Mol Genet* 1994;3:279–83.
- [477] Lewis MA, Quint E, Glazier AM, Fuchs H, De Angelis MH, Langford C, van Dongen S, Abreu-Goodger C, Piipari M, Redshaw N, Dalmay T, Moreno-Pelayo MA, Enright AJ, Steel KP. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat Genet* 2009;41:614–8.
- [478] Hughes AE, Bradley DT, Campbell M, Lechner J, Dash DP, Simpson DA, Willoughby CE. Mutation altering the miR-184 seed region causes familial keratoconus with cataract. *Am J Hum Genet* 2011;89:628–33.
- [479] de Pontual L, Yao E, Callier P, Faivre L, Drouin V, Cariou S, Van Haeringen A, Genevieve D, Goldenberg A, Oufadem M, Manouvrier S, Munnich A, Vidigal JA, Vekemans M, Lyonnet S, Henrion-Caude A, Ventura A, Amiel J. Germline deletion of the miR-17 approximately 92 cluster causes skeletal and growth defects in humans. *Nat Genet* 2011;43:1026–30.
- [480] Abdel-Salam GM, Abdel-Hamid MS, Issa M, Magdy A, El-Kotoury A, Amr K. Expanding the phenotypic and mutational spectrum in microcephalic osteodys-

- plastic primordial dwarfism type I. *Am J Med Genet A* 2012;158A:1455–61.
- [481] He H, Liyanarachchi S, Akagi K, Nagy R, Li J, Dietrich RC, Li W, Sebastian N, Wen B, Xin B, Singh J, Yan P, Alder H, Haan E, Wiczorek D, Albrecht B, Puffenberger E, Wang H, Westman JA, Padgett RA, Symer DE, de la Chapelle A. Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science* 2011;332:238–40.
- [482] Makrythanasis P, Antonarakis SE. Pathogenic variants in non-protein-coding sequences. *Clin Genet* 2013;84:422–8.
- [483] Collins LJ, Penny D. The RNA infrastructure: dark matter of the eukaryotic cell? *Trends Genet* 2009;25:120–8.
- [484] Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. Widely distributed non-coding purifying selection in the human genome. *Proc Natl Acad Sci U S A* 2007;104:12410–5.
- [485] Kryukov GV, Schmidt S, Sunyaev S. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet* 2005;14:2221–9.
- [486] Chen CT, Wang JC, Cohen BA. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* 2007;80:692–704.
- [487] Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. *Nature* 2009;461:199–205.
- [488] Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 2009;19:212–9.
- [489] Glinskii AB, Ma J, Ma S, Grant D, Lim CU, Sell S, Glinsky GV. Identification of intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell cycle progression/differentiation pathways in multiple common human disorders. *Cell Cycle* 2009;8:3925–42.
- [490] Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;106:9362–7.
- [491] Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigo R, Gingeras TR, Antonarakis SE, Reymond A. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res* 2007;17:746–59.
- [492] Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;8:e1000294.
- [493] Attanasio C, Reymond A, Humbert R, Lyle R, Kuehn MS, Neph S, Sabo PJ, Goldy J, Weaver M, Haydock A, Lee K, Dorschner M, Dermitzakis ET, Antonarakis SE, Stamatoyannopoulos JA. Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. *Genome Biol* 2008;9:R168.
- [494] Quemener S, Chen JM, Chuzhanova N, Benech C, Casals T, Macek Jr M, Bienvenu T, McDevitt T, Farrell PM, Loumi O, Messaoud T, Cuppens H, Cutting GR, Stenson PD, Giteau K, Audrezet MP, Cooper DN, Ferec C. Complete ascertainment of intragenic copy number mutations (CNMs) in the CFTR gene and its implications for CNM formation at other autosomal loci. *Hum Mutat* 2010;31:421–8.
- [495] Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet* 2015;24:R102–10.
- [496] Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum Mol Genet* 2016;25:R157–65.
- [497] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485:376–80.
- [498] Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 2015;161:1012–25.
- [499] Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schopflin R, Kraft K, Kempfer R, Jerkovic I, Chan WL, Spielmann M, Timmermann B, Wittler L, Kurth I, Cambiaso P, Zuffardi O, Houge G, Lambie L, Brancati F, Pombo A, Vingron M, Spitz F, Mundlos S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 2016;538:265–9.
- [500] Wong C, Dowling CE, Saiki RK, Higuchi RG, Erlich HA, Kazazian Jr HH. Characterization of beta-thalassemia mutations using direct genomic sequencing of amplified single copy DNA. *Nature* 1987;330:384–6.
- [501] Kozak M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res* 1984;12:857–72.
- [502] Mehdi H, Manzi S, Desai P, Chen Q, Nestlerode C, Bontempo F, Strom SC, Zarnegar R, Kamboh MI. A functional polymorphism at the transcriptional initiation site in beta2-glycoprotein I (apolipoprotein H) associated with reduced gene expression and lower plasma levels of beta2-glycoprotein I. *Eur J Biochem* 2003;270:230–8.

- [503] Cazzola M, Skoda RC. Translational pathophysiology: a novel molecular mechanism of human disease. *Blood* 2000;95:3280–8.
- [504] Girelli D, Corrocher R, Bisceglia L, Olivieri O, De Franceschi L, Zelante L, Gasparini P. Molecular basis for the recently described hereditary hyperferritinemia-cataract syndrome: a mutation in the iron-responsive element of ferritin L-subunit gene (the “Verona mutation”). *Blood* 1995;86:4050–3.
- [505] Athanassiadou A, Papachatzopoulou A, Zoumbos N, Maniatis GM, Gibbs R. A novel beta-thalassaemia mutation in the 5′ untranslated region of the beta-globin gene. *Br J Haematol* 1994;88:307–10.
- [506] Ho PJ, Rochette J, Fisher CA, Wonke B, Jarvis MK, Yardumian A, Thein SL. Moderate reduction of beta-globin gene transcript by a novel mutation in the 5′ untranslated region: a study of its interaction with other genotypes in two families. *Blood* 1996;87:1170–8.
- [507] Sgourou A, Routledge S, Antoniou M, Papachatzopoulou A, Psiouri L, Athanassiadou A. Thalassaemia mutations within the 5′UTR of the human beta-globin gene disrupt transcription. *Br J Haematol* 2004;124:828–35.
- [508] Conne B, Stutz A, Vassalli JD. The 3′ untranslated region of messenger RNA: a molecular ‘hotspot’ for pathology? *Nat Med* 2000;6:637–41.
- [509] Moi P, Loudianos G, Lavinha J, Murru S, Cossu P, Casu R, Oggiano L, Longinotti M, Cao A, Pirastu M. Delta-thalassaemia due to a mutation in an erythroid-specific binding protein sequence 3′ to the delta-globin gene. *Blood* 1992;79:512–6.
- [510] Chen JM, Ferec C, Cooper DN. A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes I: general principles and overview. *Hum Genet* 2006a;120:1–21.
- [511] Pirastu M, Saglio G, Chang JC, Cao A, Kan YW. Initiation codon mutation as a cause of alpha thalassemia. *J Biol Chem* 1984;259:12315–7.
- [512] Kozak M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 1991;266:19867–70.
- [513] Jacobson EM, Concepcion E, Oashi T, Tomer Y. A Graves’ disease-associated Kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: a case for translational pathophysiology. *Endocrinology* 2005;146:2684–91.
- [514] Kozak M. Emerging links between initiation of translation and human diseases. *Mamm Genome* 2002;13:401–10.
- [515] Wolf A, Caliebe A, Thomas NS, Ball EV, Mort M, Stenson PD, Krawczak M, Cooper DN. Single base-pair substitutions at the translation initiation sites of human genes as a cause of inherited disease. *Hum Mutat* 2011;32:1137–43.
- [516] Clegg JB, Weatherall DJ, Milner PF. Haemoglobin Constant Spring—a chain termination mutant? *Nature* 1971;234:337–40.
- [517] Hamby SE, Thomas NS, Cooper DN, Chuzhanova N. A meta-analysis of single base-pair substitutions in translational termination codons (‘nonstop’ mutations) that cause human inherited disease. *Hum Genomics* 2011;5:241–64.
- [518] Zia A, Moses AM. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC Bioinformatics* 2011;12:299.
- [519] Mort M, Ivanov D, Cooper DN, Chuzhanova NA. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat* 2008;29:1037–47.
- [520] Benz EJ, Forget BG, Hillman DG, Cohen-Solal M, Pritchard J, Cavallese C, Prenskey W, Housman D. Variability in the amount of beta-globin mRNA in beta0 thalassemia. *Cell* 1978;14:299–312.
- [521] Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* 2004;5:89–99.
- [522] Inacio A, Silva AL, Pinto J, Ji X, Morgado A, Almeida F, Faustino P, Lavinha J, Liebhauer SA, Romao L. Nonsense mutations in close proximity to the initiation codon fail to trigger full nonsense-mediated mRNA decay. *J Biol Chem* 2004;279:32170–80.
- [523] Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, Ferreira PG, Smith KS, Zhang R, Zhao F, Banks E, Poplin R, Ruderfer DM, Purcell SM, Tukiainen T, Minikel EV, Stenson PD, Cooper DN, Huang KH, Sullivan TJ, Nedzel J, Consortium GT, Geuvadis C, Bustamante CD, Li JB, Daly MJ, Guigo R, Donnelly P, Ardlie K, Sammeth M, Dermitzakis ET, McCarthy MI, Montgomery SB, Lappalainen T, MacArthur DG. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 2015;348:666–9.
- [524] Inoue K, Khajavi M, Ohyama T, Hirabayashi S, Wilson J, Reggin JD, Mancias P, Butler IJ, Wilkinson MF, Wegner M, Lupski JR. Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nat Genet* 2004;36:361–9.
- [525] Ozisik G, Mantovani G, Achermann JC, Persani L, Spada A, Weiss J, Beck-Peccoz P, Jameson JL. An alternate translation initiation site circumvents an amino-terminal DAX1 nonsense mutation leading to a mild form of X-linked adrenal hypoplasia congenita. *J Clin Endocrinol Metab* 2003;88:417–23.
- [526] Frischmeyer PA, Dietz HC. Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet* 1999;8:1893–900.
- [527] Pacho F, Zambruno G, Calabresi V, Kiritsi D, Schneider H. Efficiency of translation termination in humans

- is highly dependent upon nucleotides in the neighbourhood of a (premature) termination codon. *J Med Genet* 2011;48:640–4.
- [528] Alber T. Mutational effects on protein stability. *Annu Rev Biochem* 1989;58:765–98.
- [529] Pakula AA, Sauer RT. Genetic analysis of protein stability and function. *Annu Rev Genet* 1989;23:289–310.
- [530] Wacey AI, Cooper DN, Liney D, Hovig E, Krawczak M. Disentangling the perturbational effects of amino acid substitutions in the DNA-binding domain of p53. *Hum Genet* 1999;104:15–22.
- [531] Bross P, Corydon TJ, Andresen BS, Jorgensen MM, Bolund L, Gregersen N. Protein misfolding and degradation in genetic diseases. *Hum Mutat* 1999;14:186–98.
- [532] Gregersen N, Bross P, Jorgensen MM, Corydon TJ, Andresen BS. Defective folding and rapid degradation of mutant proteins is a common disease mechanism in genetic disorders. *J Inherit Metab Dis* 2000;23:441–7.
- [533] Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* 1987;51:975–85.
- [534] Stamatoyannopoulos G. Human hemoglobin switching. *Science* 1991;252:383.
- [535] Vyas P, Vickers MA, Simmons DL, Ayyub H, Craddock CF, Higgs DR. Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin. *Cell* 1992;69:781–93.
- [536] Driscoll MC, Dobkin CS, Alter BP. Gamma delta beta-thalassemia due to a de novo mutation deleting the 5' beta-globin gene activation-region hypersensitive sites. *Proc Natl Acad Sci U S A* 1989;86:7470–4.
- [537] Liebhaber SA, Griese EU, Weiss I, Cash FE, Ayyub H, Higgs DR, Horst J. Inactivation of human alpha-globin gene expression by a de novo deletion located upstream of the alpha-globin gene cluster. *Proc Natl Acad Sci U S A* 1990;87:9431–5.
- [538] Lecointre R, Lima S, Varlet MN, Combe C. Immunoglobulin treatment for neonatal hemochromatosis: a case report in a context of immunoglobulin delivery quotas. *Ann Pharm Fr* 2009;67:304–9.
- [539] D'Haene B, Attanasio C, Beysen D, Dostie J, Lemire E, Bouchard P, Field M, Jones K, Lorenz B, Menten B, Buysse K, Pattyn F, Friedli M, Ucla C, Rossier C, Wyss C, Speleman F, De Paepe A, Dekker J, Antonarakis SE, De Baere E. Disease-causing 7.4 kb cis-regulatory deletion disrupting conserved non-coding sequences and their interaction with the FOXL2 promoter: implications for mutation screening. *PLoS Genet* 2009;5:e1000522.
- [540] Chen J, Wildhardt G, Zhong Z, Roth R, Weiss B, Steinberger D, Decker J, Blum WF, Rappold G. Enhancer deletions of the SHOX gene as a frequent cause of short stature: the essential role of a 250 kb downstream regulatory domain. *J Med Genet* 2009;46:834–9.
- [541] Gordon CT, Tan TY, Benko S, Fitzpatrick D, Lyonnet S, Farlie PG. Long-range regulation at the SOX9 locus in development and disease. *J Med Genet* 2009;46:649–56.
- [542] Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, McBride D, Golzio C, Fisher M, Perry P, Abadie V, Ayuso C, Holder-Espinasse M, Kilpatrick N, Lees MM, Picard A, Temple IK, Thomas P, Vazquez MP, Vekemans M, Roest Crollius H, Hastie ND, Munnich A, Etchevers HC, Pelet A, Farlie PG, Fitzpatrick DR, Lyonnet S. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat Genet* 2009;41:359–64.
- [543] Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, Lie RT, Wilcox AJ, Fitzpatrick DR, Green ED, Mossey PA, Little J, Steegers-Theunissen RP, Pennacchio LA, Schutte BC, Murray JC. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 2008;40:1341–7.
- [544] Haiman CA, Le Marchand L, Yamamoto J, Stram DO, Sheng X, Kolonel LN, Wu AH, Reich D, Henderson BE. A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* 2007;39:954–6.
- [545] Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Taberner J, Baselga J, Regan MM, Manak JR, Shivdasani R, Coetzee GA, Freedman ML. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 2009;41:882–4.
- [546] Tuupainen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymäki I, Mecklin JP, Jarvinen H, Ristimäki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 2009;41:885–90.
- [547] Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* 2010;30:1411–20.
- [548] Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 2002;30:233–7.

- [549] Lewinsky RH, Jensen TG, Moller J, Stensballe A, Olsen J, Troelsen JT. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Hum Mol Genet* 2005;14:3945–53.
- [550] Olds LC, Sibley E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum Mol Genet* 2003;12:2333–40.
- [551] Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008;4:e1000214.
- [552] Venturin M, Moncini S, Villa V, Russo S, Bonati MT, Larizza L, Riva P. Mutations and novel polymorphisms in coding regions and UTRs of CDK5R1 and OMG genes in patients with non-syndromic mental retardation. *Neurogenetics* 2006;7:59–66.
- [553] Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proenca C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, Charpentier G, Dina C, Durand E, Elliott P, Hadjadj S, Jarvelin MR, Laitinen J, Lauritzen T, Marre M, Mazur A, Meyre D, Montpetit A, Pisinger C, Posner B, Poulsen P, Pouta A, Prentki M, Ribel-Madsen R, Ruokonen A, Sandbaek A, Serre D, Tichet J, Vaxillaire M, Wojtaszewski JF, Vaag A, Hansen T, Polychronakos C, Pedersen O, Froguel P, Sladek R. Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat Genet* 2009;41:1110–5.
- [554] Warren ST, Nelson DL. Advances in molecular analysis of fragile X syndrome. *JAMA* 1994;271:536–42.
- [555] Borrell-Pages M, Zala D, Humbert S, Saudou F. Huntington's disease: from huntingtin function and dysfunction to therapeutic strategies. *Cell Mol Life Sci* 2006;63:2642–60.
- [556] Davis BM, McCurrach ME, Taneja KL, Singer RH, Housman DE. Expansion of a CUG trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. *Proc Natl Acad Sci U S A* 1997;94:7388–93.
- [557] Day JW, Ranum LP. Genetics and molecular pathogenesis of the myotonic dystrophies. *Curr Neurol Neurosci Rep* 2005;5:55–9.
- [558] McVey JH, Michaelides K, Hansen LP, Ferguson-Smith M, Tilghman S, Krumlauf R, Tuddenham EG. A G→A substitution in an HNF I binding site in the human alpha-fetoprotein gene is associated with hereditary persistence of alpha-fetoprotein (HPAFP). *Hum Mol Genet* 1993;2:379–84.
- [559] Wijmenga C, Hewitt JE, Sandkuijl LA, Clark LN, Wright TJ, Dauwerse HG, Gruter AM, Hofker MH, Moerer P, Williamson R, et al. Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat Genet* 1992;2:26–30.
- [560] Deidda G, Cacurri S, Grisanti P, Vigneti E, Piazzi N, Felicetti L. Physical mapping evidence for a duplicated region on chromosome 10qter showing high homology with the facioscapulohumeral muscular dystrophy locus on chromosome 4qter. *Eur J Hum Genet* 1995;3:155–67.
- [561] Lemmers RJ, van der Vliet PJ, Klooster R, Sacconi S, Camano P, Dauwerse JG, Snider L, Straasheijm KR, van Ommen GJ, Padberg GW, Miller DG, Tapscott SJ, Tawil R, Frants RR, van der Maarel SM. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* 2010;329:1650–3.
- [562] Richards M, Coppee F, Thomas N, Belayew A, Upadhyaya M. Facioscapulohumeral muscular dystrophy (FSHD): an enigma unravelled? *Hum Genet* 2011.
- [563] Kleinjan DJ, van Heyningen V. Position effect in human genetic disease. *Hum Mol Genet* 1998;7:1611–8.
- [564] Fantes J, Redeker B, Breen M, Boyle S, Brown J, Fletcher J, Jones S, Bickmore W, Fukushima Y, Mannens M, Danes S, van Heyningen V, Hanson I. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Hum Mol Genet* 1995;4:415–22.
- [565] Pfeifer D, Kist R, Dewar K, Devon K, Lander ES, Birren B, Korniszewski L, Back E, Scherer G. Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region. *Am J Hum Genet* 1999;65:111–24.
- [566] Velagaleti GV, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, Withers M, Lupski JR, Stankiewicz P. Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. *Am J Hum Genet* 2005;76:652–62.
- [567] de Kok YJ, Vossenaar ER, Cremers CW, Dahl N, Laporte J, Hu LJ, Lacombe D, Fischel-Ghodsian N, Friedman RA, Parnes LS, Thorpe P, Bitner-Glindzicz M, Pander HJ, Heilbronner H, Graveline J, den Dunnen JT, Brunner HG, Ropers HH, Cremers FP. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. *Hum Mol Genet* 1996;5:1229–35.
- [568] Spitz F, Montavon T, Monso-Hinard C, Morris M, Ventruto ML, Antonarakis S, Ventruto V, Duboule D. A t(2;8) balanced translocation with breakpoints near the human HOXD complex causes mesomelic dysplasia and vertebral defects. *Genomics* 2002;79:493–8.

- [569] Beysen D, Raes J, Leroy BP, Lucassen A, Yates JR, Clayton-Smith J, Ilyina H, Brooks SS, Christin-Maitre S, Fellous M, Fryns JP, Kim JR, Lapunzina P, Lemyre E, Meire F, Messiaen LM, Oley C, Splitt M, Thomson J, Peer YV, Veitia RA, De Paepe A, De Baere E. Deletions involving long-range conserved nongenic sequences upstream and downstream of FOXL2 as a novel disease-causing mechanism in blepharophimosis syndrome. *Am J Hum Genet* 2005;77:205–18.
- [570] Crisponi L, Deiana M, Loi A, Chiappe F, Uda M, Amati P, Bisceglia L, Zelante L, Nagaraja R, Porcu S, Ristaldi MS, Marzella R, Rocchi M, Nicolino M, Lienhardt-Roussie A, Nivelon A, Verloes A, Schlessinger D, Gasparini P, Bonneau D, Cao A, Pilia G. The putative forkhead transcription factor FOXL2 is mutated in blepharophimosis/ptosis/epicanthus inversus syndrome. *Nat Genet* 2001;27:159–66.
- [571] Lettice LA, Horikoshi T, Heaney SJ, van_Baren MJ, van_der_Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A* 2002;99:7548–53.
- [572] Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* 2005;76:8–32.
- [573] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved elements in the human genome. *Science* 2004;304:1321–5.
- [574] Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 2004;5:456–65.
- [575] Dermitzakis ET, Reymond A, Antonarakis SE. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* 2005;6:151–7.
- [576] Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 2002;420:578–82.
- [577] Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongsong EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 2003;424:788–93.
- [578] Loots GG, Kneissel M, Keller H, Baptist M, Chang J, Collette NM, Ovcharenko D, Plajzer-Frick I, Rubin EM. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res* 2005;15:928–35.
- [579] Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet* 2003;34:157–65.
- [580] Pascoe L, Jeunemaitre X, Lebrethon MC, Curnow KM, Gomez-Sanchez CE, Gasc JM, Saez JM, Corvol P. Glucocorticoid-suppressible hyperaldosteronism and adrenal tumors occurring in a single French pedigree. *J Clin Invest* 1995;96:2236–46.
- [581] Nathans J, Piantanida TP, Eddy RL, Shows TB, Hogness DS. Molecular genetics of inherited variation in human color vision. *Science* 1986;232:203–10.
- [582] Francis NJ, McNicholas B, Awan A, Waldron M, Reddan D, Sadlier D, Kavanagh D, Strain L, Marchbank KJ, Harris CL, Goodship TH. A novel hybrid CFH/CFHR3 gene generated by a microhomology-mediated deletion in familial atypical hemolytic uremic syndrome. *Blood* 2011.
- [583] Bartram CR, de Klein A, Hagemeijer A, van Agthoven T, Geurts van Kessel A, Bootsma D, Grosveld G, Ferguson-Smith MA, Davies T, Stone M, et al. Translocation of c-abl oncogene correlates with the presence of a Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* 1983;306:277–80.
- [584] Delattre O, Zucman J, Plougastel B, Desmaze C, Melot T, Peter M, Kovar H, Joubert I, de Jong P, Rouleau G, et al. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* 1992;359:162–5.
- [585] Rabbitts TH. Chromosomal translocations in human cancer. *Nature* 1994;372:143–9.
- [586] Frank SA, Nowak MA. Problems of somatic mutation and cancer. *Bioessays* 2004;26:291–9.
- [587] Erickson RP. Somatic gene mutation and human disease other than cancer. *Mutat Res* 2003;543:125–36.
- [588] Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 1993;75:1027–38.
- [589] Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M, et al. Mutations of a mutS homolog

- in hereditary nonpolyposis colorectal cancer. *Cell* 1993;75:1215–25.
- [590] Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD, et al. Mutation of a mutL homolog in hereditary colon cancer. *Science* 1994;263:1625–9.
- [591] Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colon-ic carcinogenesis. *Nature* 1993;363:558–61.
- [592] Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, Fan RS, Zborowska E, Kinzler KW, Vogelstein B, et al. Inactivation of the type II TGF- β receptor in colon cancer cells with microsatellite instability. *Science* 1995;268:1336–8.
- [593] Schmutte C, Jones PA. Involvement of DNA methylation in human carcinogenesis. *Biol Chem* 1998;379:377–88.
- [594] Upadhyaya M, Han S, Consoli C, Majounie E, Horan M, Thomas NS, Potts C, Griffiths S, Ruggieri M, von Deimling A, Cooper DN. Characterization of the somatic mutational spectrum of the neurofibromatosis type 1 (NF1) gene in neurofibromatosis patients with benign and malignant tumors. *Hum Mutat* 2004;23:134–46.
- [595] Greenblatt MS, Grollman AP, Harris CC. Deletions and insertions in the p53 tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage/misalignment model. *Cancer Res* 1996;56:2130–6.
- [596] Jago N, Thomas G, Hamelin R. Short direct repeats flanking deletions, and duplicating insertions in p53 gene in human cancers. *Oncogene* 1993;8:209–13.
- [597] Kolomietz E, Meyn MS, Pandita A, Squire JA. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes Cancer* 2002;35:97–112.
- [598] Oldenburg J, Rost S, El-Maarri O, Leuer M, Olek K, Muller CR, Schwaab R. De novo factor VIII gene intron 22 inversion in a female carrier presents as a somatic mosaicism. *Blood* 2000;96:2905–6.
- [599] Ivanov D, Hamby SE, Stenson PD, Phillips AD, Kehrer-Sawatzki H, Cooper DN, Chuzhanova N. Comparative analysis of germline and somatic microlesion mutational spectra in 17 human tumor suppressor genes. *Hum Mutat* 2011;32:620–32.
- [600] Zlotogora J. Germ line mosaicism. *Hum Genet* 1998;102:381–6.
- [601] Hall JG. Review and hypotheses: somatic mosaicism: observations related to clinical genetics. *Am J Hum Genet* 1988;43:355–63.
- [602] Kehrer-Sawatzki H, Cooper DN. Mosaicism in sporadic neurofibromatosis type 1: variations on a theme common to other hereditary cancer syndromes? *J Med Genet* 2008;45:622–31.
- [603] Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet* 2015;31:382–92.
- [604] Pham J, Shaw C, Pursley A, Hixson P, Sampath S, Roney E, Gambin T, Kang SH, Bi W, Lalani S, Bacino C, Lupski JR, Stankiewicz P, Patel A, Cheung SW. Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10,362 consecutive cases. *Eur J Hum Genet* 2014;22:969–78.
- [605] Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K, Turner J, Cannons JL, Bick D, Blakemore L, Blumhorst C, Brockmann K, Calder P, Cherman N, Deardorff MA, Everman DB, Golas G, Greenstein RM, Kato BM, Keppler-Noreuil KM, Kuznetsov SA, Miyamoto RT, Newman K, Ng D, O'Brien K, Rothenberg S, Schwartzentruber DJ, Singhal V, Tirabosco R, Upton J, Wientroub S, Zackai EH, Hoag K, Whitewood-Neal T, Robey PG, Schwartzberg PL, Darling TN, Tosi LL, Mullikin JC, Biesecker LG. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* 2011;365:611–9.
- [606] Lindhurst MJ, Parker VE, Payne F, Sapp JC, Rudge S, Harris J, Witkowski AM, Zhang Q, Groeneveld MP, Scott CE, Daly A, Huson SM, Tosi LL, Cunningham ML, Darling TN, Geer J, Gucuv Z, Sutton VR, Tziotziou C, Dixon AK, Helliwell T, O'Rahilly S, Savage DB, Wakelam MJ, Barroso I, Biesecker LG, Semple RK. Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. *Nat Genet* 2012;44:928–33.
- [607] Lee JH, Huynh M, Silhavy JL, Kim S, Dixon-Salazar T, Heiberg A, Scott E, Bafna V, Hill KJ, Collazo A, Funari V, Russ C, Gabriel SB, Mathern GW, Gleeson JG. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nat Genet* 2012;44:941–5.
- [608] Lynch M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* 2010;107:961–8.
- [609] Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010;328:636–9.
- [610] Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. De novo rates and selection of large copy number variation. *Genome Res* 2010;20:1469–81.
- [611] Crow JF. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* 2000;1:40–7.

- [612] Becker J, Schwaab R, Moller-Taube A, Schwaab U, Schmidt W, Brackmann HH, Grimm T, Olek K, Oldenburg J. Characterization of the factor VIII defect in 147 patients with sporadic hemophilia A: family studies indicate a mutation type-dependent sex ratio of mutation frequencies. *Am J Hum Genet* 1996;58:657–70.
- [613] Grimm T, Meng G, Liechti-Gallati S, Bettecken T, Muller CR, Muller B. On the origin of deletions and point mutations in Duchenne muscular dystrophy: most deletions arise in oogenesis and most point mutations result from events in spermatogenesis. *J Med Genet* 1994;31:183–6.
- [614] Sayres MA, Venditti C, Pagel M, Makova KD. Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution* 2011;65:2800–15.
- [615] Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P. Variation in genome-wide mutation rates within and between human families. *Nat Genet* 2011;43:712–4.
- [616] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012;488:471–5.
- [617] Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands C, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, Slagboom PE, Boomsma DI, Ye K, Guryev V, Arndt PF, Kloosterman WP, de Bakker PIW, Sunyaev SR. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* 2015;47:822–6.
- [618] Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, Dominiczak A, Morris A, Porteous D, Smith B, Stratton MR, Consortium UK, Hurles ME. Timing, rates and spectra of human germline mutation. *Nat Genet* 2016;48:126–33.
- [619] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilcic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium, I.B.C., Consortium, I.M.-S., PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
- [620] Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, Veltman JA, Hoischen A, Vissers LE, Gilissen C. Post-zygotic point mutations are an under-recognized source of de novo genomic variation. *Am J Hum Genet* 2015;97:67–74.
- [621] Goldmann JM, Wong WS, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LE, Hoischen A, Roach JC, Vockley JG, Veltman JA, Solomon BD, Gilissen C, Niederhuber JE. Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* 2016;48:935–9.
- [622] Wexler NS, Young AB, Tanzi RE, Travers H, Starosta-Rubinstein S, Penney JB, Snodgrass SR, Shoulson I, Gomez F, Ramos Arroyo MA, et al. Homozygotes for Huntington's disease. *Nature* 1987;326:194–7.
- [623] Zlotogora J. Dominance and homozygosity. *Am J Med Genet* 1997;68:412–6.
- [624] Cogan JD, Phillips 3rd JA, Schenkman SS, Milner RD, Sakati N. Familial growth hormone deficiency: a model of dominant and recessive mutations affecting a monomeric protein. *J Clin Endocrinol Metab* 1994;79:1261–5.
- [625] Spritz RA, Giebel LB, Holmes SA. Dominant negative and loss of function mutations of the c-kit (mast/stem cell growth factor receptor) proto-oncogene in human piebaldism. *Am J Hum Genet* 1992;50:261–9.
- [626] Patel PI, Roa BB, Welcher AA, Schoener-Scott R, Trask BJ, Pentao L, Snipes GJ, Garcia CA, Francke U, Shooter EM, Lupski JR, Suter U. The gene for the peripheral myelin protein PMP-22 is a candidate for Charcot-Marie-Tooth disease type 1A. *Nat Genet* 1992;1:159–65.
- [627] Aldred MA, Trembath RC. Activating and inactivating mutations in the human GNAS1 gene. *Hum Mutat* 2000;16:183–9.
- [628] Mooers BH, Logue JS, Berglund JA. The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc Natl Acad Sci U S A* 2005;102:16626–31.
- [629] Armanios M, Chen JL, Chang YP, Brodsky RA, Hawkins A, Griffin CA, Eshleman JR, Cohen AR, Chakravarti A, Hamosh A, Greider CW. Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc Natl Acad Sci U S A* 2005;102:15960–4.

- [630] Kim HJ, Nam SH, Kim HJ, Park HS, Ryoo HM, Kim SY, Cho TJ, Kim SG, Bae SC, Kim IS, Stein JL, van Wijnen AJ, Stein GS, Lian JB, Choi JY. Four novel RUNX2 mutations including a splice donor site result in the cleidocranial dysplasia phenotype. *J Cell Physiol* 2006;207:114–22.
- [631] Dobyns WB, Filauro A, Tomson BN, Chan AS, Ho AW, Ting NT, Oosterwijk JC, Ober C. Inheritance of most X-linked traits is not dominant or recessive, just X-linked. *Am J Med Genet A* 2004;129:136–43.
- [632] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarrroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53.
- [633] Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 2010;42:570–5.
- [634] Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;40:695–701.
- [635] Bodmer W, Tomlinson I. Rare genetic variants and the risk of cancer. *Curr Opin Genet Dev* 2010;20:262–7.
- [636] Mefford HC, Shafer N, Antonacci F, Tsai JM, Park SS, Hing AV, Rieder MJ, Smyth MD, Speltz ML, Eichler EE, Cunningham ML. Copy number variation analysis in single-suture craniosynostosis: multiple rare variants including RUNX2 duplication in two cousins with metopic craniosynostosis. *Am J Med Genet A* 2010;152A:2203–10.
- [637] Lehner B. Molecular mechanisms of epistasis within and between genes. *Trends Genet* 2011;27:323–31.
- [638] Wolf U. Identical mutations and phenotypic variation. *Hum Genet* 1997;100:305–21.
- [639] Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* 2013;132:1077–130.
- [640] Parchi P, Petersen RB, Chen SG, Autilio-Gambetti L, Capellari S, Monari L, Cortelli P, Montagna P, Lugaresi E, Gambetti P. Molecular pathology of fatal familial insomnia. *Brain Pathol* 1998;8:539–48.
- [641] Kajiwara K, Berson EL, Dryja TP. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* 1994;264:1604–8.
- [642] Morell R, Spritz RA, Ho L, Pierpont J, Guo W, Friedman T.B., Asher Jr. J.H.. Apparent digenic inheritance of Waardenburg syndrome type 2 (WS2) and autosomal recessive ocular albinism (AROA). *Hum Mol Genet* 1997;6:659–64.
- [643] Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, Lupski JR. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* 2001;293:2256–9.
- [644] Nichols WC, Ginsburg D. von Willebrand disease. *Medicine (Baltimore)* 1997;76:1–20.
- [645] Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 2011.
- [646] Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigo R. Identifying protein-coding genes in genomic sequences. *Genome Biol* 2009;10:201.
- [647] Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, Crawford GE. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011;9:e1001046.
- [648] Ying H, Huttley G. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome Biol Evol* 2011;3:938–49.
- [649] Millar DS, Lewis MD, Horan M, Newsway V, Easter TE, Gregory JW, Fryklund L, Norin M, Crowne EC, Davies SJ, Edwards P, Kirk J, Waldron K, Smith PJ, Phillips 3rd JA, Scanlon MF, Krawczak M, Cooper DN, Procter AM. Novel mutations of the growth hormone 1 (GH1) gene disclosed by modulation of the clinical selection criteria for individuals with short stature. *Hum Mutat* 2003;21:424–40.
- [650] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;33(Suppl.):228–37.

Genes in Families

Jackie Cook

Consultant in Clinical Genetics, Sheffield Clinical Genetics Service, Sheffield Children's NHS Foundation Trust,
Sheffield, United Kingdom

7.1 INTRODUCTION

This chapter provides an overview of genetic conditions within families and how to approach and analyze a family history of genetic disease. Many of the topics touched upon are covered in more detail elsewhere in this volume. The principles of Mendelian inheritance for single nuclear gene disorders have been known for many years, but new insights are being constantly gained from new technologies, particularly whole-genome sequencing. An increasing understanding of nontraditional inheritance patterns and the study of the genome rather than single genes are adding to our knowledge and enabling clinicians to provide information to a growing number of families. Most of our current understanding of genetic disease is related to pathogenic variations affecting the exomes or coding regions of the genes. It is likely that, with the advent of whole-genome sequencing, we will start to identify a whole new group of genetic disorders and/or explain disorders of unknown etiology that are caused by alterations in the non-protein-coding parts of the genome, including sequences involved in the regulation of gene expression both in time and at the tissue site.

Genetic counseling is the process by which patients and relatives at risk of a disorder that may be hereditary are advised of the consequences of the disorder, the probability of developing and transmitting it, and the ways in which this may be prevented, avoided, or ameliorated. To achieve these aims, an accurate diagnosis and detailed information regarding the family

history are essential. The basis for establishing a diagnosis depends on medical history, examination, and investigation. The diagnostic information is combined with information obtained from the family pedigree to determine the mode of inheritance of the disorder and to calculate the risk of recurrence, so that family members can be appropriately counseled.

A family history of a genetic condition may be due to:

1. a pathogenic variant within a single nuclear gene;
2. a pathogenic variant within a mitochondrial gene;
3. a contiguous gene deletion or duplication involving one or many genes;
4. a chromosomal rearrangement resulting in unbalanced products at meiosis;
5. polygenic and multifactorial inheritance;
6. multiple genetic conditions contributing to the overall phenotype.

7.2 PEDIGREE CONSTRUCTION

Accurate documentation of the family history is an essential part of genetic assessment, and the best method of recording this information is by constructing a family pedigree. Pedigrees are universally used in patients' genetic records, journal articles, and textbooks as the means of relaying information in an easily interpretable visual format. Pedigrees also provide the basis for calculations required for both recurrence risk estimation in individual families and linkage analysis in gene-mapping studies.

In a pedigree, squares are used to represent males and circles represent females. Generations are indicated by Roman numerals and individuals within each generation by Arabic numbers. Despite the universal use of the pedigree as a method of recording information and as an analytical tool, there is still considerable variation in the use of symbols relating to both routine medical information (pregnancy, spontaneous abortion, and termination of pregnancy) and new reproductive technologies (artificial insemination by donor semen, donor ovum, and surrogate motherhood).

An example of a family pedigree is shown in Fig. 7.1, using the symbols illustrated in Fig. 7.2. When drawing a pedigree, it is usually simplest to start with the person seeking advice (the consultand). In some cases, the consultand will be an apparently healthy relative seeking information about how the condition may affect him or her or their offspring. Pedigree details are completed for both sides of the family, including previous and subsequent generations.

Conventionally, the paternal lineage is placed on the left and the maternal lineage on the right. Within sibships, individuals are listed from left to right in birth order. The affected person (proband), through whom the family has been ascertained, or the relative seeking advice (the consultand) is normally indicated in the pedigree by an arrow. Name, date of birth, and a summary of relevant medical details should be recorded for all family members. Further information, including medical records, may be required for affected individuals to confirm actual diagnoses. Not all relevant information may be volunteered by the consultand; indeed certain sensitive information—such as a previous termination of pregnancy or illegitimacy—may be deliberately withheld if the accompanying partner is

not aware of it. Important details that should be asked about include assisted conception, previous miscarriages, stillbirths, terminations, children adopted into or out of the family, and consanguinity. Ethnic origin should be recorded, as some genetic conditions are more prevalent in particular ethnic groups, and this information may provide a clue to the likely diagnosis. The family history should be completed on both sides of the family for a consultand couple, even if the presenting condition is clearly traced to one side, as unrelated pregnancy losses or other relevant conditions may become apparent on the other side of the family that could have greater reproductive impact for the couple than the disorder for which they are seeking information.

7.3 UNIFACTORIAL INHERITANCE/SINGLE-GENE DISORDERS

Unifactorial inheritance refers to those disorders that are due to the inheritance of a single pathogenic gene variant. The first descriptions of unifactorial inheritance were made by Mendel in 1865, when he published the results of his experiments on the garden pea in his paper “Versuche uber Pflanzen Hybriden” (“Experiments on Plant Hybrids”). His work was largely ignored until it was popularized by Bateson in 1901, from which time the term “Mendelian inheritance” became synonymous with unifactorial inheritance.

From the ratios that Mendel described in his experiments on the garden pea, and the work of subsequent researchers, including Bateson, four main conclusions were drawn [1]:

1. Genes come in pairs (Mendel termed them factors), one inherited from each parent.

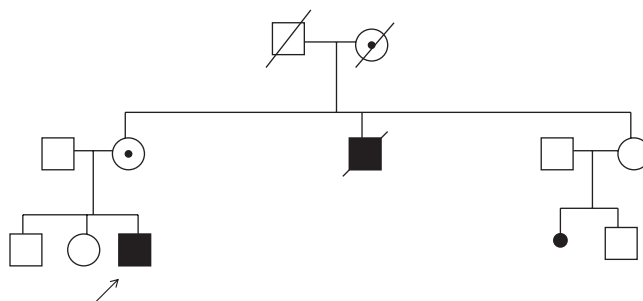


Figure 7.1 Example of a family pedigree for Duchenne muscular dystrophy.

- Individual genes can have different alleles, some of which (dominant traits) exert their effects over others (recessive traits)—the principle of dominance. In Mendel's own words "those characters which are transmitted entire, or almost unchanged in the hybridisation, and therefore in themselves constitute the characters of the hybrid, are termed the dominant, and those which become latent in the process, recessive."
- At meiosis, alleles segregate from each other, with each gamete receiving only one allele—the principle of segregation, or Mendel's first law.
- The segregation of different pairs of alleles is independent—the principle of independent assortment, or Mendel's second law.

With time, these principles have had to be modified. For example, although most genes come in pairs, for genes on the sex chromosomes males have only one allele, that is, they are termed hemizygous. Also, although alleles of a gene on different chromosomes show independent assortment, genes that are physically close together on the same chromosome do not—a phenomenon that has allowed mapping of genes in the human genome through linkage studies. These principles, however, still form a useful set of rules designed to explain the inheritance of many inherited characteristics and disorders.

Disorders inherited in a Mendelian fashion are categorized according to whether the gene is on an autosome, the chromosomes shared in common between males and females, or a sex chromosome and whether the trait is dominant or recessive.

7.4 DOMINANCE AND RECESSIVENESS

7.4.1 Definition of Dominance

The concepts of dominance and recessiveness are fundamental to the understanding of Mendelian inheritance. Dominance is not a property intrinsic to a particular allele but describes the relationship between it and the corresponding allele on the homologous chromosome. If the phenotypes associated with the genotypes AA and AB are the same, but differ from the phenotype of BB, allele A is dominant to allele B and, conversely, allele B is recessive to allele A. Therefore, allele A manifests in the heterozygous state. An example of a dominant allele is that for Huntington disease, with most individuals affected with the disease being heterozygous for a variant allele. Individuals have been identified who have been shown by molecular techniques to be homozygous for the variant allele by virtue of both parents being

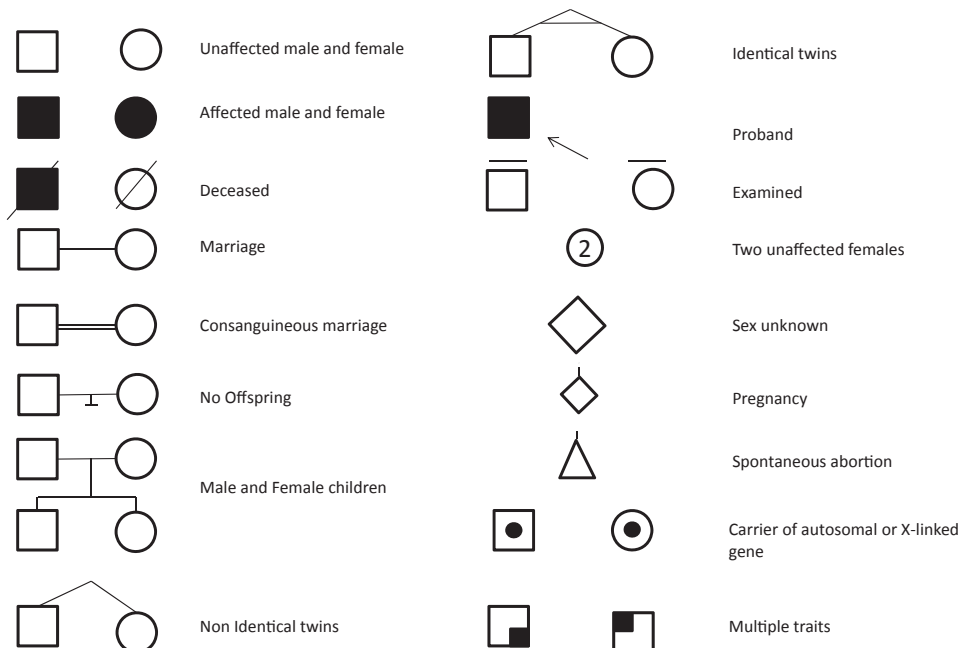


Figure 7.2 Examples of symbols commonly used in drawing a pedigree.

affected with Huntington disease. Such individuals do not appear different phenotypically from heterozygotes for the disorder.

7.4.1.1 Incomplete Dominance

If the phenotype of the heterozygous state, AB, is intermediate between the phenotypes of AA and BB, allele A is said to be incompletely dominant or semidominant to allele B. The skeletal dysplasia achondroplasia causes rhizomelic shortening of the limbs, a characteristic facies with midface hypoplasia, exaggerated lumbar lordosis, limitation of hip and elbow extension, genu varum, and trident hand. It was conventionally thought to be due to a dominant allele, but homozygotes for the variant gene have a much more severe skeletal dysplasia, resulting in early death from respiratory obstruction due to a small thoracic cage and neurologic deficit due to hydrocephalus. Therefore, achondroplasia is an example of incomplete or semidominance. Homozygotes for most dominant alleles causing human genetic disorders occur so rarely that it is not known whether they exhibit complete or incomplete dominance.

7.4.1.2 Codominance

If the phenotype of AB displays the phenotypic features of both the homozygotic states, then alleles A and B are said to be codominant. The human ABO blood group system exhibits codominance. The system consists of three alleles A, B, and O. Both A and B are dominant in relation to O, and therefore blood group A can have the genotype AA or AO. Blood group B can have the genotype BB or BO. However, neither A nor B shows dominance over the other, and therefore individuals with the genotype AB have the phenotypic characteristics of both blood group A and blood group B.

7.4.2 Mechanisms of Dominance

Most pathogenic variants result in an allele that is recessive to the wild-type allele; the phenotype is therefore expressed only in the homozygous state. This is because most pathogenic variants result in an inactive gene product, but the reduced level of activity due to the remaining wild-type allele is sufficient to achieve the effects of that gene product. An example is a gene for an enzyme that is required only in small amounts as a catalyst for a metabolic pathway. Although in some recessive inborn errors of metabolism it is possible to identify heterozygotes biochemically, more often than not, the only way to identify carriers is by direct mutation analysis, using molecular genetic techniques.

There are several mechanisms by which a pathogenic variant can lead to a dominant allele whose phenotype is expressed in the heterozygous state [2,3].

7.4.2.1 Loss-of-Function Variants

For most variant alleles, loss of function will usually exhibit recessive behavior. Where a reduced amount or reduced activity of the gene product results in the phenotypic features, this is termed haploinsufficiency (e.g., in a critical rate-limiting step of a metabolic pathway). Haploinsufficiency can be due to a number of different mechanisms, including the following:

1. deletion of a whole allele;
2. a pathogenic variant causing gene inactivation;
3. a regulatory variant resulting in failure of transcription;
4. incorrect translational control;
5. decreased mRNA or protein stability; for example, a premature termination codon can lead to non-sense-mediated mRNA decay.

Any reduction in the amount of gene product will result in that pathway not being able to function at full activity. The same appears to apply to regulatory genes that could have a threshold level of activity. *PAX3* is a gene coding for a DNA-binding protein, and single-nucleotide variants in this gene result in Waardenburg syndrome type 1, characterized by deafness and pigmentary disturbances. Certain variants in *PAX3* have been shown to abolish all protein function of that allele, so the phenotype must be due to a dosage effect as it manifests in the heterozygous state.

Other examples in which the quantitative amount of a gene product is important are genes that produce proteins in large quantities. An example is the *C1NH* gene for C1 esterase inhibitor, in which pathogenic variants cause the disorder hereditary angioneurotic edema. C1 esterase inhibitor is removed rapidly from the circulation at a rate independent of its concentration. Therefore, although heterozygotes produce 50% of the normal amount, they have only 15%–20% of the normal amount in the circulation, leading to the clinical manifestations of the disorder.

7.4.2.2 Gain-of-Function Variants

7.4.2.2.1 Increased gene dosage. This mechanism involves an excess of gene product leading to a disease phenotype. Although the gene dosage of critical regions or genes has been invoked as the cause of the phenotypic

features associated with the autosomal trisomies, there are few examples involving single-gene disorders. One example involves the *PMP22* gene, which codes for the peripheral myelin protein 22. Duplication of the DNA sequence of one allele is associated with hereditary motor and sensory neuropathy type 1A.

7.4.2.2.2 Ectopic or temporally altered messenger RNA expression. Genes can be expressed or turned off at different times throughout development and the life of the individual and are also differentially expressed in different tissues. Ectopic or temporally altered messenger RNA is expressed when a pathogenic variant occurs that affects the time or place of gene expression and usually involves a regulatory part of the gene. For example, during development in erythroid precursor cells, there is a switch from the production of γ -globin to the production of δ - and β -globin. This switch is controlled, at least in part, by the binding of transcription factors to the γ -globin promoter. Single-nucleotide variants in the globin promoter region prevent the normal switch, resulting in the disorder of hereditary persistence of fetal hemoglobin.

7.4.2.2.3 Increased protein activity. Pathogenic variants can lead to proteins with a prolonged half-life or proteins that have lost their normal constitutive inhibitory regulatory activity. If a pathogenic variant occurs in a part of a gene that codes for the protein sequence acting as the recognition site for proteolytic degradation, this will not take place, with the protein remaining active. Many proteins possess domains that allow their activity to be reversibly inhibited. For example, skeletal muscle sodium channels undergo voltage-sensitive regulation, and variants in the gene *SCN4A*, which codes for the α subunit of the sodium channels, result in the disorder hyperkalemic periodic paralysis, characterized by muscle myotonia and paralysis due to loss of regulatory inactivation of the sodium channel.

7.4.2.2.4 Dominant-negative variants. If a variant allele interferes with the wild-type allele, this is termed a dominant-negative variant. This could occur in a multimeric protein in which an abnormal subunit has an intact binding domain but altered catalytic activity, affecting the function of the entire multimer. If a protein is a dimer, one variant and one wild-type allele would result in only 25% normal dimers, with up to a 75% reduction in activity.

Many structural proteins are multimers (e.g., the various types of collagen proteins). Each of the collagen

subunit genes has a central portion coding for repeating tripeptide units that are essential for the assembly of the collagen molecule. The disorder osteogenesis imperfecta is caused by single-nucleotide variants in the central portion of one of the collagen subunit genes *COL1A1* or *COL1A2* leading to a structural deformation that causes disruption of the whole collagen protein.

7.4.2.2.5 Toxic protein alterations. Toxic protein alterations are pathogenic variants that cause structural alterations in proteins, thus disrupting normal function and leading to toxic products that poison the cell. An example is hereditary amyloidosis, in which variants in the transthyretin gene *TTR* lead to resistance to proteolysis, with resultant increased stability of the protein. The protein then undergoes multimerization and accumulates in the cell as fibrils, causing disruption of the cell.

7.4.2.2.6 New protein functions. Some variants have been found to confer a new function on a gene product. For example, a fatal bleeding disorder was found to be caused by a missense pathogenic variant in the $\alpha 1$ -antitrypsin gene, in which methionine was replaced by arginine at position 358, the effect of which was to convert $\alpha 1$ -antitrypsin, normally an inhibitor of elastase, into an inhibitor of thrombin. This thrombin-inhibitory activity was not compensated for by an increase in endogenous coagulant production, resulting in a severe bleeding disorder [4].

7.4.2.3 Recessive Variants With Dominant Effects

The mechanisms described so far show how variants can cause dominant effects at a cellular level by the effects on the proteins produced. It is possible to have variants that show a dominant pattern of inheritance in families, yet are recessive at the cellular or molecular level; that is, the gene is inactivated but has no other effect. The classic example of this is the retinoblastoma gene *RBI*, inactivation of which can lead to the formation of the developmental eye tumor retinoblastoma. Families can show a dominant mode of inheritance for this disorder, yet cells heterozygous for the variant are completely normal, the variant itself being recessive.

The dominant pattern of inheritance of familial retinoblastoma is the result of transmission of a first variant with a second somatic variant occurring in the normal allele of at least one retinal cell during a critical period of development, leading to the formation of a retinoblastoma—the “two-hit” hypothesis [5]. There are several ways in which the normal allele in somatic cells can be

inactivated. These include single-nucleotide variants, deletions, translocations, and mitotic nondisjunction, resulting in the loss of a whole chromosome. It is now known that the two-hit hypothesis applies to most of the dominantly inherited familial cancer syndromes in which the germline variant in a tumor suppressor gene is recessive and an acquired variant in a somatic cell in the corresponding allele leads to the development of a tumor.

7.5 AUTOSOMAL DOMINANT INHERITANCE

Autosomal dominant inheritance refers to disorders caused by genes located on the autosomes, thereby affecting both males and females. The variant alleles are dominant to the wild-type alleles, so the disorder is manifest in the heterozygote (i.e., an individual who possesses both the wild-type and the variant allele). The necessary characteristics to be certain a disorder is inherited in an autosomal dominant manner are listed in Table 7.1 and depicted in Fig. 7.3.

TABLE 7.1 Characteristics of an Autosomal Dominant Inherited Disorder
Successive or multiple generations in a family are affected
Males and females are both affected in approximately equal proportions
Males and females can both be responsible for transmission
There is at least one instance of male-to-male transmission

7.5.1 Recurrence Risks

Because individuals with autosomal dominant disorders are heterozygous for a variant and a normal allele, there is a 1 in 2 (50%) chance a gamete will carry the normal allele and a 1 in 2 (50%) chance a gamete will carry the variant allele. Assuming that the individual’s partner will contribute a normal allele, there is a 1 in 2 (50%) chance that the offspring, regardless of sex, will inherit the disorder with each pregnancy (Fig. 7.4).

7.5.2 Penetrance

There can be marked variability in the clinical manifestations of autosomal dominant disorders, and they can demonstrate reduced penetrance (i.e., not every person with the variant allele shows features of the disorder). The penetrance of a disorder is an index of the proportion of individuals with a variant allele who manifest the disorder. An allele is said to be nonpenetrant if an individual known to be heterozygous for the allele, either by pedigree analysis or by molecular investigation, shows no signs of the disorder when subjected to appropriate clinical investigation. The penetrance of some genes is dependent on the age of the individual, as in Huntington disease, in which the penetrance is age dependent or is said to show delayed penetrance.

The penetrance of a disorder is usually expressed as the proportion or percentage of the individuals carrying the gene who develop the disorder. If the penetrance is known for a particular condition, the risk to the offspring of an apparently unaffected individual can be calculated. In practice, the risk is usually less than 10%, as an unaffected relative is not likely to carry

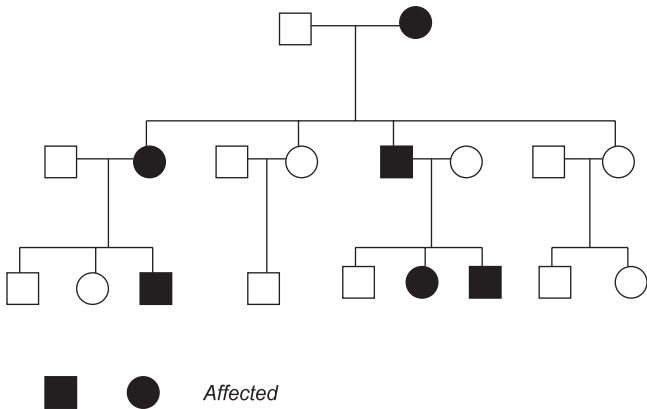


Figure 7.3 Pedigree consistent with autosomal dominant inheritance.

the gene if the penetrance is high, and a gene carrier is not likely to develop the disorder if the penetrance is low.

7.5.3 Expressivity

The expressivity of a gene is the degree to which a particular phenotype is expressed in an individual. Many autosomal dominant disorders show variable expressivity such that individuals in the same family who carry an identical variant can vary considerably in the severity of their disorder. For example, in the autosomal dominant disorder neurofibromatosis type 1, the number of neurofibromas that an individual develops can vary dramatically from a few to many hundreds even within the same family. The variability seen in autosomal dominant disorders may present as both inter- and intrafamilial differences. Intrafamilial variability may reflect the action of modifying genes, but interfamilial variability can also be due to allelic heterogeneity at a single locus. A problem encountered in genetic counseling is that a mildly affected individual, such as a parent with only skin manifestations of tuberous sclerosis, may have a severely affected child. This situation is seen in many autosomal dominant disorders, in that the mildly affected individuals are more likely to reproduce than the severely affected individuals.

7.5.4 Anticipation

A disorder is said to demonstrate anticipation if the phenotype of the variant allele increases in severity as it is passed down the generations. An example of a disorder

that demonstrates anticipation is myotonic dystrophy. A typical three-generation family with myotonic dystrophy showing anticipation is shown in Fig. 7.5. Another example of anticipation is seen in Huntington disease, in which the onset of symptoms is often seen to occur earlier with each succeeding generation.

7.5.5 Sex Influence

Sex influence involves the expression of an autosomal allele that occurs more frequently in one sex than the other. An example in humans is early pattern baldness, with males affected more frequently than females, an effect probably mediated by hormonal differences.

7.5.6 Sex Limitation

Some traits are manifested only in individuals of one sex, an extreme situation known as sex limitation. This can occur when a gene affects an organ possessed by only one of the sexes (e.g., unicornuate uterus or ovarian cancer).

7.5.7 Pleiotropy

Pleiotropy refers to the phenomenon in which a single gene is responsible for a number of distinct and seemingly unrelated phenotypic effects. For example, the allele causing neurofibromatosis type 1 can produce abnormalities of skin pigmentation, neurofibromas of the peripheral nerves, short stature, macrocephaly, skeletal abnormalities, and fits. Each of the pleiotropic effects of an allele can show reduced or nonpenetrance and variable expressivity.

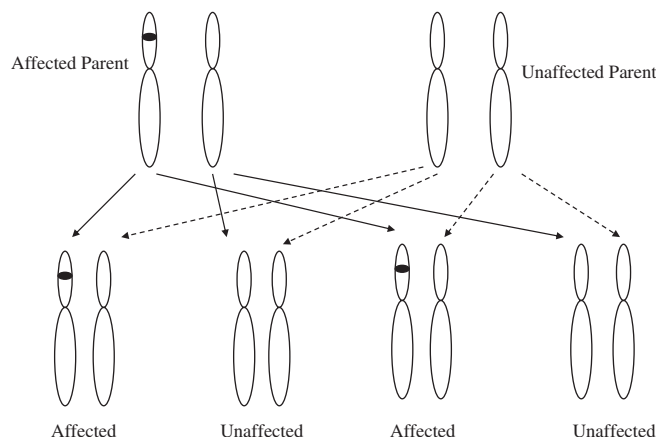


Figure 7.4 Recurrence risks in autosomal dominant inheritance.

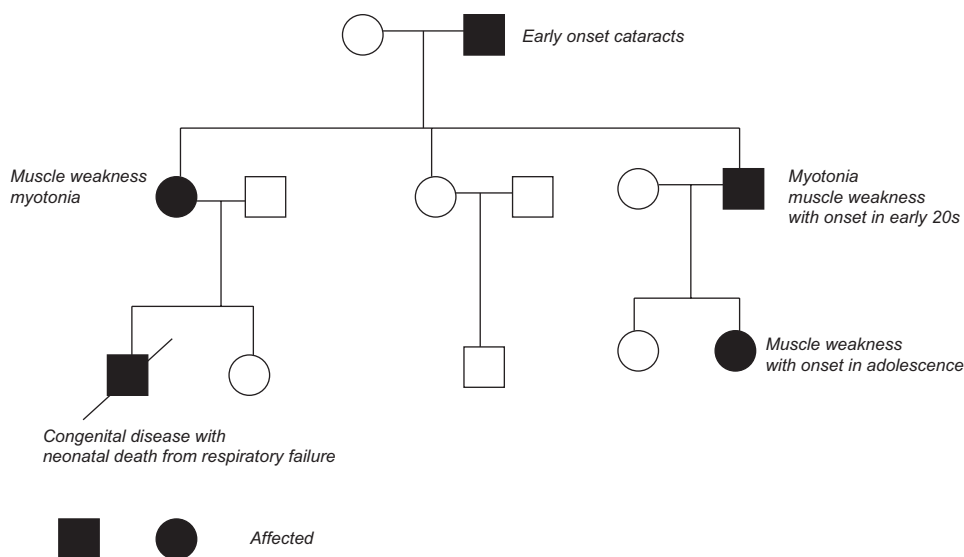


Figure 7.5 Pedigree for a family with myotonic dystrophy demonstrating anticipation.

7.5.8 Mechanisms of Reduced Penetrance and Variable Expressivity

Some of the underlying mechanisms accounting for reduced penetrance and variable expressivity have been known for some time, while others are only now becoming apparent with the advent of modern molecular genetic techniques such as whole-genome sequencing.

7.5.8.1 Environmental Factors

Environmental influences can affect the expression of genes. These can involve factors in the internal environment, such as hormones, or in the external environment, such as the effects of certain drugs, for example, barbiturates in acute intermittent porphyria. This disorder is characterized by attacks of abdominal pain, constipation, and psychiatric disturbances and is due to a pathogenic variant in the *HMBS* gene coding for an enzyme involved in heme biosynthesis. Attacks can be precipitated by certain drugs, including phenobarbital and the sulfonamides. Avoidance of these precipitating factors will result in nonpenetrance of the disorder.

7.5.8.2 Somatic Variants

Retinoblastoma has already been mentioned as an example of one of the familial cancer genes in which a “second-hit” somatic mutation needs to occur for the disorder to manifest.

7.5.8.3 Unstable DNA Triplet-Repeat Sequences

There is a group of dominant genetic disorders in which the variant is an unstable DNA triplet-repeat sequence that expands in successive meioses and whose size correlates with the severity of the disorder. The larger the expansion, the more unstable it becomes, which can result in increasingly larger increases in size as it is passed down through the generations. This accounts for the anticipation seen in such disorders as myotonic dystrophy and Huntington disease.

7.5.8.4 Genetic Background

New techniques of examining the whole genome have shown numerous examples of the ways in which both variants within the same gene and variants in other genes that interact with the variant allele influence the penetrance and expression of a disorder. Analysis of the effects of this type of interaction is complex and can involve large numbers of genes, each producing a small effect on the phenotype.

An example of how a variant in the other allele of a gene can influence the phenotype is seen in autosomal dominant polycystic kidney disease (ADPKD). ADPKD is caused by dominant pathogenic variants in one of two genes, *PKD1* and *PKD2*. In general, this is an adult-onset disease but in some families a parent with an adult-onset disease can have a child with renal cysts developing in infancy. In some families this has been shown to be due

to the inheritance of the variant allele from the affected parent and a milder missense variant from the unaffected parent. This hypomorphic or incompletely penetrant allele is not disease causing in the heterozygous state but when paired with a pathogenic variant causes increased disease severity [6].

Also in ADPKD, a comparison of monozygotic twins to siblings showed greater variance in time to end-stage renal failure between siblings, suggesting the influence of genetic modifiers involving genes other than the PKD genes. Differences due to genetic background and the presence or absence of genetic modifiers was estimated to account for anything between 18% and 59% of the phenotypic variance up to end-stage renal failure.

Similarly, numerous studies of families carrying a variant in one of the high-risk familial cancer genes, such as *BRCA1* and *BRCA2*, which cause a high risk for breast and ovarian cancer, have shown that large numbers of variants in other genes contribute to the risk of developing cancer associated with the *BRCA* variant.

7.5.9 New Dominant Variants

While nonpenetrance can be a possible cause of a dominant disorder arising in the offspring of completely normal parents, an alternative explanation is that a variant has arisen during transmission of the gene, that is, it represents a new or de novo germline variant. Genome and exome sequencing studies of parent-offspring trios have provided new insights into the frequency of de novo variants. Different studies have suggested anything from 44 up to 82 de novo single-nucleotide variants in the average genome, with one or two affecting the coding sequence [7].

This appears to be more common for certain disorders than others and indeed mutagenesis does not occur randomly across the genome. This can be explained by intrinsic characteristics of the genome, such as timing of replication, transcriptional activity, and chromatin state. An example of a gene with a high mutation rate is the *FGFR3* gene, causing achondroplasia. In achondroplasia both parents are of normal stature in 80% of families. This also reflects the reduced reproductive fitness of adults with achondroplasia. The observation that achondroplasia occurs more frequently in the last-born children of a sibship was suggested by Penrose as attributable to increased paternal age associated with new variants, on the basis that “older germ-cells, possessed by older parents, might be more likely to show deterioration in the form of genetical changes” [8]. This

suggestion has been confirmed using genome sequencing, and approximately 80% of all de novo germline single-nucleotide variants arise on the paternal allele. Advanced paternal age at conception is the major factor linked to the increase in de novo variants in the offspring. One of the reasons for this is thought to be that spermatogonial cells divide throughout life, resulting in the progressive accumulation of variants due to errors during DNA replication.

Some dominant variants are universally lethal before the affected individual reaches reproductive age, and therefore are always seen as new dominant variants. Several lethal disorders that were previously considered to be recessive are now known in many instances to be due to new dominant variants, such as the perinatal lethal form (type II) of osteogenesis imperfecta caused by variants in *COL1A1* and *COL1A2*.

New dominant variants that occur during gametogenesis are associated with a negligible recurrence risk for future siblings. However, if the disorder is compatible with survival to reproductive age (and the possibility of reproduction), the recurrence risk for the offspring of the affected individual is 50%.

7.5.10 Gonadal or Somatic Mosaicism

In the case of certain new dominant variants, there is a small but significant risk that a second child will be affected despite both parents being clinically normal. The ability to determine the molecular basis of many of these disorders has shown that this finding can be explained by the phenomenon of somatic mosaicism. This is when there are two genetically different types of cell in an individual, one carrying the variant allele and the other not. If two genetically different types of cell occur within the gonads, this is referred to as gonadal mosaicism; however, there may be no mosaicism present in the gonads, where all cells may carry the variant gene, and the mosaicism may be present in other tissues. This situation would more correctly be referred to as somatic mosaicism.

The mutational event occurs after fertilization, with the degree and tissue specificity of the mosaicism being dependent upon the time when it occurred. There is evidence that this usually occurs early in development, as commitment of primordial cells to the germline occurs before tissue allocation, and in studies of individuals with gonadal mosaicism, up to 50% also have the variant in a somatic cell line.

The frequency of gonadal mosaicism differs between disorders. As of this writing, it is unclear why it is a frequent finding in some conditions and not others. Studies on the recurrence of lethal osteogenesis imperfecta show genetic heterogeneity, with some cases due to recessive inheritance and some due to dominant parental mosaicism. The rate of parental mosaicism in families in which a dominant variant was identified in the first affected child has been reported in one study as 16% with a recurrence rate of 1.3% [9]. Recurrence of achondroplasia to normal stature parents, where there is a high new germline mutation rate, has been reported in only a few rare cases. This is an important point to remember when providing recurrence risk advice in genetic counseling. Counseling in cases of gonadal mosaicism is complex because the recurrence risk is dependent on the percentage of cells carrying the variant in the gonads, which can be difficult to determine.

Somatic mosaicism may also be present in an individual who manifests the phenotype of an autosomal dominant disorder. A study looking at individuals who were the first members of their families to develop the signs of neurofibromatosis type 2, a disorder characterized by bilateral vestibular schwannomas, came up with a mosaicism rate of 33% for classical neurofibromatosis type 2 with bilateral tumors and 60% for those presenting with unilateral tumors, with the variant often detected in tumor material and not in lymphocyte DNA

[10]. Again, the risk of passing the variant on to the next generation is determined by whether the variant is present in the gonads, and this information is not straightforward to ascertain.

7.6 AUTOSOMAL RECESSIVE INHERITANCE

Autosomal recessive inheritance refers to disorders due to genes located on the autosomes, but in which the variant alleles are recessive to the wild-type alleles and are therefore not evident in the heterozygous state, being manifest only in the homozygous state. The necessary characteristics to be certain that a disorder is inherited in an autosomal recessive manner are listed in Table 7.2 and depicted in Fig. 7.6. The parents of an individual with an autosomal recessive disorder are heterozygous for the variant allele and are usually referred to as being carriers for the disorder.

TABLE 7.2 Characteristics of an Autosomal Recessive Inherited Disorder
Both males and females are affected
The disorder normally occurs in only one generation, usually within a single sibship
The parents can be consanguineous

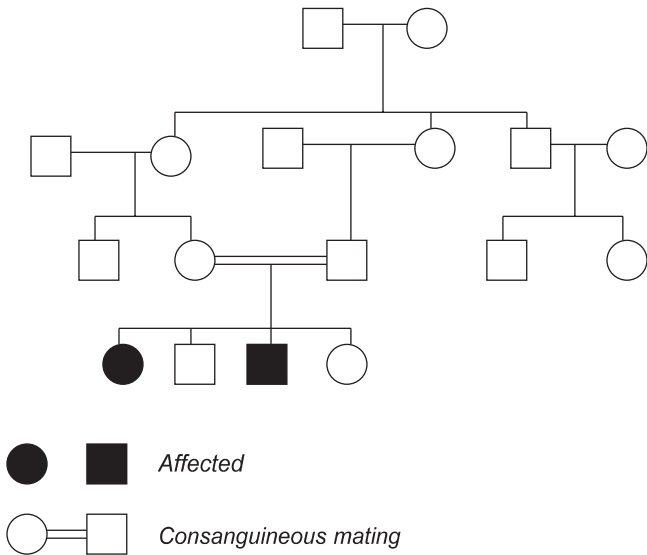


Figure 7.6 Pedigree consistent with autosomal recessive inheritance.

7.6.1 Consanguinity

If a couple are consanguineous, they have at least one ancestor in common in the preceding few generations. First cousins share approximately 1/8 of their alleles in common. This means that they are more likely to carry identical alleles inherited from their common ancestor and could both transmit an identical allele to their offspring, who would then be homozygous for that allele. A consanguineous couple has an increased risk that their offspring will be affected with a recessive disorder. The rarer a particular disorder is in a population, the more likely the parents are to be consanguineous. For example, cystic fibrosis is a common autosomal recessive disorder in whites in Western Europe, with an incidence of approximately 1 in 2000. The incidence of consanguinity in the parents of children with cystic fibrosis is not appreciably greater than that in the general population. By contrast, with very rare autosomal recessive disorders such as alkaptonuria, 8 of the first 19 families originally described by Garrod were consanguineous [11].

7.6.2 Recurrence Risks

When two parents, who each carry a variant allele for a genetic disorder, reproduce, there is an equal chance that the gametes will contain the variant or the wild-type allele. There are four possible combinations of these gametes, resulting in a 1 in 4 (25%) chance of having a homozygous affected offspring, a 1 in 2 (50%) chance of having a heterozygous unaffected carrier offspring, and a 1 in 4 (25%) chance of having a homozygous unaffected offspring (Fig. 7.7).

When an individual with an autosomal recessive disorder has children, they will produce only gametes containing the variant allele. Since it is most likely that their partner will be homozygous for the wild-type allele, the partner will always contribute a wild-type allele and therefore all the children will be heterozygous carriers and unaffected (Fig. 7.8). If, however, an affected individual has children with a partner who happens to be heterozygous for the variant allele, there will be a 50% chance of transmitting the disorder, depending on whether the partner contributes a variant or a wild-type allele (Fig. 7.9). Such a pedigree is said to exhibit pseudodominance (Fig. 7.10).

In autosomal recessive disorders, the difficulty lies not with risk estimation, but in determining the underlying mode of inheritance, as these disorders usually present as isolated cases with little contributory information to be gleaned from the pedigree. Carrier risks to other relatives can be calculated from the pedigree, and carrier testing may be appropriate for disorders with a high variant allele frequency or when consanguineous marriages are planned. The risk to members of the extended family for having an affected child will depend on their own risk calculated from the pedigree data and the population-based carrier risk appropriate to their partner. Higher risks in consanguineous partnerships are dependent on the degree of relationship between the partners. In disorders of unknown genetic etiology, consanguinity suggests, but does not prove, an autosomal recessive mode of inheritance.

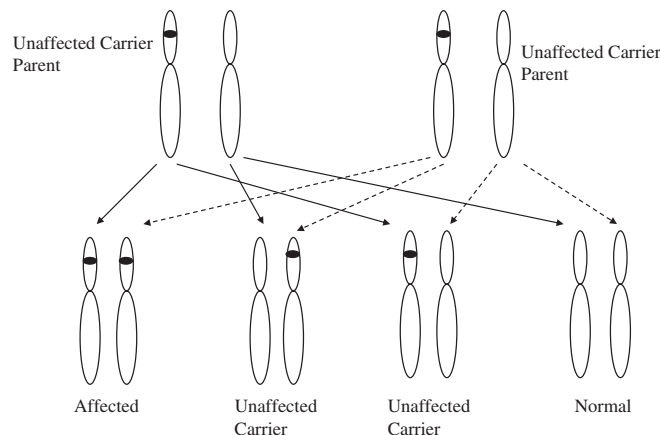


Figure 7.7 Recurrence risks in autosomal recessive inheritance.

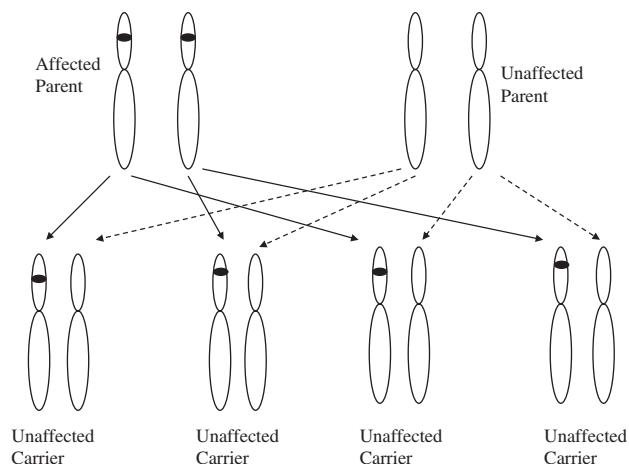


Figure 7.8 Recurrence risks for an individual with an autosomal recessive disorder and a normal partner.

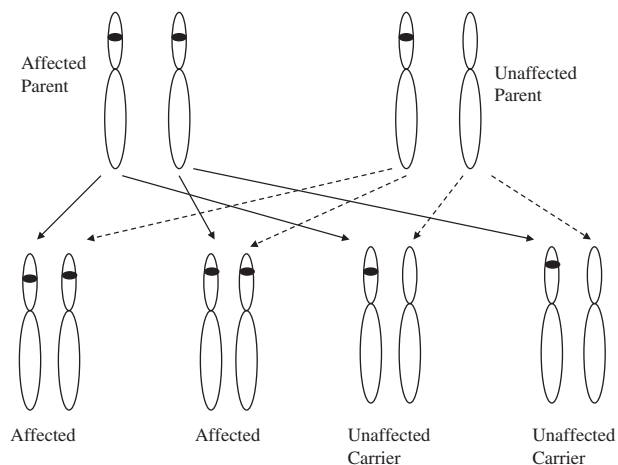


Figure 7.9 Recurrence risks for an individual with an autosomal recessive disorder and a carrier partner.

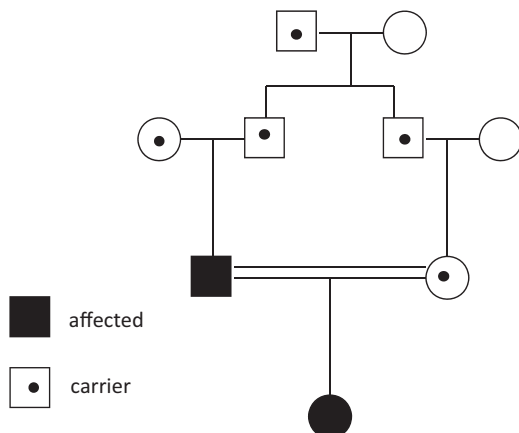


Figure 7.10 Pedigree of an autosomal recessive disorder showing pseudodominance.

7.6.3 Genetic Heterogeneity

It is not unusual in some recessive disorders, such as sensorineural deafness, for two affected individuals to have children. Assortative mating occurs in such instances because of social circumstances in which individuals with the same disability, such as deafness or visual impairment, are often educated together or share the same social facilities. If their disorder were due to a variant in the same autosomal recessive gene, all their offspring would be affected. In a number of studies involving the offspring of parents with inherited sensorineural deafness, however, a significant proportion of such unions led to offspring with normal hearing. Although in some instances this could be due to other causes (e.g., acquired causes being mistaken for

inherited deafness), in most instances the gene causing the deafness in the two parents is different, a phenomenon known as genetic heterogeneity. Each parent will transmit the variant allele for their own deafness, but a wild-type allele of the gene involved in their partner's deafness. Therefore, the child is heterozygous for the two variant alleles, referred to as double heterozygosity. This type of genetic heterogeneity involving different genes is known as locus heterogeneity.

Different modes of inheritance have also been documented for a number of clinically defined disorders with similar phenotypes. For example, autosomal dominant, autosomal recessive, and X-linked recessive inheritance have all been documented in hereditary spastic paraplegia, hereditary motor and sensory neuropathy, and retinitis pigmentosa, depending on the causative gene. Without information from genetic testing, locus heterogeneity makes it difficult to determine the risks of recurrence for phenotypes that follow both dominant and recessive inheritance unless the mode of inheritance is clearly defined by the family pedigree. Next-generation sequencing and the use of gene panels have revolutionized testing in these circumstances. It is now possible to analyze very large numbers of genes causing a particular phenotype in a single test, enabling accurate risk analysis.

Heterogeneity can also exist at the same locus; thus, an individual affected with a recessively inherited disorder can have two different variants in the two alleles of the gene and is often called a compound heterozygote. Most individuals with recessive disorders are compound heterozygotes unless a specific variant is especially prevalent in a particular population or the affected individual is the offspring of a consanguineous relationship, in which case the allele is likely to be identical by descent. This is known as allelic or mutational heterogeneity. The specific variants that an affected individual possesses can, in fact, determine the severity of the disorder, as in cystic fibrosis, in which individuals who are homozygous for the most common variant in the cystic fibrosis gene, phe508del, have a higher incidence of pancreatic insufficiency. As the variants underlying disease are identified, it is becoming increasingly apparent that in many cases the exact nature of the variant will determine the phenotype—a phenomenon known as genotype–phenotype correlation.

In some cases different variants in a particular gene may act in a dominant or in a recessive manner. For example, a wide variety of phenotypes and inheritance patterns have been associated with variants in the Lamin

A/C gene [12]. These include variants acting in an autosomal dominant fashion, causing limb girdle muscular dystrophy type 1B, dilated cardiomyopathy, familial partial lipodystrophy, and Hutchinson–Gilford progeria syndrome, as well as variants acting in an autosomal recessive fashion, such as type 2B1 axonal neuropathy, and mandibuloacral dysplasia.

7.6.4 Uniparental Disomy

Uniparental disomy (UPD) refers to the presence of both homologues of a chromosome pair or chromosomal region in a diploid offspring being derived from a single parent. If the two homologues are identical due to an error in meiosis II, this is known as uniparental isodisomy, while if the two homologues are different but still from the same parent due to a meiosis I error, this is known as uniparental heterodisomy. UPD has been reported as a rare cause of the autosomal recessive disorder cystic fibrosis in the offspring of a couple in whom only one parent was a heterozygous carrier of the variant allele. The affected offspring received both chromosome 7 homologues with the variant allele from that parent. The recurrence risk in this situation would be negligible.

7.6.5 New Variants

An autosomal recessive disorder may potentially be due to the inheritance of a variant from one parent, with a *de novo* variant occurring at the same locus on the chromosome inherited from the other parent. This is likely to be a very rare phenomenon, but accounts for some cases of spinal muscular atrophy because of a predisposition to generate deletions in the 5q13 region involving the *SMN1* gene [13]. The recurrence risk in this situation would relate to the risk of a repeated *de novo* mutation, and is therefore negligible.

7.7 SEX-LINKED INHERITANCE

Strictly speaking, sex-linked inheritance refers to the inheritance patterns shown by genes on the sex chromosomes. If the gene is on the X chromosome, it is said to show X-linked inheritance, and if on the Y chromosome, Y-linked or holandric inheritance.

7.7.1 X-Linked Recessive Inheritance

This form of inheritance is conventionally referred to as sex-linked inheritance. It refers to phenotypes due to recessive genes on the X chromosome. Males have a single X chromosome and are therefore hemizygous

for most of the alleles on the X chromosome, so that, if they have a variant allele, they will manifest the disorder. Females, on the other hand, will usually manifest the disorder only if they are homozygous for the variant allele, and if heterozygous will usually be unaffected. Since it is rare for females to be homozygous for a variant allele, X-linked recessive disorders usually affect only males.

The necessary characteristics to be certain a disorder is inherited in an X-linked recessive manner are listed in Table 7.3 and portrayed in Fig. 7.11.

7.7.1.1 Recurrence Risks

If a male affected with an X-linked recessive disorder survives to reproduce, he will always transmit his X chromosome with the variant allele to his daughters, who will be obligate carriers. An affected male will always transmit his Y chromosome to a son, and therefore none of his sons will be affected (Fig. 7.12). A carrier female has one X chromosome with the wild-type allele and one X chromosome with the variant allele;

therefore, her sons have a 1 in 2 (50%) chance of being affected, while her daughters have a 1 in 2 (50%) chance of being carriers (Fig. 7.13).

For a female to be affected with an X-linked recessive disorder, her mother would have to be a carrier and her father affected with the disorder. Obviously this situation is encountered only very rarely. Another possibility is that her mother is a carrier and the X chromosome transmitted by her father undergoes a new mutation. A female can also be affected by an X-linked recessive disorder if she has a single X chromosome (i.e., Turner syndrome), in which case she will be hemizygous for alleles on the X chromosome, like a male.

7.7.1.2 X Inactivation

Early in embryonic development, one of the X chromosomes in females is inactivated in each cell, with the result that the female, like the male, has a single functional X chromosome. This phenomenon was first described by Mary Lyon in 1961 and is sometimes referred to as lyonization. In individuals with X-chromosome aneuploidies, all but one of the X chromosomes are inactivated in each cell. The process of X inactivation is controlled by a region of the X chromosome, the X-inactivation center, situated on the proximal portion of the long arm. X inactivation usually occurs as a random process, such that in approximately 50% of the cells in a female the maternally derived X chromosome is active, and in the other 50% the paternally derived X chromosome is active. In females who are carriers of an X-linked recessive variant, approximately one-half of the cells will actively express the variant allele. Occasionally, this can

TABLE 7.3 Characteristics of an X-Linked Recessive Inherited Disorder
Males are affected almost exclusively
Transmission occurs through unaffected or carrier females to their sons
Male-to-male transmission is not observed
Affected males are at risk of transmitting the disorder to their grandsons through their obligate carrier daughters

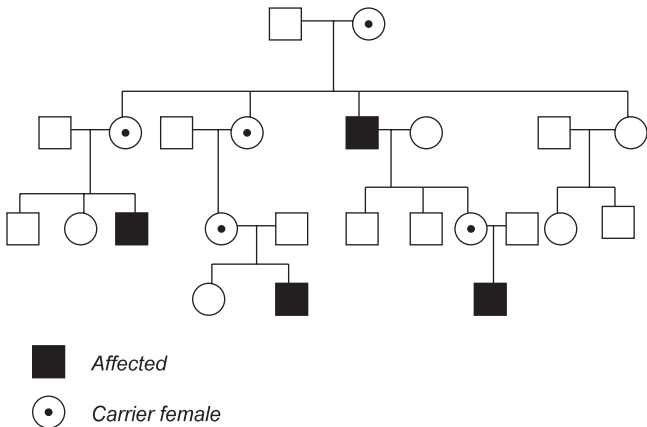


Figure 7.11 Pedigree consistent with X-linked recessive inheritance.

be demonstrated clinically. In X-linked retinitis pigmentosa, for example, careful fundoscopic examination of a female carrier can show a mosaic pattern of pigmentation.

7.7.1.3 Manifesting Female Carriers of X-Linked Recessive Disorders

Although female carriers of X-linked recessive disorders are usually asymptomatic, they can manifest signs of the disorder. They are usually much less severely affected than males, however. There are a number of different mechanisms by which a female heterozygote can manifest signs of an X-linked recessive disorder, but the underlying cause for each one is nonrandom or skewed X inactivation. In this situation, there is a departure

from the normal random process of X inactivation, with a greater proportion of one X chromosome being inactivated than the other. If, in most cells, the active X chromosome is the one with the variant allele, a female may manifest the disorder.

7.7.1.3.1 Mechanisms of nonrandom X inactivation. A number of mechanisms can lead to nonrandom X inactivation:

1. *Chance:* Skewed X inactivation can occur by chance.
2. *Monozygotic twinning:* There have been several reports of monozygotic female twins, both heterozygous for a dystrophin gene deletion, one of whom was a manifesting carrier for Duchenne muscular dystrophy and the other an unaffected carrier. In some cases it has been demonstrated that, in the

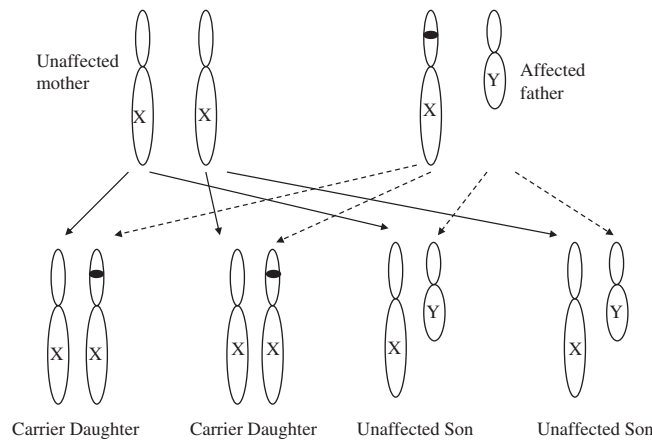


Figure 7.12 Recurrence risks for a male with an X-linked recessive disorder.

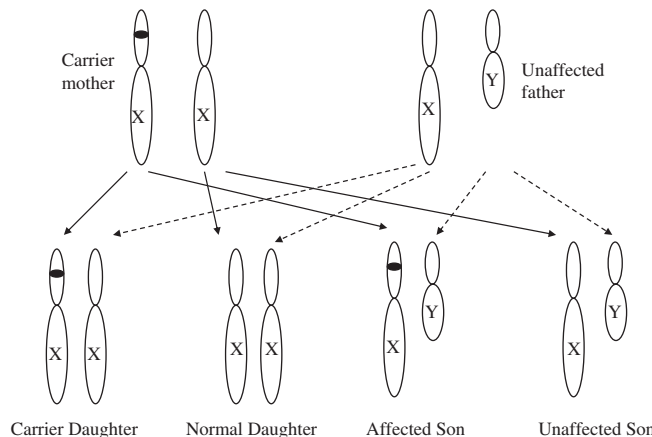


Figure 7.13 Recurrence risks for a female carrier of an X-linked recessive disorder.

lymphocytes and fibroblasts of the affected twin, the majority of active X chromosomes had the deletion, while in the unaffected twin, most active X chromosomes possessed the intact gene or there was random X inactivation. A number of hypotheses have been suggested to explain this observation [14]. One hypothesis is that random X inactivation produces two clusters of cells in the initial cell mass with opposite X-inactivation patterns, which stimulate the monozygotic twinning event. Another hypothesis is that the twinning event leads to unequal allocation of cells, leading to catch-up growth and skewed X inactivation in the twin with fewer cells

- 3. *Cytogenetic abnormalities*: In females with an X-autosome translocation, the normal X chromosome will be preferentially inactivated, maintaining the diploid state for the autosome involved in the translocation. If the translocation disrupts or interferes with the expression of a gene on the X chromosome, or if the X chromosome involved in the translocation carries a variant allele, that female will manifest the disorder. The finding of X-autosome translocations in manifesting female carriers of Duchenne muscular dystrophy was instrumental in the mapping and cloning of the dystrophin gene.
- 4. *Elimination of cells expressing the variant allele*: If a gene on the X chromosome is required for cell survival, the normal gene will always be found on the active X chromosome and the defective gene on the inactive X chromosome in the mature cell population, even though X inactivation occurred as a random process. This has been demonstrated in a number of disorders, including X-linked severe combined immunodeficiency, and can be used in the determination of carrier status for females at risk for that condition, although carrier status is now more easily identified by direct molecular testing for the variant.

7.7.1.4 Gonadal Mosaicism

As with autosomal dominant disorders, gonadal mosaicism is an important phenomenon in X-linked recessive disorders, occurring particularly frequently in Duchenne muscular dystrophy, in which it has been shown to occur in both male and female gametogenesis. It is important to take this into account when advising mothers of apparently sporadically affected males with Duchenne muscular dystrophy of recurrence risks.

Even though the mother does not carry the variant in lymphocytes, the risk to another son who inherits the same X as his affected brother can be up to 20% due to gonadal mosaicism.

7.7.1.5 New Variants

Variants arising during meiosis may occur in male or female germ cells. If a particular variant arises largely in male germ cells, as in Lesch-Nyhan syndrome [15] and hemophilia A [16], the majority of mothers of affected boys will be carriers, and risks to sisters will be 50% regardless of how many unaffected brothers they have. In Duchenne muscular dystrophy, the overall mutation rate appears to be equal in males and females, but it has been suggested that most single-nucleotide variants occur in spermatogenesis, while most deletions arise in oogenesis [17]. Therefore, in isolated cases, the recurrence risk might be higher in nondeletion cases, as these mothers would be more likely to be carriers.

7.7.2 X-Linked Dominant Inheritance

X-linked dominant inheritance is a relatively less frequent form of inheritance and is caused by dominant alleles on the X chromosome. The phenotype will manifest in both hemizygous males and heterozygous females. Random X inactivation usually means that females are less likely to be severely affected than hemizygous males, unless they are homozygous for the variant allele. The characteristics of an X-linked dominant inherited disorder are listed in Table 7.4 and depicted in Fig. 7.14.

7.7.2.1 Recurrence Risks

The offspring of either sex have a 1 in 2 (50%) chance of inheriting the disorder from affected females (Fig. 7.15). The situation is different for males affected by X-linked dominant disorders, whose daughters will always inherit the gene and whose sons cannot inherit

TABLE 7.4 Characteristics of an X-Linked Dominant Inherited Disorder
Daughters of affected males always inherit the disorder
Sons of affected males never inherit the disorder
Affected females can transmit the disorder to offspring of both sexes
An excess of affected females exists in pedigrees for the disorder

the gene (Fig. 7.16). An example of an X-linked dominant disorder is vitamin D-resistant rickets. X inactivation results in females with this disorder having less severe skeletal changes than those that occur in affected males.

An exception to the rule that females are less severely affected than males is seen in craniofrontonasal dysplasia, in which heterozygous females have a coronal craniosynostosis and affected hemizygous males do not have a craniosynostosis. The condition is caused by variants in the ephrin B1 gene *EFNB1*, and it has been proposed that in heterozygous females, patchwork loss of ephrin B1 disturbs tissue boundary formation at the developing coronal suture, whereas in males deficient in ephrin B1, an alternative mechanism maintains the normal boundary [18].

7.7.2.2 X-Linked Dominant Lethal Alleles

In some disorders due to variant alleles of genes on the X chromosome, affected males are never or very rarely seen (e.g., incontinentia pigmenti and Goltz syndrome). This is thought to be due to a lethal effect of the variant allele in the hemizygous male, resulting in nonviability of the conceptus during early embryonic development. As a consequence, if an affected female were to have children, one would expect a sex ratio of 2:1, female to male, in the offspring and that one-half of the females would be affected, while none of the male offspring would be affected (Fig. 7.17). The majority of the mothers of females with these X-linked dominant lethal disorders are unaffected, and the variant alleles are therefore thought to arise as new mutations.

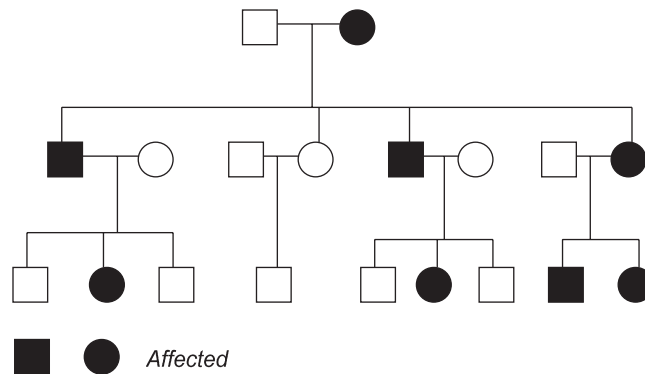


Figure 7.14 Pedigree consistent with X-linked dominant inheritance.

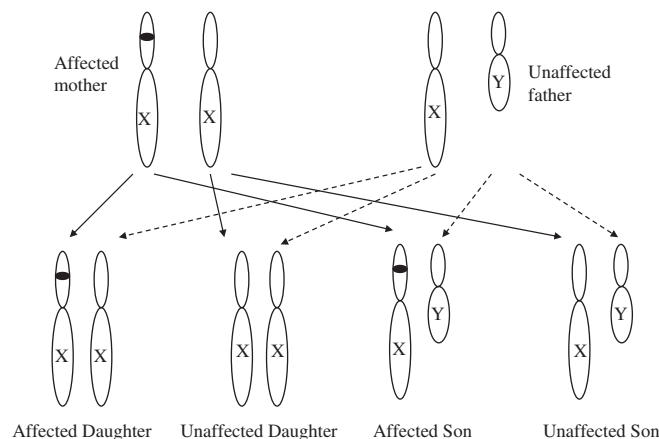


Figure 7.15 Recurrence risks for a female affected with an X-linked dominant disorder.

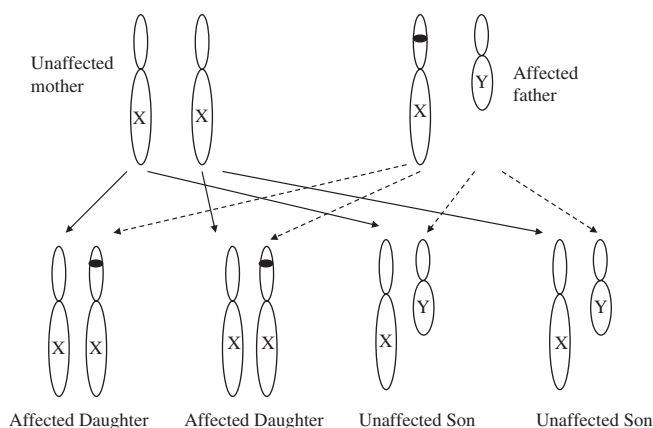


Figure 7.16 Recurrence risks for a male affected with an X-linked dominant disorder.

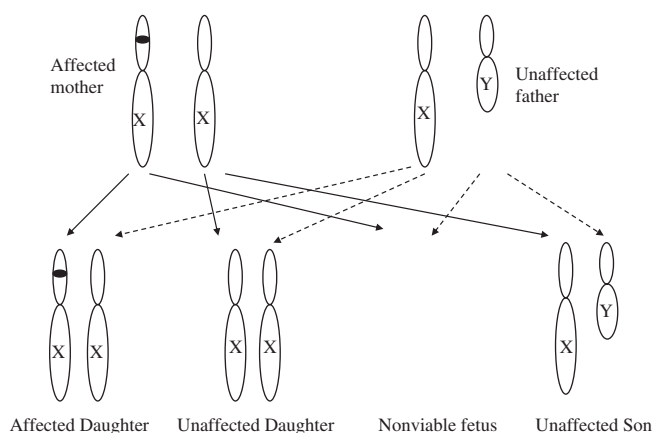


Figure 7.17 Recurrence risks for a female affected with an X-linked dominant disorder lethal in males.

Rett syndrome is an X-linked dominant condition caused by variants in the *MECP2* gene. Girls with classical Rett syndrome have severe mental retardation developing after a period of relatively normal development. The variants that cause classical Rett syndrome are lethal in males; however, other variants in the gene can give a pattern of neuroencephalopathy and profound intellectual impairment in males that behaves as an X-linked recessive condition—another example of genotype–phenotype correlation.

7.7.3 Y-Linked (Holandric) Inheritance

Y-linked, or holandric, inheritance refers to genes carried on the Y chromosome. They therefore will be present only in males, and the disorder would be passed on to all their sons but never their daughters (Figs. 7.18

and 7.19). Genes involved in spermatogenesis have been mapped to the Y chromosome, but a male with a variant in a Y-linked gene involved in spermatogenesis would probably be infertile or hypofertile, making it difficult to demonstrate Y-linked inheritance. This situation may well change with the use of techniques such as intracytoplasmic sperm injection to treat male infertility, which will result in the transmission of the infertility to male offspring.

7.8 PARTIAL SEX LINKAGE

A small region of sequence identity exists between the X and the Y chromosomes, located at the tips of the long and short arms, known as the pseudoautosomal regions of the sex chromosomes. A high rate of recombination

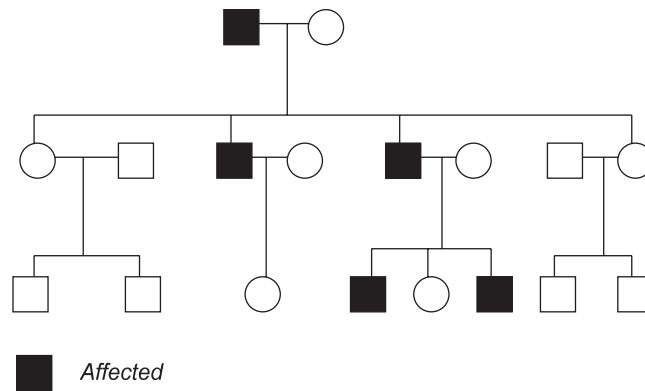


Figure 7.18 Pedigree consistent with Y-linked inheritance.

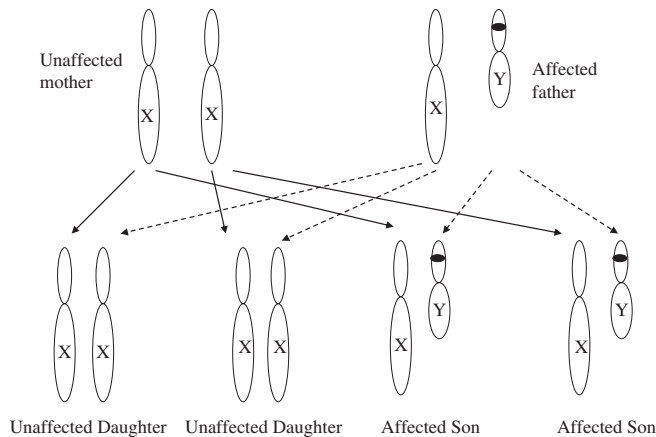


Figure 7.19 Recurrence risks for a male affected with a Y-linked disorder.

at the telomeres of the short arms is thought to be obligatory for normal meiosis of these chromosomes. The genes within these regions, known as pseudoautosomal genes, escape X inactivation in the female; therefore, both sexes have two active alleles at these loci. The pseudoautosomal gene *SHOX* has been postulated to account for some of the features seen in the numerical sex chromosome disorder Turner syndrome. As a result of the high recombination frequency, variant genes located within the pseudoautosomal region can be transferred from the Y chromosome to the X chromosome and vice versa. Haploinsufficiency of the *SHOX* gene is the cause of Leri–Weill dyschondrosteosis and, in one study, about half of the segregations investigated showed a transfer of the *SHOX* variant to the alternate sex chromosome. Therefore, the condition can be inherited as either an X-linked dominant condition or, more rarely, a Y-linked

condition. Affected men can transmit the variant to a son as well as to a daughter [19].

7.9 NONTRADITIONAL INHERITANCE

New techniques in molecular genetics are starting to reveal mechanisms that result in unexpected patterns of inheritance.

7.9.1 Genomic Imprinting and Epigenetic Mechanisms

Genomic imprinting is a phenomenon in which gene expression depends on parental origin. Imprinting is due to an epigenetic mechanism, in which the primary DNA sequence of the gene is not changed, but transcriptional regulation is affected by mechanisms such as DNA methylation, histone acetylation, and histone

methylation. The imprint is erased during the early development of the male and female germ cells and reset prior to germ cell maturation so, for example, the imprint of a paternally imprinted gene is reset during oogenesis to a maternal imprint. In contrast to the biallelic expression of most genes, imprinted genes demonstrate monoallelic expression. This process modifies the transmission and expression of certain genetic diseases and should be borne in mind as a possible mechanism underlying disorders that do not follow typical Mendelian inheritance.

Many disorders due to defects affecting imprinted genes arise *de novo*, but they can also be familial. For example, the familial paraganglioma syndrome is due to variants in the succinate dehydrogenase subunits B, C, and D. *SDHD* is an imprinted gene, with the paternal allele active in each cell and the maternal allele inactive. The pathogenic variant is dominant and an affected parent (male or female) has a 1 in 2 chance of passing it on to each child but the child is at increased risk to develop paragangliomas only if the variant is inherited from the father (Fig. 7.20). If the variant is inherited from the mother, the risk to develop paraganglioma is negligible but, if her son inherits it and passes it on, his children are at increased risk.

Prader–Willi syndrome (PWS) and Angelman syndrome are probably the best-known syndromic examples of disorders due to imprinted genes. These disorders affect different genes in the imprinted region at 15q11–q13. The first finding was that deletions within the region on the paternally derived chromosome led to

PWS, while deletions of the maternally derived chromosome led to Angelman syndrome. This is because the genes in this region that cause PWS are active on the paternal allele only and the gene that causes Angelman syndrome, *UBE3A*, is active only on the maternal allele. The underlying mechanism is often a *de novo* deletion or single-nucleotide variant within the imprinted gene itself, but in some cases disorders are due to epigenetic mechanisms such as UPD or the consequence of an imprinting center defect. Maternal UPD results in PWS and paternal UPD in Angelman syndrome. Some familial cases of Angelman syndrome are due to variants in the gene *UBE3A*. A pathogenic variant inherited from a mother will cause Angelman syndrome but, when inherited from the father, the condition will not manifest. The imprint is reset during male and female gametogenesis. Therefore, in familial Angelman syndrome, if an unaffected female carries a variant in *UBE3A* on her paternal allele it is reset at gametogenesis, so that it is now on the active allele and children who inherit it will have Angelman syndrome.

Some genes show tissue-specific imprinting, for example, the *GNAS* gene, which is associated with Albright hereditary osteodystrophy (AHO). AHO is due to G(s) α inactivating variants, imprinted in a tissue-specific manner, with expression in the proximal renal tubules, thyroid, pituitary, and ovaries being from the maternal allele. Maternally inherited variants lead to AHO with endocrine involvement (pseudohypoparathyroidism type 1A), whereas paternally inherited variants lead to AHO alone. Pseudohypoparathyroidism

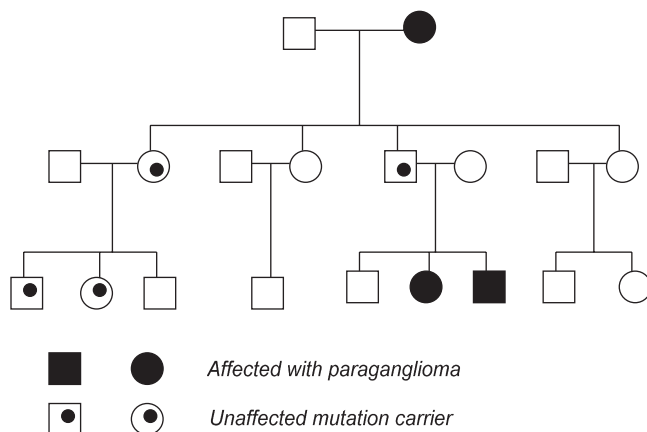


Figure 7.20 Pedigree showing familial paragangliomas and the effects of imprinting of the *SDHD* gene.

type 1B (parathormone resistance without AHO) can be caused by a deletion of the imprinted promoter regions of the gene that control gene expression [20].

7.9.2 Digenic Inheritance

Genetic or locus heterogeneity by which different genes can cause clinically identical disorders has been discussed previously in this chapter. However, these cases are considered to be monogenic in that, in any one family, only one locus is thought to harbor pathogenic variants. Digenic inheritance is the simplest form of inheritance in complex genetic disease and occurs when the phenotype is due to the coinheritance of variants in two distinct genes, i.e., the affected individuals are double heterozygotes [21].

Retinitis pigmentosa was the first inherited condition in which digenic inheritance was convincingly described [22]. Although the families described were initially thought to be compatible with autosomal dominant retinitis pigmentosa with reduced penetrance, the families showed a number of unusual features:

1. In each family, the disorder originated in the offspring of unaffected individuals.
2. Affected individuals transmitted the disorder to statistically significantly fewer than 50% of their offspring.

On molecular testing, it was found that both affected and unaffected individuals carried a variant in the *peripherin/RDS* gene. Affected individuals were also heterozygous for a variant in the *ROM1* gene. These genes encode two of the polypeptide subunits of an oligomeric transmembrane protein complex present at the photoreceptor outer segment disc rims. Abnormal *peripherin/RDS* protein can assemble with wild-type *ROM1* protein to form structurally normal complexes but cannot assemble with abnormal *ROM1* protein. Therefore, only the combination of the two heterozygous variants is pathogenic [23].

Another example of digenic inheritance is seen in facioscapulohumeral muscular dystrophy type 2 (FSHD2). The more common FSHD1 is a dominant disorder associated with a contraction of the D4Z4 array on chromosome 4 from the normal 11–100 repeats down to 1–10 D4Z4 units, when that array is on a specific, permissive chromosome 4 haplotype. The permissive chromosome 4 haplotype contains a single-nucleotide polymorphism in the D4Z4-adjacent sequence at the distal end of the array and is found in about 50%

of the general population. Each D4Z4 unit contains a copy of the *DUX4* gene, a transcription factor gene, and contraction of the D4Z4 array on the permissive haplotype leads to chromatin relaxation and overexpression of *DUX4*. Overexpression of *DUX4* is toxic to regenerating and developing muscle cells. FSHD2 occurs in individuals who inherit a loss-of-function variant in the *SMCHD1* gene on chromosome 18 and a normal-sized D4Z4 array on a chromosome 4 haplotype permissive for *DUX4* expression. Reduction of *SMCHD1* gene expression results in D4Z4 CpG hypomethylation, presumably resulting in overexpression of *DUX4*. Both the *SMCHD1* variant and the chromosome 4 permissive haplotype need to be present for disease expression [24].

7.9.3 Mitochondrial Inheritance

The nuclear chromosomes are not the only source of coding DNA sequences within the cell. Mitochondria possess their own DNA, which, as well as coding for mitochondrial transfer RNA and ribosomal RNA, also carries the genes for 13 structural proteins that are all mitochondrial enzyme subunits. Variants within these genes have been shown to cause disease (e.g., Leber hereditary optic neuropathy). The inheritance pattern of mitochondrial DNA (mtDNA) is, however, very different from that of nuclear DNA, as mitochondria are exclusively maternally inherited. Therefore, mitochondrial variants can be transmitted only through females, although they can affect both sexes equally. Genetic assessment of mitochondrial disorders is complicated by the great variability of these disorders and a pedigree that is seldom conclusive of maternal transmission. That is because there are a number of different genetic mechanisms that can underlie mitochondrial disorders [25]. These include:

1. variants of the mtDNA, including single-nucleotide variants, deletions, and duplications;
2. dominant variants in nuclear genes involved in mitochondrial structure, function, and mtDNA maintenance;
3. recessive variants in nuclear genes involved in mitochondrial structure, function, and mtDNA maintenance.

In Leber hereditary optic neuropathy, the pattern of maternal inheritance is well documented. The commonest mtDNA variant is a single-nucleotide variant in base pair 1178 of the ND4 gene of complex I of the respiratory chain. Two other common variants have also been

described (G3460A and T14484C), which also involve genes encoding complex I subunits of the respiratory chain. More than 95% of cases are the result of one of these three variants. Women with the variant will transmit it to all offspring, but only around 1 in 2 males and 1 in 10 females with the variant develop loss of vision. Affected and carrier males do not transmit the variant to their offspring. An example of a pedigree for a family with Leber's is shown in Fig. 7.21.

A single cell contains many copies of the mitochondrial genome, and heteroplasmy for a mitochondrial variant is usual. Heteroplasmy is the presence in the cell of both wild-type and variant copies of a gene. When all the mtDNA carries the variant it is known as homoplasmy. The proportion of variant mtDNA in leukocytes is not a reliable indicator of which individuals will develop symptoms. As with other mitochondrial disorders, this presents problems in counseling asymptomatic individuals known to carry the variant. The recurrence risks for mitochondrial disorders involving mtDNA are difficult to determine. Large-scale mtDNA deletions, as in Kearns–Sayre syndrome, are usually sporadic, while single-nucleotide variants, as in Leber hereditary optic neuropathy, are more likely to be maternally transmitted.

7.9.4 Multiple Genetic Disorders in a Family

Whole-genome and -exome sequencing has led to the understanding that the phenotype in some families is not due to a single genetic condition but to multiple genetic conditions in both individuals and families. Each condition will be inherited independently within the family, giving a complex clinical picture [26]. The

combination of disorders can include different types of genetic change, such as a single-gene disorder in combination with a chromosomal copy number variant.

7.10 CHROMOSOMAL DISORDERS

Factors that indicate the possibility of a familial chromosomal disorder in a pedigree include a family history of infertility, multiple spontaneous miscarriages, and malformed stillbirths or live-born infants with multiple congenital malformations occurring in a pattern that does not conform to that of Mendelian inheritance. Fig. 7.22 illustrates a kindred in which there is segregation of a reciprocal translocation in both balanced and unbalanced forms. The possibility of a subtle rearrangement not detected by routine cytogenetic analysis but leading to recurrent chromosome imbalance in offspring should be borne in mind, and a number of techniques can be used to identify this situation, such as fluorescence in situ hybridization, comparative genomic hybridization, and microarray analysis.

It is difficult to give precise reproductive risks to a person known to carry a balanced reciprocal translocation. The risk of viable chromosome imbalance will depend on the size and origin of the unbalanced segment generated, and there is an association between the mode of ascertainment and the risk of recurrence, although this is not absolute. Ascertainment through the live birth of an abnormal baby is associated with a higher risk than ascertainment through miscarriage or infertility. Information from an individual pedigree is seldom sufficient to derive a family-specific risk based on reproductive

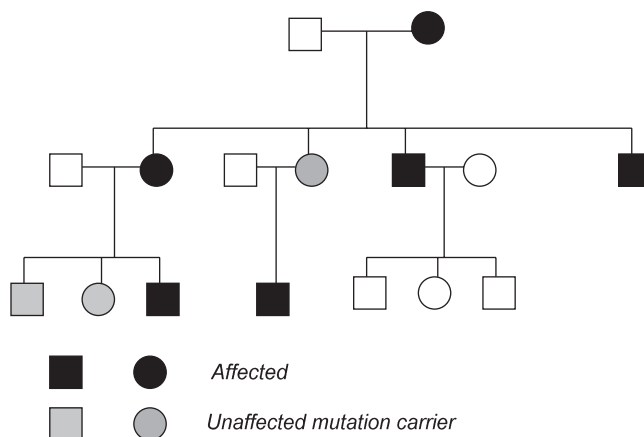


Figure 7.21 Pedigree showing Leber hereditary neuropathy due to a mitochondrial mutation.

outcomes in known carriers, and estimation of risk is mainly based on the likely segregation pattern at meiosis and the chance of an imbalance being viable.

The majority of chromosomal deletions and duplications present as sporadic cases. Although very low, the recurrence risk may be increased above the normal age-related risk. After the birth of a child with Down syndrome due to trisomy 21 to a mother under the age of 35 years, the recurrence risk is 0.5% for trisomic Down syndrome and 1% for all chromosomal abnormalities. For mothers over the age of 35 years, the birth of a child with Down syndrome does not appear to increase recurrence risk significantly compared with the population-related risks. The risk for other family members is not increased.

7.11 POLYGENIC AND MULTIFACTORIAL INHERITANCE

A large group of relatively common disorders have a considerable genetic predisposition but do not follow clear-cut patterns of inheritance within families. These include many birth defects and chronic diseases of later life. Liability to these disorders appears to be due to the inheritance of multiple gene variants, each of small effect (polygenic inheritance), or the interaction of several genetic and environmental influences (multifactorial inheritance). Clusters of cases within a family may simulate a Mendelian pattern of inheritance.

It is often difficult to be precise about recurrence risks in multifactorial disorders. The risk is greatest among

first-degree relatives, is usually small for second-degree relatives, and often does not exceed the general population risk for third-degree relatives. The risk increases when multiple family members are affected. Risks are also influenced by the incidence of the disorder in the general population and the sex of the patient and relatives in disorders that have unequal sex incidence, such as pyloric stenosis and Hirschsprung disease. In these disorders recurrence is higher in the siblings of the less affected sex, probably reflecting a greater genetic load to cause the disease in that sex. The severity of the disorder also influences risk, for example, the recurrence risk is greater for bilateral cleft lip and palate than for unilateral cleft lip alone.

In many disorders, empirical risks have been derived from family studies that provide data on observed recurrences. Risks derived in this way will be less accurate in disorders that are genetically heterogeneous and may not be applicable to populations different from those in which the family studies were performed. Nevertheless, empirical risks provide a useful basis for discussing levels of risk during genetic counseling. Risk tables for some of the common congenital malformations, such as cleft lip, with or without cleft palate, pyloric stenosis and neural tube defects have been published.

Considerable heterogeneity occurs within groups of disorders traditionally considered to be multifactorial or polygenic, such as diabetes, epilepsy, and congenital heart disease, with a proportion of cases attributable to single-gene defects. Identification of a specific genetic etiology, such as the presence of a submicroscopic deletion

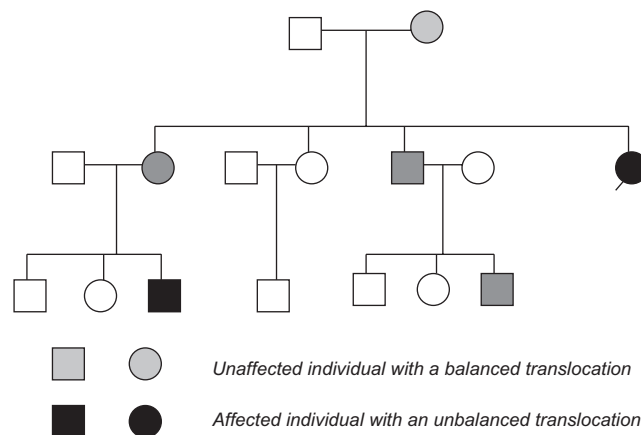


Figure 7.22 Pedigree illustrating the segregation of a balanced and unbalanced reciprocal chromosome translocation.

at chromosome 22q11 in some cases of nonsyndromic congenital heart disease, is important in providing genetic advice appropriate to a particular case. Hirschsprung disease provides a good example of a polygenic disorder in which involvement of several single loci has now been identified in a proportion of families. Data from initial family studies suggested sex-modified polygenic inheritance, and empirical recurrence risks have been produced for genetic counseling based on the sex of the index case and relatives and the length of the aganglionic segment involved; however, a mode of inheritance compatible with an incompletely penetrant autosomal dominant gene was suggested in some families. Linkage to a gene on chromosome 10 was subsequently demonstrated and variants in the *RET* oncogene were demonstrated. Variants in other genes, notably, the endothelin receptor type B gene *EDNRB*, have also been implicated. Thus in a subset of families with Hirschsprung disease, there is a major unifactorial predisposition, a situation that is seen in many other multifactorial disorders, e.g., Alzheimer and Parkinson disease.

7.12 ISOLATED CASES

In the days before widespread molecular genetic testing, isolated, presumed genetic, cases within a family posed problems in terms of genetic counseling unless the condition had a well-recognized inheritance pattern. Nowadays modern molecular genetics techniques are identifying the causes of many of these conditions, although they are still not giving the answers in all cases. Many studies of whole-exome and -genome sequencing are coming up with a similar diagnosis rate of about one-third. Therefore, there is clearly a long way to go in understanding all the mechanisms of genetic disease in humans.

We now understand that there may be many causes of an isolated case within a family. These include the following:

1. The disorder may be due to an autosomal dominant gene variant, arising by a new mutation, transmitted through a nonpenetrant or very mildly affected parent or by a clinically unaffected parent who carries a mosaic germline variant. The situation may also represent misattributed paternity. Risk to siblings varies between 0% and 50%, depending on the origin of the variant, while risk to the offspring of the affected individual would be 50%.
2. The disorder may be caused by an autosomal recessive gene variant with a 25% recurrence risk for

siblings unless due to UPD or a de novo variant. If carrier state can be confirmed by molecular or biochemical analysis, cascade screening of other family members may be appropriate when the population carrier frequency is high or consanguineous marriages are planned.

3. The disorder may be due to an X-linked gene, usually presenting in a hemizygous male but affecting females if the X-inactivation pattern is skewed or the gene acts dominantly. Isolated cases may represent de novo variants, which are frequent in lethal X-linked recessive disorders, or may be transmitted by asymptomatic mothers who are carriers or who are gonadal mosaics for the variant.
4. The disorder may be due to two or more independently inherited monogenic or chromosomal defects.
5. The disorder may be due to an mtDNA variant representing a sporadic case or maternal transmission.
6. The disorder may be due to a chromosomal abnormality. Many of these, including the common trisomies due to nondisjunction, have a low risk of recurrence, but unbalanced karyotypes due to familial chromosomal rearrangements may carry high risks of recurrence, and investigation of relatives is required.
7. The disorder may be polygenic, and recurrence risks depend on the disorder. These are based on empirical data derived from family studies.
8. The disorder may have a nongenetic etiology with no increase in the risk of recurrence, unless due to a teratogenic agent to which further pregnancies will also be exposed.

This list will increase with new technologies and new understanding. The use of whole-genome sequencing is likely to uncover a significant number of genetic conditions that are due to gene regulation rather than the gene itself. It is hoped that the future will provide more and more answers to families seeking an explanation for the genetic disease within their family.

REFERENCES

- [1] Peters J. Classic papers in genetics. London: Prentice-Hall; 1959.
- [2] Wilkie A. The molecular basis of genetic dominance. *J Med Genet Genomics* 1994;3189–98.
- [3] Veitia RA, Caburet S, Birchler JA. Mechanisms of Mendelian dominance. *Clin Genet* 2017. <https://doi.org/10.1111/cge13107>.

- [4] Sheffield WP, Bhakta V. The M358R variant of α (1)-proteinase inhibitor inhibits coagulation factor VIIA. *Biochem Biophys Res Commun* 2016;470(3):710–3.
- [5] Knudson A. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 1971. 168820–168823.
- [6] Harris P, Rossetti S. Determinants of renal disease variability in ADPKD. *Adv Chron Kidney Dis* 2010;17(2):131–9.
- [7] Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol* 2016;17:241.
- [8] Penrose L. Parental age and mutation. *Lancet* 1955;312–3.
- [9] Pyott SM, Pepin MG, Schwarze U, Yang K, Smith G, Byers P. Recurrence of perinatal lethal osteogenesis imperfecta in sibships: parsing the risk between parental mosaicism for dominant mutations and autosomal recessive inheritance. *Genet Med* 2011;13(2):125–30.
- [10] Evans DGR, Ramsden RT, Shenton A, Gokhale C, Bowers NL, Huson SM, Pichert G, Wallace A. Mosaicism in neurofibromatosis type 2: an update of risk based on uni/bilaterality of vestibular schwannoma at presentation and sensitive mutation analysis including multiple ligation-dependent probe amplification. *J Med Genet* 2007;44:424–8.
- [11] Garrod AE. The incidence of alkaptonuria: a study in chemical individuality. *Lancet* 1902:1616–20.
- [12] Rankin J, Ellard S. The laminopathies: a clinical review. *Clin Genet* 2006;70(4):261–74.
- [13] Rodrigues NR, Owen N, Talbot K. Deletions in the survival motor neuron gene on 5q13 in autosomal recessive spinal muscular atrophy. *Hum Mol Genet* 1995;4:631–4.
- [14] Machin G. Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. *Am J Med Genet* 1996;61:216–28.
- [15] Francke U, Felsenstein J, Gartler SM, Migeon BR, Dancis J, Seegmiller JK, Bakay F, Nyhan WL. The occurrence of new mutants in the X-linked recessive Lesch–Nyhan disease. *Am J Hum Genet* 1976;28:123–37.
- [16] Bröcker-Vriends AHJT, Rosendaal FR, van Houwelingen JC, Bakker E, van Ommen GJ, van de Kamp JJ, Briet E. Sex ratio of the mutation frequencies in haemophilia A: coagulation assays and RFLP analysis. *J Med Genet* 1991;28:672–80.
- [17] Lee T, Takeshima Y, Kusunoki N, Awano H, Yagi M, Matsuo M, Iijima K. Differences in Carrier frequency between mothers of Duchenne and Becker muscular dystrophy patients. *J Hum Genet* 2014;59:46–50.
- [18] Twigg SR, Kan R, Babbs C, Bochukova EG, Robertson SP, Wall SA, Morriss-Kay GM, Wilkie AO. Mutations of Ephrin-B1 (EFNB1), a marker of tissue boundary formation, cause craniofrontonasal syndrome. *Proc Natl Acad Sci U S A* 2004;101:8652–7.
- [19] Kant SG, van der Kamp HJ, Kriek M, Bakker E, Bakker B, Hoffer MJ, van Bunderen P, Losekoot M, Maas SM, Wit JM. The jumping SHOX gene-crossover in the pseudoautosomal region resulting in unusual inheritance of Leri–Weill dyschondrosteosis. *J Clin Endocrinol Metab* 2011;96(2):E356–9.
- [20] Liu J, Che, n M, Deng C, Bourc’his D, Nealon JG, Erlichman B, Bestor TH, Weinstein LS. Identification of the control region for tissue-specific imprinting of the stimulatory G protein alpha-subunit. *Proc Natl Acad Sci U S A* 2005;102(15):5513–8.
- [21] Saffer AA. Digenic inheritance in medical genetics. *J Med Genet* 2013;50:641–52.
- [22] Kajiwarra K, Berson E, Dryja T. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* 1994;264:1604–7.
- [23] Goldberg AF, Molday RS. Defective subunit assembly underlies a digenic form of retinitis pigmentosa linked to mutations in peripherin/RDS and ROM-1. *Proc Natl Acad Sci U S A* 1996;93:13726–30.
- [24] Lemmers RJLF, Tawil R, Petek LM, Balog J, Block GJ, Santen GWE, Amell AM, van der Vliet PJ, Almomani R, Straasheijm KR, Krom YD, Klooster R, Sun Y, den Dunnen JT, Helmer Q, Donlin-Smith CM, Padberg GW, van Engelen BGM, de Greef JC, Aartsma-Rus AM, Frants RR, de Visser M, Desnuelle C, Sacconi S, Filippova GN, Bakker B, Bamshad MJ, Tapscott SJ, Miller DG, van der Maarel SM. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes fascioscapulohumeral muscular dystrophy type 2. *Nat Genet* 2012;44:1370–4.
- [25] Chinnery P. Mitochondrial disease in adults: what’s old and what’s new? *EMBO Mol Med* 2015;7:1503–12.
- [26] Balci TB, Hartley T, Xi Y, Beaulieu CL, Bernier FP, Dupuis L, Horvath GA, Mendoza-Londono R, Prasad C, Richer J, Yang X-R, Armour CM, Bareke E, Fernandez BA, McMillan HJ, Lamont RE, Majewski J, Parboosingh JS, Prasad AN, Rupar CA, Schwartzentruber J, Smith AC, Tetreault M, FORGE Canada Consortium, Care4Rare Canada Consortium, Innes AM, Boycott KM. Debunking Occam’s razor: diagnosing multiple genetic disease in families by whole-exome sequencing. *Clin Genet* 2017;92:281–9.

Analysis of Genetic Linkage

Rita M. Cantor

Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, United States

8.1 INTRODUCTION TO LINKAGE ANALYSIS

Linkage analysis is a well-established genetic method used to map the genes for heritable traits to their chromosome locations. It is part of a larger process that has been referred to as “reverse genetics,” because the approach works in the reverse order from our model of how genes operate, biologically. That is, while genes act in a forward fashion to produce a trait, reverse genetics starts with the trait and uses linkage analysis along with other analytic methods to identify the predisposing genes. Reverse genetics became feasible in the 1990s, when a very extensive panel of multiallelic markers that spanned the human genome was established. Since 2008, the genome-wide markers in use have evolved from multiallelic to biallelic single-nucleotide polymorphisms (SNPs), where their spacing is much denser. The whole genome is analyzed and the approach is referred to as a full-genome linkage scan.

The genome-wide markers are genotyped and tested in a study sample of pedigrees, and those showing the strongest statistical evidence of linkage exceeding a predetermined threshold localize the trait gene to the chromosome segment where the markers reside. The resolution at the locus is usually quite poor, as many genes will reside within a linked region. Nevertheless, a statistically significant linkage result limits the search for the predisposing gene to those in the linked region, thus reducing cost and follow-up time. Once a trait gene is mapped by linkage, other strategies such as fine mapping, linkage analysis with

additional markers, targeted association analysis, and sequencing of the chromosome region can be used to identify the gene of interest. Using the advances made by the Human Genome Project, reverse genetics has been very effective in identifying genes causing rare Mendelian disorders that result from fully penetrant single genes.

Linkage analysis has two requirements. First, one must identify and study families containing individuals who exhibit a heritable trait of interest. Then, a substantial number of family members, both affected and unaffected, must be genotyped for genetically informative markers. A critical step is to establish that the trait of interest is heritable and to assess its mode of inheritance. The hallmark of a heritable trait is that it is shared to a greater degree by genetically close relatives compared with distant relatives. Comparison of trait concordance rates in monozygotic and dizygotic twin pairs is the classic approach to assess heritability, but other designs have also been used. The initial genetic markers were multiallelic single tandem repeats, which are very informative. However, these are now genotyped rarely, and less-informative biallelic SNPs genotyped on arrays have replaced them.

In a linkage analysis, the families can vary in size from nuclear families or sibships with at least two genotyped children to larger pedigrees with complicated structures in which a substantial percentage of the members are genotyped and measured for the trait. The trait can be binary, having only two values, such as the absence or presence of a disease, or quantitative, continuous, and possibly normally distributed, such as height. Statistical algorithms are applied to the family

structure, marker, and trait data to test if a trait value cosegregates with a particular marker allele within each family more often than one would expect by chance alone. The segregating marker allele may differ among families and the results are combined over all of the families. Chromosome regions where statistically greater than expected marker/trait cosegregation occurs are considered to be linked to the trait. The analytic linkage algorithms can be simple or complex and easy or difficult to apply and interpret. The design of a good linkage study involves the consideration and coordination of all of these factors.

For Mendelian traits that follow a pattern of inheritance consistent with a single gene that is X-linked, autosomal recessive, or autosomal dominant, linkage between the trait and the genetic markers is tested using a parametric linkage analysis that models the mode of inheritance of the trait in the analysis. Application of appropriate computer software results in an estimate of a recombination fraction between the trait gene and a chromosome locus. The linked region is usually relatively large, and a statistically successful linkage result is usually followed by fine mapping of additional markers in the same pedigrees within the linked region. These marker genotypes are analyzed by additional linkage or association tests to better localize the gene. For genetically complex traits, which are discussed in [Section 8.4.1](#), the initial approach is similar to the one used for Mendelian traits, but the mode of inheritance is not known and, thus, not specified.

It should be noted, however, that currently many gene identification studies are bypassing linkage analysis. Pedigrees have been reduced to single individuals, genotyping is done with very dense SNP marker panels, and trait genes are identified by testing individual SNPs using the statistical method of association analysis. Rather than testing cosegregation of marker alleles and trait values of individuals within pedigrees, the current studies capitalize on linkage disequilibrium, which is a reflection of the cosegregation of base-pair alleles over many generations due to their very close proximity.

The important underlying biological concepts and statistical methods for conducting and interpreting a linkage analysis are presented in this chapter. A very detailed discussion of the additional aspects and refinements of linkage analysis is provided in the book *Analysis of Human Genetic Linkage* by Jurg Ott [1].

8.2 LINKAGE ANALYSIS: BASIC CONCEPTS

Linkage analysis is a statistical procedure to combine data on family structures, trait values, and genetic markers. Its biological basis is the detection of recombination between markers and trait genes, and a reduced amount of recombination compared with what is expected when the marker and trait genes are segregating independently is the hallmark of linkage. Interpreting this statistically requires an understanding of likelihood functions, maximum likelihood estimation, and odds ratios. These basic concepts and their integration to produce parametric linkage analysis are discussed below.

8.2.1 Recombination: Biological Basis of Linkage Analysis

Linkage analysis is based on the biological phenomenon of genetic recombination, which occurs in the parental gametes during the process of meiosis before the eggs and sperm are produced. In a parental gamete, when a pair of chromosomes, one from each grandparent, aligns in the first metaphase, an exchange of chromosomal material often occurs via a crossover event, with the crossover location thought to be determined by chance. This recombination of genetic material results in chromosomes different from those that would be inherited from either parent alone. Thus, each child inherits a unique set of chromosomes that are recombinants of the grandparents'. Linkage analysis is based on identifying recombination events between genetic markers and trait loci and inferring whether a trait and marker alleles are traveling in close proximity on the same chromosome or are farther away or on different chromosomes. The fundamental principle of linkage analysis is that for any two loci on the same chromosome, the closer they are to each other, the less likely it is that they will undergo recombination. Linked genes are those located close enough to each other on a chromosome that an expected crossover rate within the genetic material separating them at meiosis is less than 50%. Although recombination rates are not uniform across the genome, this principle has provided an effective biological model for linkage analysis.

8.2.2 Linkage Analysis Simplified: Inbred Mouse Strains

Families segregating a trait of interest are essential for linkage analysis. The patterns of allele frequencies in

human genes and their cosegregation derive from millennia of nonexperimental mating, resulting in our current human population. Thus, unlike experiments with inbred mouse strains, all genetic studies in humans can only be observational. However, to illustrate some basic concepts used in linkage analysis, we first discuss the approach in inbred strains where the ideas are very straightforward. How these principles are applied to analyze linkage in human pedigrees is discussed in greater detail later in the chapter.

In experimental species, controlled crosses can be optimally designed to investigate recombination between loci. For two biallelic markers A,a and B,b , homozygous parental inbred strains can be crossed, where a phase-known genotype of AB/AB is crossed with a phase-known genotype of ab/ab , yielding a first generation (F1) with phase-known AB/ab individuals in which the alleles inherited from each parent are located on distinct chromosomes divided by the line “/” in our notation here. F1 individuals can be crossed back to either parental line, say, $AB/ab \times AB/AB$ (backcross), or they can be crossed among themselves, $AB/ab \times AB/ab$ (intercross). Because the chromosomal origin of each allele is unambiguous, recombination events can be directly identified in each offspring by genotyping markers. Consequently, estimation of the recombination fraction, r , the observed number of crossovers between the two loci divided by the possible number of crossovers in that interval, is very simple to estimate by counting. If the estimate of r is close to zero, we can infer that the two markers are very close together; while if it is 50%, we can infer that the two markers are either very far apart on the same chromosome or on different chromosomes. This is because the expected crossover on the same chromosome or reshuffling of chromosomes during meiosis results in a 50% chance of seeing the parental genotype combination in the child. Linkage is inferred when r is significantly less than 50%.

The option of counting recombinants is usually not possible in humans. A situation analogous to a murine backcross consists of a nuclear family in which the observed parental genotypes for markers at two loci are $AaBb \times aabb$. Note, however, that the phase of these genotypes is not known for the double heterozygote: that parent may be either AB/ab or Ab/aB . This precludes simple counting as a method for estimating r . Historically, these two possible phases

have been called *coupling* AB/ab and *repulsion* Ab/aB . The possibilities of coupling and repulsion are modeled by the statistical methods of parametric linkage analysis. Since there is a 50% chance that the gametes are in coupling and a 50% chance they are in repulsion, the evidence for linkage is examined under both assumptions and the evidence under the two possibilities is combined.

8.2.3 Parametric Linkage Analysis: Statistical Concepts

The parametric method to sequentially test for linkage in pedigrees was adapted and applied by Newton Morton in his 1955 paper [2]. It is a procedure in which r is the parameter of interest, and each pedigree is analyzed separately, using the same analytic algorithm separately at fixed values of r . The results are combined sequentially over the families that are tested until a decision regarding linkage is reached. Using Morton's sequential linkage approach, linkage is inferred if the evidence from the tested families results in a test statistic, referred to as the LOD score, that exceeds 3.0 and is ruled out if the LOD score falls below -2.0 . These values are equivalent to the odds of 1000:1 for linkage or 100:1 against linkage, respectively. They are derived from the prior probability of linkage to a region and the multiple tests that are being conducted. Although many new approaches, algorithms, and extensions to the algorithm have been developed since that time, the method remains in use. Since many of the important concepts and approaches to linkage analysis were developed in relation to this algorithm, those concepts are presented in the context of the algorithm to detect parametric linkage.

8.2.3.1 Likelihoods, Maximum Likelihood Estimation, and Statistical Significance

We begin by defining the concept of likelihood in the context of linkage analysis. In general, if H is a hypothesis (e.g., two loci are linked with a recombination fraction of r) and D is the data collected to test the hypothesis (marker and trait information in families), statistical theory tells us that the likelihood of this hypothesis, $L(H)$, is proportional to the probability of observing the data when we assume that the hypothesis is true, $p(D|H)$. Constructing a likelihood $L(D|H)$ requires representing in symbols a model of how the data (trait phenotypes and marker genotypes generated

in the pedigrees) are expected to cosegregate in the families. The principle of maximum likelihood states that the hypothesis or model with greatest value for the likelihood is that for which the probability of the data that are observed is maximized. For linkage, we assess the value of r giving the maximum value for the likelihood for the available data. That is, this maximum is obtained by finding that value of r for which the probability of the experimental data in the model containing r , $p(D, r)$, is the largest. Originally the likelihoods were calculated for specific values of r , which are 0.01, 0.05, 0.10, 0.20, 0.30, and 0.40. With the advent of efficient computer programs, the maximum likelihood estimate of r is obtained currently using a numerical approach, where an algorithm or procedure is used to analytically climb the surface of the likelihood function until a maximum is reached. Thus, r may be estimated at a value that does not equal any of the aforementioned fixed values.

As with all estimates derived from a limited sample of experimental data, a significant sampling error is associated with r , and linkage cannot be inferred from this value alone, even if it is less than 0.5. Rather, a statistical test that contrasts the likelihoods of linkage and independent segregation of the marker and the trait is conducted. The likelihood of the latter hypothesis is proportional to the probability of the observations when r is 0.5, which is the null hypothesis of no linkage. The ratio of the likelihoods reflects the odds in favor of linkage at the maximum likelihood estimate of r . Its log is referred to as the LOD score, discussed below.

8.2.3.2 LOD Scores

The LOD score represents the logarithm in base 10 of the odds of linkage of a trait gene at a recombination fraction r with a particular marker locus compared with a recombination fraction of 0.5 between the marker and the trait gene. The term LOD is derived from the first letters in the log of the odds and the method it represents provides a different way of assessing the significance of the linkage signal, other than a p value. The LOD score approach was applied by Newton Morton to the linkage problem in his paper presenting the method in 1955, and is from the field of sequential methods in statistics, where all the evidence is not assessed in one large analysis, but data are collected and used sequentially to reach a decision. This reflects an efficient and economical method for studying

families—once a decision is reached it is not necessary to study additional families. In symbols, the LOD score for linkage takes the form, $\log[L(D, r)/L(D, 0.5)]$, where D is the marker and trait data in the families and r is the recombination fraction. We infer that there is linkage when the LOD score exceeds 3.0, a value that was calculated by Morton by taking into account the fact that any two loci in the whole genome have a certain prior probability of being on the same chromosome. Other criteria have been proposed for more densely spaced markers, and Nyholt [3] has addressed the issue of significance levels for different statistical methods of linkage analyses more fully. Together with the maximum likelihood estimate of the recombination frequency and its associated LOD score, it remains customary to report a LOD score table, where the LOD score is computed for a set of predetermined recombination fractions of 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, and 0.40. Using this convention, the results from independent studies can be combined at these particular recombination fractions without reanalysis of the combined set of families. Once the LOD score exceeds 3.0 at any recombination fraction, linkage is declared, and the distance between the trait gene and the marker is inferred to be that recombination fraction among the ones previously listed exhibiting the largest LOD score. If r is 0.05 at the highest LOD score, it means that the trait and the marker are 5 centimorgans (cM) apart. A Morgan is defined as the distance along a chromosome in which exactly one recombination is expected. A centimorgan is 0.01 of a Morgan. A rough rule is that the map distance of a centimorgan is equivalent to 1 million base pairs in physical distance. Thus, the recombination fraction of 5 represents a distance of 5 million base pairs. There are likely to be many genes in a region of this size, and follow-up work is required to identify the specific gene involved.

8.2.3.3 Modeling Traits with Penetrance Functions

Computer software for parametric linkage analyses requires that the mode of inheritance of the trait of interest be specified in terms of the penetrances of the genotypes at the “causal” gene. To do this, it is usually assumed that the trait gene is biallelic. Penetrance is the probability that the trait genotype will lead to the trait phenotype, and the penetrance values vary between 0.0 and 1.0. If we assume that the two alleles are d and D , the inheritance pattern is modeled by setting the values of

the penetrances of the three possible genotypes, DD , Dd , and dd . In a simple model of a dominant binary trait, the penetrances of DD and Dd are each 1.0 in those with the trait. That is, those with one or two copies of allele D will surely exhibit the trait. Since the dd genotype will never lead to the trait, its penetrance is 0.0. The penetrance values for those without the trait are 0.0, 0.0, and 1.0 for DD , Dd , and dd , respectively. That is, only those individuals with the dd genotype will develop the trait. For a recessive disorder, where D is again the trait-predisposing allele, the penetrances for DD , Dd , and dd in those with the trait are 1.0, 0.0, and 0.0, respectively, and penetrances in those without it are 0.0, 1.0, and 1.0, respectively. These very simple models ignore the possibility of alternative trait-predisposing genes, phenocopies that do not have a genetic basis for the phenotype but exhibit it anyway, and reduced penetrance whereby the predisposing genotypes result in the trait only some of the time. These options are discussed in greater detail in [Sections 8.3.1 and 8.3.2](#).

8.2.3.4 Designing and Conducting Parametric Linkage Analyses

The purpose of linkage analysis is to accrue statistical evidence regarding the cosegregation of a trait and marker alleles within families. It should be noted that the trait can cosegregate with a different marker allele in each family and linkage would be established. To select families, each should have some members with the trait of interest. Mendelian traits that are relatively rare and have a known mode of inheritance are investigated in extended families. Ascertainment of large pedigrees is usually the only way an adequate number of affected individuals can be obtained to provide the linkage analysis with sufficient statistical power. Linkage analysis of large pedigrees is usually performed with the assistance of computer programs such as LINKAGE and SAGE/LODPAL. [Tables 8.1 and 8.2](#) give some of the most commonly used computer programs that are used to test for parametric linkage. Websites for the computer programs discussed here are provided in the Bibliography. The LINKAGE software is appropriate for the analysis of binary traits in large pedigrees. Other programs, such as Option 2 of MENDEL, will conduct the same analysis, although the format of the input files and output reports of the programs may differ.

Knowledge about the mode of inheritance should have the greatest influence on the selection of families.

Families that include individuals who are inbred with many homozygotes provide the most powerful sample for recessive traits, while multigenerational families provide the best power for dominant traits. It is also important to recognize that informative families are those that are segregating the trait in all branches, so that the ability to detect the trait in each branch of the family is an important first step before collecting and genotyping DNA. However, bilateral families, in which both sides of the family have the trait, will confound a linkage analysis. It is also important to recognize that a single individual in a sibship where phase is not known through the grandparents will not provide information regarding linkage.

Markers and methods of genotyping have evolved. Multiallelic markers are the most informative; however, these are rarely genotyped for current studies because of cost and the lack of labs typing these markers. Instead labs using SNP genotype arrays are much more plentiful and the genotypes are much less costly. SNPs are biallelic, and therefore less informative; however, conducting linkage analysis on a very dense set of SNPs is likely to reflect the presence of a causal allele(s) through linkage disequilibrium within the pedigrees.

The steps to conduct a linkage analysis of a binary trait using the LODPAL computer program are clearly delineated in a chapter by Cantor [\[4\]](#).

8.3 EXTENDING PARAMETRIC LINKAGE ANALYSIS

Since the methods for linkage have been formalized, computer programs have been extended to make linkage analyses more precise and powerful. The methods to test linkage for a Mendelian binary trait have been extended to model the inheritance of traits that result from a single gene for which the genotype penetrance values are neither 0 nor 1. This occurs when the risk genotypes do not always lead to the development of the trait (reduced penetrance) or when there are people who exhibit the trait but do not have the trait genotype (phenocopies). Second, the methods have also been extended to reflect genetic heterogeneity, the existence of genes at multiple loci leading to the development of the same trait. If it is not recognized, genetic heterogeneity can mask a true linkage signal. Third, incorporating the genetic information from several markers simultaneously can improve the linkage signal and better localize the trait

TABLE 8.1 Linkage Analysis Software for Binary Traits in Large Pedigrees and Nuclear Families and Indicating Whether the Analysis Is Parametric and Model Based or Model Free

Linkage of Binary Traits in	Program Name	Model	Extra Features
Large pedigrees	LINKAGE	Based	HOMOG: locus homogeneity testing
	MENDEL Option 2	Based	Two-point and multipoint
	SIMWALK	Both	Uses MCMC algorithm to analyze multiple markers in large pedigrees
Nuclear families	SAGE/LODLINK	Based	Two-point
	SAGE/MLOD	Based	Multipoint
	GENEHUNTER/ESTIMATE	Free	Moderate size pedigrees
	SAGE/LODPAL	Free	Parent-of-origin, covariates possible
	MERLIN	Based	Empirical p values
	SAGE/SIBPAL	Free	Empirical p values

TABLE 8.2 Linkage Analysis Software for Quantitative Traits in Large Pedigrees and Nuclear Families Indicating Whether the Analysis Is Parametric and Model Based or Model Free

Linkage of Quantitative Traits in	Program Name	Model	Extra Features
Large pedigrees	LOKI	MCMC algorithm	Multiple QTL
	SOLAR	Variance components	Bivariate QTL
Nuclear families	GENEHUNTER/	Both	Moderate size pedigrees, parent-of-origin effects
	a. MAPMAKER/SIBS		
	b. NPL		
	c. HASEMAN-ELSTON	Free	Empirical p values
	MERLIN/REGRESS		
	SAGE/SIBPAL	Free	Empirical p values

QTL, quantitative trait loci.

gene. Multipoint analyses accomplish this goal. Each of these important factors is discussed in the following sections.

8.3.1 Incomplete Penetrance and Phenocopies

Penetrance specifies the probability that an individual with one of the possible trait genotypes will exhibit the trait. Age of onset or gender-specific trait risks can be incorporated with a penetrance model through multiple liability classes with differing penetrance estimates. For example, disease penetrances for DD, Dd, and dd can be 80%, 40%, and 10% in males and 20%, 10%, and 2% in females. For unaffecteds, the DD, Dd, and dd penetrances would be 20%, 60%, and 90% in males and 80%,

90%, and 98% in females, as the two penetrances within a liability class for a given genotype must always sum to 1.0. Phenocopies reflect the occurrence of a trait indistinguishable from the trait of interest, but resulting from other causes. A phenocopy rate can be incorporated by assuming that affected individuals with a normal genotype have a small probability of expressing the disease, such as 5%. The power of linkage analysis can be affected significantly by phenocopies when they account for more than a small percentage of the cases. If penetrance values are unknown, the careful investigator will take the precaution of verifying that his or her inference of linkage does not critically depend on the assumed penetrances, by testing linkage over a range of reasonable penetrance estimates.

8.3.2 Genetic Heterogeneity

Defining inherited conditions solely on the basis of their clinical manifestations may obscure the fact that the trait under analysis is the result of distinct genetic etiologies that segregate among the families. Unless the trait is the result of different alleles at a single locus, heterogeneity may drastically reduce the power of a linkage analysis when it is not considered. There are several ways to address this. The first is to make a predefined partition of the families using a particular form of the phenotype or a factor such as their ethnicities. One then conducts a linkage analysis in each group separately and reports the results for each group. At a particular locus, one can establish locus heterogeneity by also testing for linkage in the entire sample. Using these three linkage analyses at a single locus, a likelihood ratio test is constructed, contrasting the product of the maximized likelihoods in each subset to the maximum likelihood obtained for the total sample. The test statistic is represented by $2 \log_e [\prod_i p(D_i|r_i)/p(D|r)]$, where i indexes the two sets of families. That is, twice the natural log of the product of the likelihoods of the two data sets with two different recombination fractions compared with the likelihood of all of the data combined with a single recombination fraction. Under the null hypothesis of homogeneity, a single recombination fraction, this statistic follows a χ^2 distribution with $n - 1$ *df* (degrees of freedom). Here n is the number of classes into which the data have been partitioned, which is 2 in this case, resulting in 1 *df*.

A more formal statistical approach to modeling and estimating the degree of heterogeneity among a sample of families is available through the LINKAGE software referred to as HOMOG. Here a test of heterogeneity can be used when the alternative to homogeneity is that the families belong to two etiologic classes, one linked and the other unlinked to the marker locus, although they do not have to be divided a priori as with the test described earlier. Assuming that a proportion, m , of the families exhibit linkage, while $1 - m$ are unlinked, the likelihood of the observations can be expressed in terms of two parameters, the recombination fraction r for the linked form and the admixture proportion, m . A likelihood ratio test can be formulated by contrasting the likelihood obtained when both parameters are estimated to that obtained under the hypothesis of homogeneity, where m is fixed to unity and only the recombination fraction, r , is estimated. This follows a χ^2 distribution

with 1 *df*. The two parameters m and r can be estimated simultaneously.

8.3.3 Multipoint Parametric Linkage Analysis: Location Scores

With the advent of dense panels of SNP markers, multipoint analyses may not be necessary. However, they can be informative for sparse sets of SNPs and multiallelic markers, and we include this section for those using such panels. When conducting a linkage analysis using sparse marker panels, combining the information from several markers from the same chromosome region simultaneously can provide greater statistical power to detect linkage and a more precise localization of the trait gene than when the markers are analyzed separately. Multipoint analyses are based on more complex probability models than single point. The parameters in the likelihood represent all of the between-locus recombination fractions among markers. Maximum likelihood methods, however, still apply and are used to estimate the recombination fractions between the trait locus and the points along the genetic map of markers in the analysis. The resulting calculations are computationally intensive, but can be carried out using software packages such as MENDEL and SAGE/MLOD. As the trait locus is moved across the marker region, the locus with the largest multipoint LOD score or location score localizes the trait gene to a more refined region than conducting several two-point analyses of sparse markers can accomplish.

8.4 LINKAGE ANALYSIS FOR COMPLEX AND QUANTITATIVE TRAITS

Linkage analysis has also been extended to include traits that are not Mendelian or binary [5,6]. Following the completion of the Human Genome Project, the remarkable success in identifying genes for Mendelian disorders resulted in a gradual shift toward analyzing traits with complex inheritance patterns. Since parametric linkage analysis is not well suited to genetically complex traits, model-free linkage analysis, in which a model of inheritance is not known and therefore not included in the analysis, has been conducted. Complexity may derive from something as simple as two disease loci, each exhibiting reduced penetrance, to something as complex as each family with the trait having risk alleles in each of 10 possible genes and requiring one of a large

number of environmental triggers. The analyses of both traits are conducted using the same model-free analytic methods. In addition, there has been an interest in the genetics of quantitative traits that may be important on their own or correlate with binary traits of interest. Methods and computer programs have been developed to identify the genetics of these traits as well.

8.4.1 Model-Free Linkage Analysis

Model-free methods test if the allele-sharing patterns in families are consistent with linkage in a very general way. That is, those in the pedigree who have the trait should display evidence of marker allele sharing to a greater degree than one would expect by chance alone in the linked regions. Most model-free tests include only individuals with the trait of interest. However, genotypes of their relatives are important to make more precise estimates of allele sharing in those exhibiting the trait. Using a study sample that includes only those individuals with the trait reduces the risk of including individuals who do not have the trait but share alleles with those who do, because of reduced penetrance. Including these individuals in the analysis would mask evidence of linkage. The most common model-free test statistics are based on allele sharing in sibling pairs. Limiting the analysis to sibling pairs has been effective because it is usually difficult to ascertain larger pedigrees with multiple members having the trait if it is complex. In addition, with multiple common risk alleles, members of large families may develop the trait as the result of several genetic etiologies, thus introducing within-pedigree heterogeneity, for which we do not have appropriate analytic approaches. Sibling pairs with a trait are usually available for study and are more likely to have the trait because of a shared genetic etiology. However, nuclear families contain less information about linkage than a large pedigree, and thus a larger number of these families will have to be studied to find linkage.

The sib-pair allele sharing linkage methods are based on a simple expectation. At a marker, the sibling pairs can inherit four possible alleles from their parents. If there is no gene predisposing to the trait in the region of the marker, the sibling pairs should share no alleles 25% of the time, one allele 50% of the time, and both alleles 25% of the time for that marker. The expected proportion of allele sharing is 50%. A statistically significant deviation from these theoretical values at a marker provides evidence of linkage of the trait to that marker.

For model-free analyses, the recombination fraction between the marker and the trait gene is assumed to be zero. Thus, an estimate of r is not included in the analysis. For several linkage statistics, the degree of allele sharing for each sibling pair is assessed, and the allele sharing estimates are averaged over the pairs. These individual estimates would be 0, $\frac{1}{2}$, or 1 for each pair if the genotypes were fully informative and all parents were completely genotyped. When data are missing or the parental genotypes do not allow for an unambiguous assignment of their inheritance, the marker allele frequencies are used in the estimate of allele sharing. The allele sharing estimates will usually be different from 0, $\frac{1}{2}$, or 1 to reflect this. In the simplest case, the average allele sharing value is tested against its theoretical value of $\frac{1}{2}$. An additional consideration is that if there are more than two siblings in a sibship, the pairs will not provide statistically independent allele sharing estimates. A weighting scheme to account for this can be employed. For example, three sibs in a sibship contribute information from three pairs, and since the information for the third pair can be derived from the first two pairs, allele sharing for these three pairs can be weighted by a factor of $\frac{2}{3}$. A subset of the programs listed in Table 8.1 implement model-free allele sharing linkage methods for binary traits in nuclear families. They are the SIBPAL program of SAGE, the MLS and NPL options of GENEHUNTER, and the MERLIN REGRESS program. These software packages allow for generation of the allele sharing test statistics at evenly spaced intervals along the entire chromosome, given the marker map, and a multipoint analysis is conducted for each chromosome. As with binary linkage, multipoint analyses require that the markers are independent, and the SNPs must be pruned to satisfy that assumption. The packages provide plots of the statistical significance of the tests across the chromosomes so that the loci with significant evidence for linkage can be identified easily.

8.4.2 Linkage Analysis of Quantitative Traits

A quantitative trait can be dichotomized into one that is binary for a parametric linkage analysis. For example, a systolic blood pressure of greater than 140 is used to classify an individual as hypertensive, and hypertension has often been the trait tested for linkage. In fact, the quantitative trait itself may be better suited to linkage analysis than the dichotomized trait, as the extent of variation can be assessed in all family members, thus

making it likely to provide increased statistical power to detect genes. In addition, the etiologic heterogeneity of genetically complex traits may be reduced by the analysis of a correlated quantitative trait that reflects only one feature of the complex disorder. For example, cardiovascular disease may be better addressed by studying a particular lipid trait, such as cholesterol, rather than the binary trait of a myocardial infarction, which is very likely to have a much broader etiology. Programs to analyze quantitative traits in large pedigrees usually require the assumption of trait normality, while those for smaller pedigrees do not require rigorous adherence to these assumptions [7]. Programs and their features are given in Table 8.2.

Variation in a quantitative trait usually results from the contributions of multiple genes with small effects modified by environmental influences. If none of the genes contributes a substantial amount to this quantitative variation, loci can be difficult to detect using linkage analysis. However, a gene contributing to a relatively large proportion of the variance of the trait, a major gene, is a good candidate for localization by quantitative trait linkage analysis. When conducting a quantitative trait linkage analysis, it is important to select families for which the trait exhibits marked variation within the pedigree. Those families having multiple members with extreme values of the trait are most likely to provide support for a major gene. Regions that are identified by the linkage analysis of a quantitative trait are usually referred to as quantitative trait loci or QTL.

A list of computer software commonly used for the linkage analysis of quantitative traits is given in Table 8.2. A variance component analysis of a quantitative trait, such as that conducted by the SOLAR software, identifies linked chromosome regions by decomposing the variance of the trait into the components that contribute to it [8]. The log of the ratio of the likelihood of the data with a major gene is compared with the likelihood of the data when there is no major gene modeled at that location. Normality of the trait distribution is an important assumption and if the trait is not normally distributed, transformation to normality is critical to a successful analysis. For nonnormal traits in smaller pedigrees, GENEHUNTER/MAPMAKER/SIBS, MERLIN/REGRESS, and SAGE/SIBPAL provide good alternatives [9]. GENEHUNTER/NPL, which uses a nonparametric method that ranks the trait values, is robust against nonnormality. The ordered subset analysis approach

implemented in OSA can be used to identify QTL for quantitative traits that are correlated with binary traits. The families in the analysis are ordered according to their scores on a quantitative correlated value and the evidence for linkage is assessed as the ordered families are sequentially included in the analysis.

8.5 LINKAGE ANALYSIS: FUTURE DIRECTIONS

Linkage analysis has been used to identify chromosome locations of human trait genes for more than 50 years. Its well-developed tools include families, markers, and statistical methods of analysis. The genes for many Mendelian disorders have been identified with these tools. However, as the number of genetic markers increased from 30 to 10 million, our ability and enthusiasm to localize and identify genes for traits of increasing genetic complexity have grown proportionally. Such marker density allows us to capitalize on linkage disequilibrium to identify trait genes, and consequently, since the early 2000s, there have been few modifications to the well-established linkage methods. The primary approach to gene identification has quickly transitioned to genome-wide association studies (GWAS), which capitalizes on the linkage disequilibrium among closely spaced markers in populations [10]. In some sense, analyzing linkage disequilibrium in populations is similar to analyzing linkage in families. To clarify, in the current generation, within a population, SNP alleles are in linkage disequilibrium because they are too close to each other on the chromosomes to have undergone significant recombination over the generations. Thus, one can view GWAS as linkage analyses of a large pedigree that is the current population under analysis. All mapping information is in the current generation, and we test for association of specific marker alleles and the trait of interest.

GWAS are based on association tests of common SNP variants whose frequencies are larger than 1%. The genotyped SNPs have been selected to tag trait variants that may not be tested directly. Although consistent gene associations have been identified for a substantial number of complex traits, their effects have been surprisingly small. Successful GWAS have required very large samples, and the detected effect sizes indicate that the risk is raised at most by about 10%–20% compared with that of the background genotype. Consequently, the genetics literature has expressed concern that GWAS

have not revealed the etiologies of traits with significant heritabilities.

Recently, concern about the small effect sizes of common variants, as well as the dramatic reduction in the cost of whole-exome sequencing with “next-generation” methods, has refocused the interest of many of those studying complex disorders toward the detection of rare genetic variants via sequencing. These variants are expected to have a frequency less than 1%, and may even consist of private mutations. They are also expected to exhibit greater penetrance values and effect sizes than the common variants. To sort through the many variants likely to be uncovered, there has been a renewed interest in the study of large pedigrees. If they exist for a complex trait, analyses will be focused on identifying those pedigrees that segregate the trait where it is acting in a quasi-Mendelian fashion. As with Mendelian disorders, model-based linkage analysis that allows for reduced penetrance and phenocopies may reveal important loci. Targeted sequencing in linked regions could reveal the genes with the segregating rare variants. With this approach, parametric linkage analysis is likely to again become a natural first step in gene-finding efforts over the next few years.

REFERENCES

- [1] Ott J. Analysis of human genetic linkage. Baltimore: Johns Hopkins University Press; 1999.
- [2] Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* 1955;7:277–318.
- [3] Nyholt DR. Invited Editorial: all LODs are not created equal. *Am J Hum Genet* 2000;67:282–8.
- [4] Cantor RM. Model-based linkage analysis of a binary trait. *Methods Mol Biol* 2012;285–300.
- [5] Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 1990;46:222–8.
- [6] Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet* 1990;46:229–41.
- [7] Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972;2:3–19.
- [8] Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998;62:1198–211.
- [9] Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995;57:439–54.
- [10] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108.

BIBLIOGRAPHY

The following are Web pages for linkage analysis software mentioned in the text and listed in [Tables 8.1 and 8.2](#).

GENEHUNTER: <http://www.broad.mit.edu/ftp/distribution/software/genehunter/>.

LINKAGE: <ftp://linkage.rockefeller.edu/software/linkage>.

LOKI: <http://www.stat.washington.edu/thompson/Genepi/Loki.shtml>.

MENDEL: <http://www.genetics.ucla.edu/software/>.

MERLIN: <http://www.sph.umich.edu/csg/abecasis/Merlin>.

OSA: <http://wwwchg.duhs.duke.edu/research/aplosa.html>.

SAGE: <http://darwin.cwru.edu/sage/>.

SIMWALK: <http://www.genetics.ucla.edu/software/simwalk>.

SOLAR: http://www.sfbr.org/Departments/genetics_detail.aspx?p=37.

Chromosomal Basis of Inheritance*

Fady M. Mikhail

Cytogenetics Laboratory, Department of Genetics, University of Alabama at Birmingham,
Birmingham, AL, United States

9.1 INTRODUCTION

The human genome is packaged into a set of chromosomes as in other eukaryotes. Chromosomes are thus the vehicles of inheritance as they contain virtually the entire cellular DNA, with the exception of the small fraction present in the mitochondria. The structure, function, and behavior of chromosomes are therefore of much interest and importance. Chromosomes are derived in equal numbers from the mother and the father. Each ovum and sperm contains a set of 23 different chromosomes, which is the haploid number (n) of chromosomes in humans. The diploid fertilized egg and virtually every cell of the body arising from it has two haploid sets of chromosomes, resulting in the diploid human chromosome number ($2n$) of 46. The human karyotype consists of 22 pairs of autosomes and a pair of sex chromosomes. The correct chromosome number in humans was determined and confirmed in 1956 [1,2].

The behavior of chromosomes during meiotic cell division provides the basis for the Mendelian laws of inheritance, whereas their abnormal behavior in cell division leads to abnormalities of chromosome number. In this chapter, we examine the current understanding of the structure, molecular organization, and behavior of human chromosomes and explore how these features contribute to chromosomal diseases.

9.2 CHROMOSOME STRUCTURE

Although the structure of human and other eukaryotic chromosomes is not understood in full detail, recent investigations have provided insights into several

aspects of chromosome structure at the molecular level. The haploid human genome consists of about 3×10^9 base pairs (bp) of DNA. Since 3000 bp of naked DNA are $\sim 1 \mu\text{m}$ long, the total length of the diploid human genome is about 2 m. As the cell nucleus is no more than $10 \mu\text{m}$ in diameter, it is necessary to fold and compact this DNA, which is accomplished by packaging it in a hierarchy of levels into chromosomes of manageable size (Fig. 9.1). Organization of the DNA into chromosomes also maintains the linear order of genes and facilitates faithful replication and segregation of genetic material during cell division. The first level of this packaging, and thus the fundamental unit of chromosome organization, is a regularly repeating protein–DNA complex called the nucleosome. The basic structural features of the nucleosome were established in the early 1970s and have been further confirmed by high-resolution analysis of its crystal structure [3]. The nucleosome has the same design in all eukaryotes and consists of a cylindrical core about 11 nm in diameter and 6 nm in height made up of two molecules each of the four core histones (H2A, H2B, H3, and H4) with 147 bp of DNA wrapped around it. A “linker” DNA connects adjacent nucleosomes. Each nucleosome is also associated with a molecule of histone H1, which changes the path of the DNA as it exits from the nucleosome, and plays a role in further condensation of chromosomal DNA. Formation of the nucleosomes achieves a sevenfold compaction of the

* This chapter is a revision of the previous edition chapter by Julie R. Korenberg and T.K. Mohandas, vol. 1, pp. 167–190, © 2007, Elsevier Ltd.

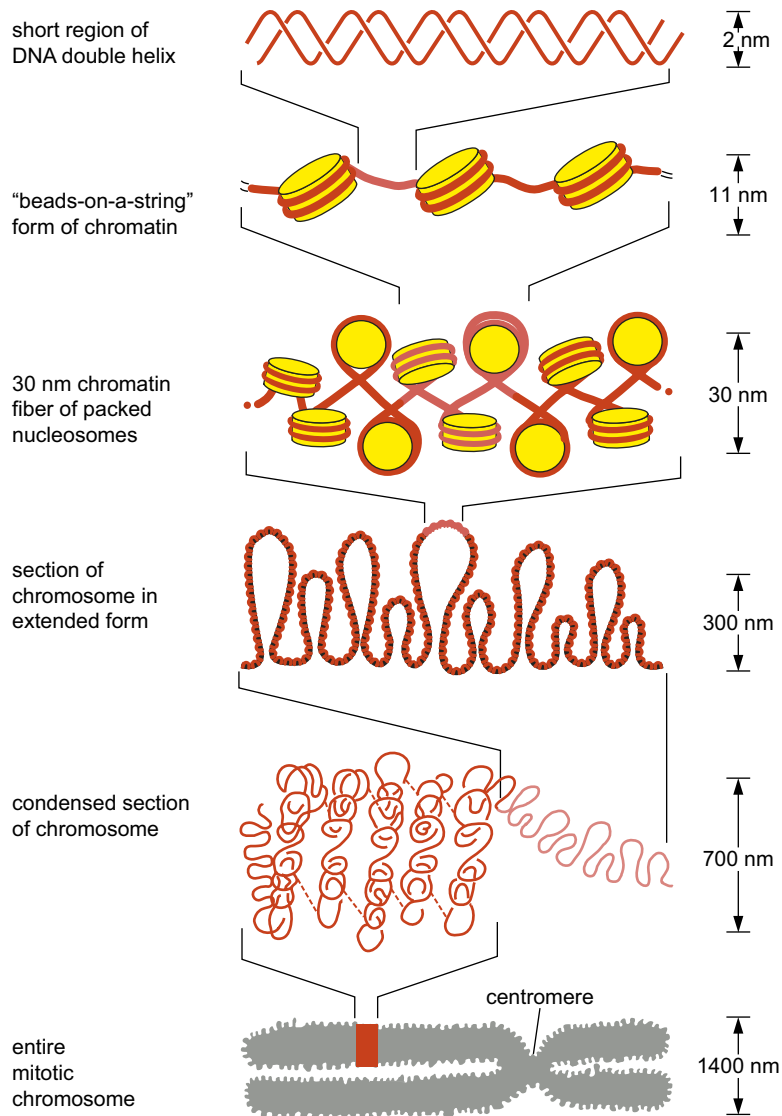


Figure 9.1 Various levels of DNA packaging in the cell. (Reproduced from *Molecular biology of the cell*, fourth ed., by Bruce Alberts, et al., Copyright © 2002 by Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. (c) 1983, 1989, 1994 by Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson. Used by permission of W. W. Norton & Company, Inc.)

DNA doublehelix. The next higher level of packaging is the chromatin fiber, visible by standard electron microscopy. This is a superhelix, 30 nm in diameter, composed of nucleosomes and histone H1. The 30-nm fiber is the basic component of interphase chromatin and metaphase chromosomes. Two models have been proposed for the formation of the 30-nm chromatin fiber. In the first model, called the "solenoid," consecutive nucleosomes are located next to each other in the fiber, folding

into a simple "one-start helix" [4]. Subsequently, a second model of the two-start helix was proposed on the basis of microscopic observations of isolated nucleosomes [5]. Although some variations exist in this model, essentially, nucleosomes are arranged in a zigzag manner, such that a nucleosome in the fiber is bound to the second neighbor, but not the first [6–8]. In 2009, it was shown that the two-start zigzag and one-start solenoid models may be present simultaneously in a 30-nm chromatin fiber

under certain conditions [9]. The structural details of the 30-nm chromatin fiber remain controversial. Formation of the 30-nm chromatin fiber achieves a nearly 50-fold compaction of the DNA doublehelix. Short AT-rich regions referred to as matrix attachment regions (MARs) that occur at about every 30–150 kb of DNA anchor the chromatin fiber to the proteins of the nuclear matrix of the interphase nucleus. Topoisomerase II, an enzyme that induces transient double-strand breaks in DNA and permits uncoiling of the two strands of the DNA duplex, is a major matrix protein. At the next level of packaging, the 30-nm chromatin fiber is arranged into loops that radiate from a core or scaffold of the metaphase chromosome. The MARs are also the site for attachment of the chromatin fiber to the nonhistone protein scaffold of the metaphase chromosome (hence also called scaffold attachment regions) [10]. Topoisomerase II is a component of the chromosome scaffold and has been shown to play a role in chromosome condensation. The other major component of the metaphase chromosome scaffold that also plays a key role in chromosome condensation is the condensin complex, a member of the SMC (structural maintenance of chromosomes) family of proteins [11]. Other members of the SMC family of proteins mediate chromosomal functions such as sister chromatid cohesion (cohesin complex) and DNA repair [11]. In addition to topoisomerase II and condensins, cations are also believed to be essential participants in chromosome condensation [12,13]. The details of the higher order structure of chromosomes are not well understood at the molecular level. However, it is clear that each chromosome contains only a single very long duplex of DNA with an estimated packaging ratio of about 1:10,000. At the highest level of compaction, the metaphase–anaphase chromosomes are most easily movable by the spindle apparatus during cell division.

9.3 CHROMOSOMES IN CELL DIVISION

Cell division and proliferation are central to growth and development of multicellular organisms. The major events in the cell cycle are replication and segregation of chromosomes. Cell division also ensures proper segregation and partitioning of the genetic material into daughter cells, thus providing the basis for Mendelian laws of inheritance. The cytologic aspects of mitosis and meiosis, the two forms of cell division in eukaryotes, have been described in great detail in numerous

studies in the past. However, the explosive growth of molecular biology since the end of the 20th century has brought the study of mitosis and meiosis to the forefront again. These investigations have elucidated biochemical aspects of cell cycle biology and chromosome mechanics [14,15]. This section describes the essential features of mitosis and meiosis relevant to inheritance. Knowledge of these features is crucial for understanding the Mendelian laws of inheritance, the construction of genetic maps, and the origin of chromosome aberrations.

9.3.1 Mitosis

In somatic cells, and in cells of the germline prior to the time they undergo their first specialized meiotic divisions, nuclear division takes place by a process called mitosis. During mitosis, each chromosome divides into two daughter chromosomes (sister chromatids), one of which segregates into each daughter cell. Therefore, the number of chromosomes per nucleus remains unchanged, producing daughter cells with identical chromosome constitutions. In cells with a generation time of 18–24 h, mitosis takes about 1–2 h and is divided into five major stages: prophase, prometaphase, metaphase, anaphase, and telophase (Fig. 9.2).

In the initial phase of mitosis, prophase, the chromosomes become visible as a result of condensation that continues throughout this phase. Each chromosome has already undergone replication during the preceding interphase, generating two sister chromatids that will become daughter chromosomes. The sister chromatids are closely held together along their length by cohesins until anaphase. The centrioles duplicate during the S phase and move apart to occupy positions at opposite ends of the cell, defining the poles of the mitotic spindle.

During prometaphase, the nuclear membrane begins to disintegrate, allowing the chromosomes to spread around the cell. In metaphase, the chromosomes become attached to microtubules of the mitotic spindle at their centromere and undergo movements that lead to their alignment in the equatorial plane of the spindle. At this stage, the chromosomes have reached their maximum state of condensation.

Anaphase begins after the chromosomes are fully aligned on the metaphase plate. Each pair of sister chromatids separates as cohesion is lost, first along the arms and finally at the centromere of each chromosome. The resultant daughter chromosomes move toward opposite poles of the spindle as a result of microtubule dynamics

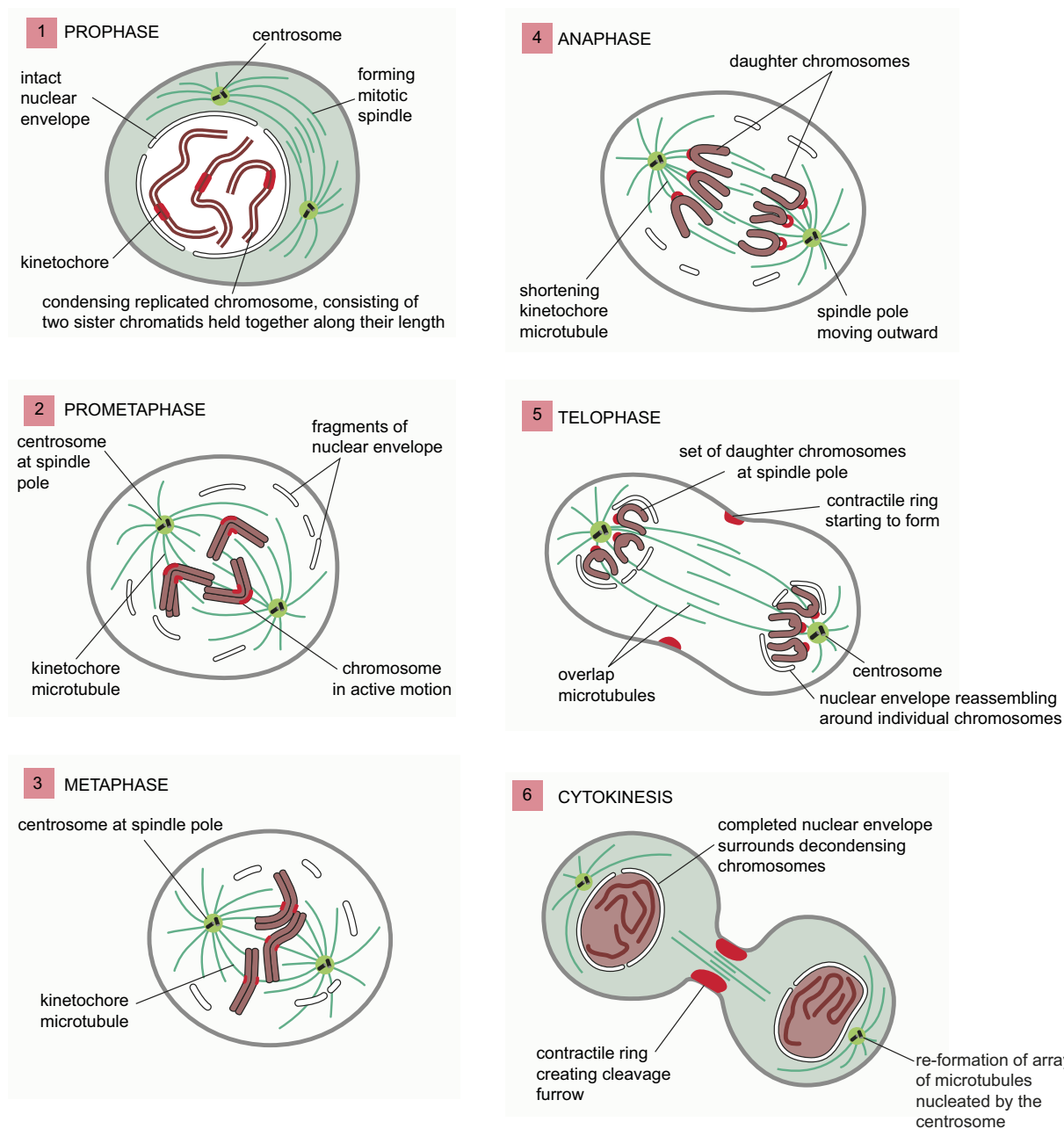


Figure 9.2 Diagrammatic representation of the stages of mitosis. (Reproduced from *Molecular biology of the cell*, fourth ed., by Bruce Alberts, et al., Copyright © 2002 by Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. (c) 1983, 1989, 1994 by Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson. Used by permission of W. W. Norton & Company, Inc.)

and the action of motor proteins. The two sister chromatids of a chromosome are held together following chromosome replication by cohesins. This is a multisubunit protein complex that ensures correct segregation of daughter chromosomes at anaphase. At the beginning of anaphase, the cohesin complex is cleaved by a protease called separase, allowing separation of the sister chromatids [16].

At telophase, each set of daughter chromosomes arrives at the centriole at one of the two ends of the mitotic spindle, and reconstitution of the nuclear membrane begins. Cytokinesis, the division of the cytoplasm, follows telophase and leads to the formation of two genetically identical daughter cells.

9.3.2 Meiosis

Meiosis is a specialized cell division in germ cells that generates gametes with the haploid set of 23 chromosomes. The final gametic set includes single representatives of each of the 23 chromosome pairs selected at random. The details of meiosis and gamete formation are somewhat different in males and females, but the basic features are the same in both and are of fundamental importance. Meiosis accounts for the major principles of Mendelian genetics: segregation, independent assortment, and recombination of linked genes. Recombination or crossing over is the exchange of genetic material between homologous nonsister chromatids, a process that adds to genetic diversity by generating new combinations of genes.

Meiosis consists of two cell divisions (meiosis I and II) and is distinguished from mitosis by the following:

1. **Homologous pairing:** Maternal and paternal homologues of each chromosome are replicated and then undergo exact pairing along their lengths during prophase of meiosis I. Such a paired unit is called a “bivalent” because there are only two centromeres, although it is composed of four chromatids.
2. **Recombination (crossing over):** Crossing over occurs at the four-strand stage between nonsister chromatids, that is, chromatids from each of the pair of homologous chromosomes. The probability of recombination increases with the physical distance between two chromosomal sites and therefore provides a basis for the genetic map.
3. **Segregation of maternal and paternal homologues:** Centromeres do not divide at the first meiotic division (meiosis I). Instead, the members of a homologous

pair go to opposite poles at anaphase of the first meiotic division. This accounts for Mendel’s first law, the segregation of homologous genetic units. The segregation of maternal and paternal homologues in each bivalent chromosome pair occurs independent of the segregation in all the other bivalents. That is, the segregation of chromosome 1 homologues is independent of that of chromosome 2 homologues and so on. This accounts for Mendel’s second law of independent assortment of genes. Meiosis I also leads to a reduction in the number of chromosomes from the diploid number ($2n = 46$) to the haploid number ($n = 23$) in the gametes.

4. **Division of the haploid set with centromere division:** The second meiotic division (meiosis II) occurs without a preceding round of DNA synthesis and chromosome duplication. In meiosis II, the two chromatids of a chromosome move to opposite daughter cells.

Meiotic prophase I is rather prolonged and can be subdivided into five stages on the basis of condensation of chromosomes and the extent of homologous pairing: leptotene, zygotene, pachytene, diplotene, and diakinesis. In leptotene, the chromosomes start to condense and are visible as long threads, but the homologues are still not paired. Chromosomal condensation continues and pairing of homologues (synapsis) begins at zygotene and is completed at pachytene, the stage at which recombination occurs (Fig. 9.3). By the pachytene stage, synapsis between homologues is completed and crossing over between nonsister chromatids occurs, during which homologous regions of DNA are exchanged. Synapsis is thought to be mediated and stabilized by the formation of the synaptonemal complex (SC) between homologous chromosomes. The SC is a protein-rich ladderlike structure that has a central element flanked by two lateral elements. The lateral elements and the central element are held together by transverse filaments. The lateral elements are formed of the axial elements of sister chromatids of the paired homologues, and the bulk of the chromosomal DNA is found in chromatin loops emanating from the outer sides of the two lateral elements. Initially, the SC was characterized by ultrastructural analysis. More recent studies have identified several protein components of the SC, although the functions of many of these are unknown. A constituent of the lateral elements is cohesin, consistent with the fact that the sister chromatids of each of the homologues are held together at pachytene. Another interesting feature

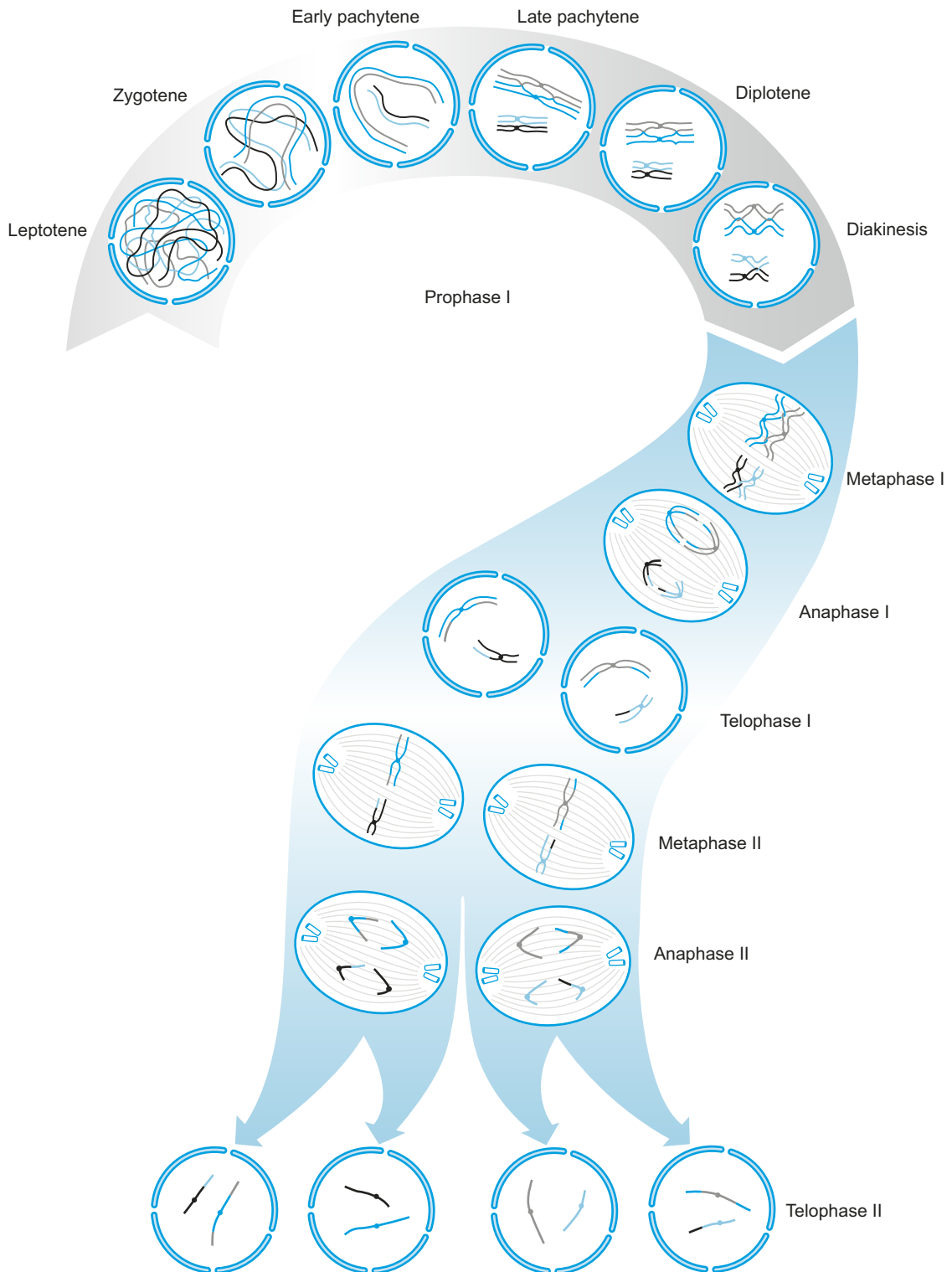


Figure 9.3 Diagrammatic representation of the stages of meiosis. (Reproduced from Turnpenny PD, Ellard S. Chapter 3: chromosomes and cell division, Figure 3.19: stages of meiosis. In: Emery's elements of medical genetics, 12th ed. Churchill Livingstone: Elsevier; 2005.)

of SC is the presence of recombination nodules along its length. These are thought to be enzyme complexes that mediate genetic recombination via DNA breakage and repair. Chiasmata, or cruciform structures, become visible at the more condensed diplotene stage as cohesion is lost along the chromosome arms except at each chiasma, the point of recombination between homologous chromosomes.

Chiasmata are still visible at diakinesis, the stage of maximal condensation, and can be used to determine the frequency as well as the location of recombination. Chiasmata, like their underlying recombination events, play an important role in the normal segregation (disjunction) of homologues, and each pair of homologues has at least one chiasma per chromosome arm. Moreover, failure of chiasma formation predisposes to non-disjunction of homologues.

During prophase I in males, pairing and crossing over between the X and the Y chromosomes are possible because of a small region of homology at the terminal ends of their arms (i.e., pseudoautosomal regions). The two chromosomes pair and cross over in these regions during prophase I.

In meiotic metaphase I, the nuclear membrane disappears and the chromosomes become aligned on the equatorial plane of the cell where they have become attached to the spindle, as in metaphase of mitosis. Then in anaphase I, the chromosomes now separate to opposite poles of the cell as the spindle contracts. In telophase I, each set of haploid chromosomes has now separated completely to opposite ends of the cell, which cleaves into two daughter gametes, so-called secondary spermatocytes or oocytes.

The meiosis II division resembles an ordinary mitotic division, except for the presence of a single set of 23 duplicated chromosomes, each with two chromatids held together at their centromere. Also, meiosis II is not strictly a genetically equal division as the two chromatids of a chromosome may not be identical as a result of genetic exchange(s) with a nonsister chromatid. At the end of the two meiotic divisions, each primary spermatocyte or oocyte has given rise to four haploid products (see Fig. 9.3). Their fates are rather different in males and females, as discussed later.

9.3.3 Spermatogenesis and Oogenesis

In the human male, the production of sperms begins at puberty and continues throughout life. Undifferentiated

stem cells of the germline, the spermatogonia, are abundant in the seminiferous tubules of the testis and show a high rate of mitotic activity throughout the adult life of a normal male. Of the two types of spermatogonia, only type A can differentiate into primary spermatocytes that enter meiosis, whereas type B are the long-lived progenitors that divide to generate daughter cells of both types A and B. In meiosis, each diploid spermatocyte gives rise to four haploid cells, each of which differentiates into a functional sperm. The entire process, from spermatogonium to sperm, takes about 70 days. The rate of sperm production may be as high as 50–100 million per day over many years, and thus the parental spermatogonia undergo many successive mitoses. It is estimated that the number of mitoses before sperm production in a 20-year-old male is about 200, while in a 45-year-old it is about 800 [17]. This provides the opportunity for the occurrence of more adverse genetic change with age in males, which is reflected in an increased mutation rate for certain inherited diseases.

The behavior of germline cells in the female is quite different from that in the male. By about the fourth month of prenatal development, about 7 million oogonia have begun to develop into primary oocytes and to enter meiosis. Primary oocytes proceed only as far as prophase of meiosis I by the time of birth, in which they remain until ovulation. This suspended stage of prophase occurs after pachytene, is referred to as dictyotene, and lasts from birth until after puberty, when small cohorts of the germ cells progress further into meiosis. The first meiotic division is stimulated by ovulation and is an unequal division in that most of the cytoplasm remains in the ovum and very little is pinched off to enter the first polar body, containing one set of homologues. Sperm penetration of the ovum stimulates the second meiotic division, leading to formation of the second polar body that contains a haploid set of chromosomes. On average, one oocyte per ovarian cycle completes the first meiotic division and proceeds to metaphase of the second meiotic division; if fertilized by a sperm, it completes the second division and embryonic development ensues. Thus, over the approximate 30-year reproductive lifetime of a female, only a few hundred oocytes complete the first meiotic division and few—if any—complete the second [18].

It is of interest to note that the frequency of point mutations and structural chromosomal changes is in general higher in male gametes and increases with age.

This increased mutation rate in males is attributed to the much larger number of cell divisions in the male germline. In contrast, changes in chromosome number increase with age in female gametes. Errors of disjunction seen with advanced maternal age appear to be related to the 13–50 years the oocytes spend in prophase before chromosome segregation. Genetic mapping studies indicate that the number as well as the positioning of crossover events influences meiotic segregation of chromosomes [19,20]. However, the molecular causes underlying age-dependent nondisjunction are still poorly understood [18].

9.4 METHODS FOR STUDYING HUMAN CHROMOSOMES

Technical innovations since the 1960s have revolutionized the study of human chromosomes. Chromosomes are normally visible only during cell division as they become condensed in preparation for orderly division. Therefore, chromosomes can be studied only in cells that are dividing *in vivo* or *in vitro*. Dividing cells are sufficiently common in some tissues *in vivo* to permit the direct study of chromosomes. This is true of meiotic divisions in the testis and embryonic ovary, and of mitotic divisions in the bone marrow, some epithelia, and tumors. However, cell culture methods have greatly extended the range of tissue and cell types from which dividing cells can be obtained *in vitro*. These include blood lymphocytes, fibroblasts from skin and other tissues, and cells from amniotic fluid or chorionic villi. Viable cells can even be obtained for a number of hours after the death of an individual or spontaneous abortion of an embryo. It is thus possible to carry out chromosome studies in a wide range of clinical situations.

The introduction of a short-term peripheral blood culture technique provided a reliable way to obtain human chromosome preparations of good quality for human cytogenetic investigations and for clinical diagnosis [21]. In this widely used technique, T lymphocytes from a small sample of peripheral blood are stimulated to divide in culture with a mitogen, such as phytohemagglutinin. The blood culture is initiated in a suitable culture medium at 37°C, and within 3 days the stimulated lymphocytes provide very large numbers of dividing cells. These are blocked in metaphase by adding a mitotic spindle poison, such as colchicine, to the culture for a few minutes. Treatment of the cells with a

hypotonic solution swells the cells and allows spreading of the chromosomes, which are then fixed and mounted on a glass slide. This makes it possible to prepare well-spread, flattened metaphase chromosome preparations on slides suitable for microscopic analysis using chromosome banding methods and molecular cytogenetic techniques.

9.4.1 Human Chromosome Identification

In the early days of human cytogenetic investigations, chromosomes were stained with Giemsa or a similar dye, yielding uniform staining along their lengths. Based on these studies, human chromosomes were classified according to their size and morphology (Figs. 9.4 and 9.5). The primary constriction represents the centromere, the chromosomal locus responsible for proper segregation of chromosomes to daughter cells during cell division. Based on the position of their centromere, human chromosomes are classified as metacentric, in which the centromere is at or near the middle of the chromosome; submetacentric, in which the centromere is located significantly off center; or acrocentric, in which the centromere is very close to one end. For all categories, the short arm of the chromosome is referred to as “p” for petite, and the long arm as “q”. In addition to the centromere or primary constriction, five pairs of the acrocentric chromosomes (numbers 13, 14, 15, 21, and 22) may exhibit secondary constrictions on their short arms (Fig. 9.6). These mark the site of each cluster of ribosomal RNA (rRNA) genes and are called nucleolar organizing regions (NORs) because, at telophase, nucleoli are formed at a subset of these sites that are transcriptionally active. The rRNA genes remain in a moderately extended state at metaphase, reflecting the late shutoff of these genes in prophase and the rapid reinitiation of their transcription after the anaphase separation of sister chromatids. Originally, the chromosomes were assigned to groups A through G according to their general size and the position of the centromere (group A, 1–3; group B, 4 and 5; group C, 6–12 and X; group D, 13–15; group E, 16–18; group F, 19 and 20; group G, 21, 22, and Y) (Fig. 9.4).

9.4.2 Chromosome Banding

The conventional (Giemsa) staining without pretreatment does not permit precise identification of each chromosome in the human complement. A major technical innovation in human cytogenetics came in 1970, when Caspersson and colleagues discovered that human

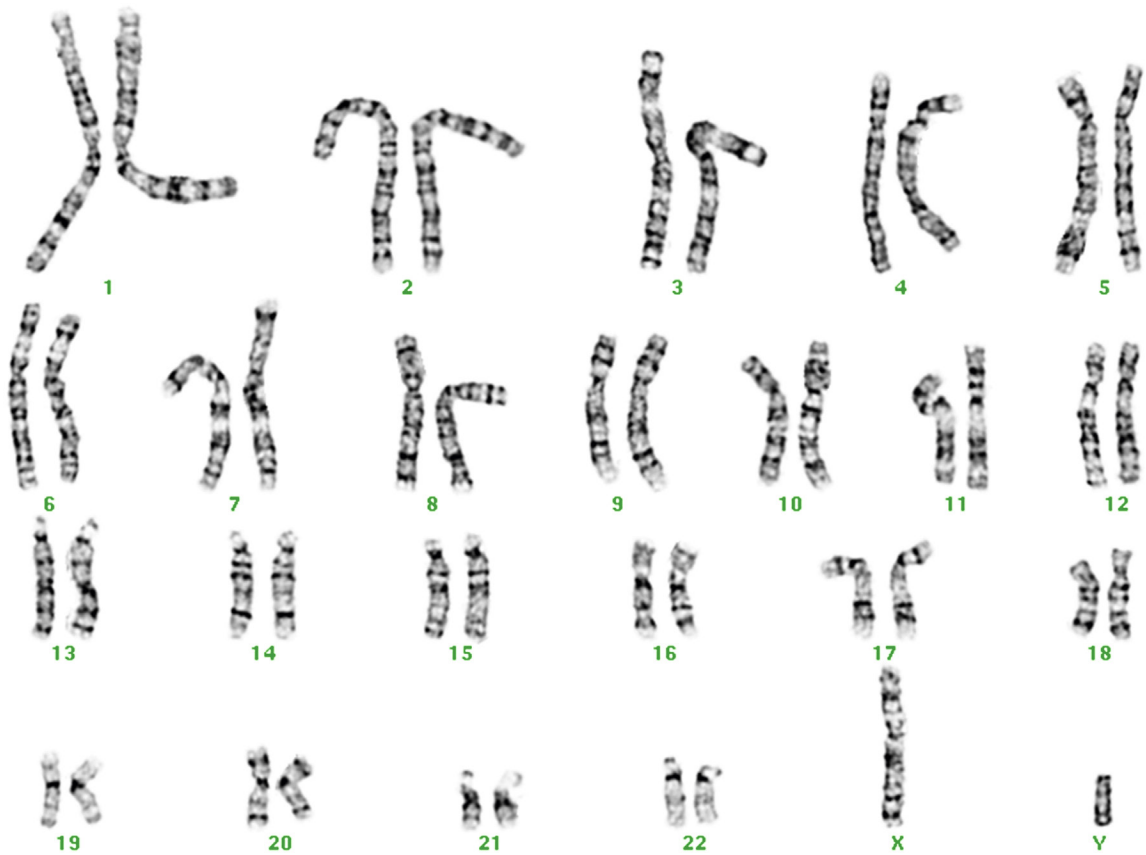


Figure 9.4 Giemsa-banded karyotype of a male cell.

chromosomes stained with quinacrine mustard, a fluorescent DNA-binding compound, and examined under ultraviolet light show characteristic variation of fluorescence intensity along the length of each chromosome, producing a banded appearance [22]. Each chromosome could then be identified by its characteristic quinacrine (Q)-banding pattern (see Fig. 9.5). Subsequently, several techniques were developed that reveal banding patterns reflecting the underlying structural features of chromosomes. Techniques such as Giemsa (G)-banding and reverse (R)-banding produce the full range of bands along each chromosome, allowing identification of individual human chromosomes. Other banding techniques produce much more restricted staining of specific subsets of chromosome bands and include centromere (C)-banding and NOR-banding. A technique that differentially stains the two sister chromatids of a chromosome is also of particular interest. Chromosome

banding methods of special interest are discussed in the following paragraphs.

Although Q-banding was the method first employed for human chromosome identification, it is rarely used today for routine chromosome analysis in clinical cytogenetics laboratories, as simpler methods have become available that do not require the use of a fluorescence microscope. A banding pattern that is almost identical to the Q-banding pattern can be produced by treatment of chromosomes with a denaturing agent or a proteolysis enzyme, prior to staining them with Giemsa. In the most consistent and commonly used version of this technique, chromosomes are treated with a dilute solution of trypsin followed by staining with Giemsa [23]. The resulting G-banding is the most widely used technique for human chromosome identification in clinical cytogenetics laboratories today (see Fig. 9.4). The G-banding patterns are also readily captured and analyzed by

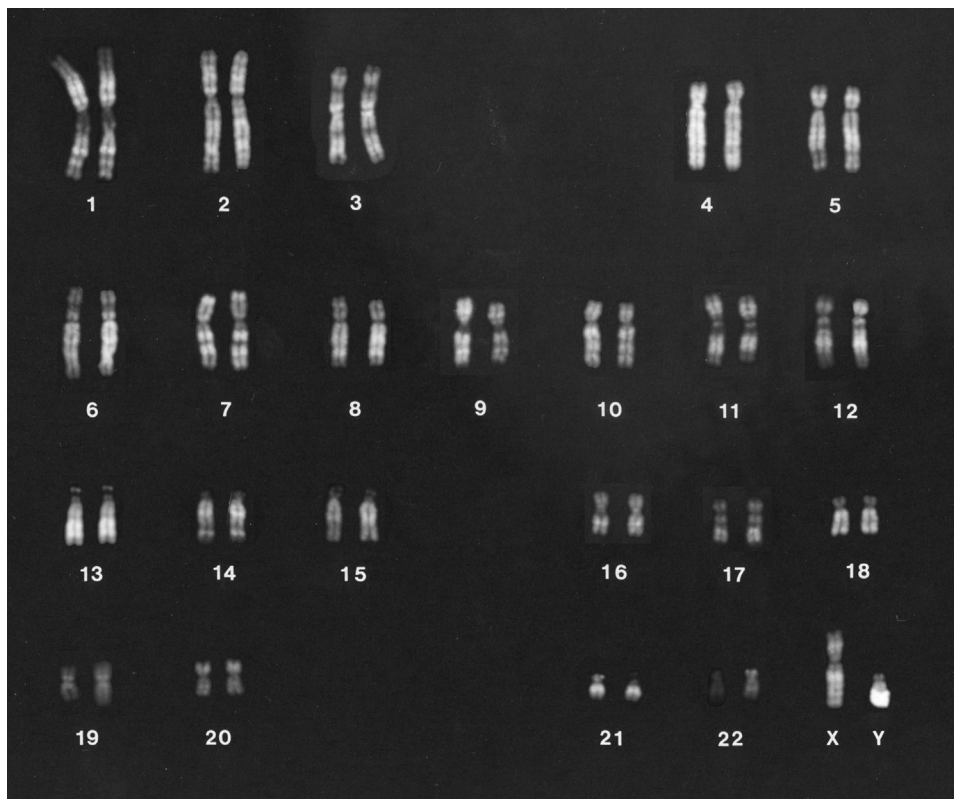


Figure 9.5 Quinacrine-banded karyotype of a male cell.

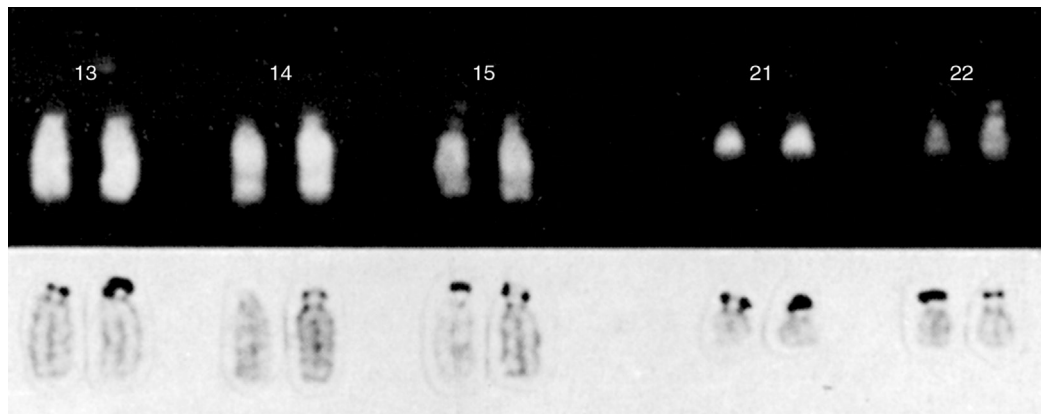


Figure 9.6 Partial karyotype of the acrocentric chromosomes from a single cell stained first by quinacrine-banding to identify each chromosome (*top*), and then by the silver NOR technique (*bottom*) to show the sites of the rRNA genes (nucleolar organizing regions, or NORs).

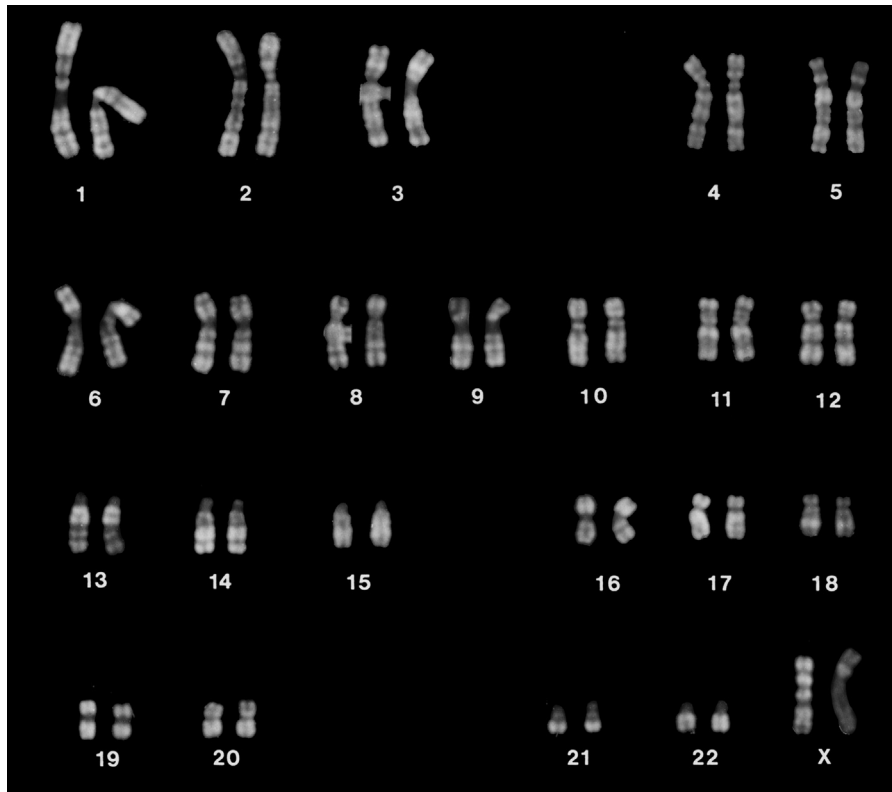


Figure 9.7 Reverse-banded karyotype of a female cell following incorporation of bromodeoxyuridine into the late-replicating regions of the chromosome.

computerized karyotyping systems used in clinical cytogenetics laboratories. As an extension of the G-banding technique, methods are now in use for obtaining longer, less condensed prometaphase chromosomes that exhibit twice as many G-bands (about 800 bands per haploid set) as the usual metaphase chromosome preparations (about 400–500 bands per haploid set), providing higher resolution to cytogenetic analysis [24].

Another banding technique of interest, although less commonly used for routine analysis, is one that produces an R-banding pattern. Many approaches have been developed to obtain R-banding, in which the staining intensity of each band is the reverse of that seen with Q- or G-banding. A commonly used method to generate R-bands is to subject chromosome preparations to moderate heat (~85°C in the presence of high salt) before staining them with Giemsa. R-banding of the highest resolution is obtained by a combination of the fluorescent dye chromomycin A3, which emits fluorescence most strongly in the R-bands, and distamycin A,

which quenches fluorescence in G-bands. An alternative R-banding technique that also provides some insight into the mechanism of chromosome banding is based on the differential replication timing of chromosome bands [25]. In this technique, growing cells are exposed to the thymidine analogue bromodeoxyuridine (BrdU) during the S phase of the cell cycle and examined following staining with the fluorochrome acridine orange. Cells that incorporate BrdU into their DNA at the late stage of the S phase are selected for observation. With acridine orange staining, the BrdU-containing chromosomal regions appear dull (G- or Q-bands), whereas the early-replicating regions that have incorporated thymidine fluoresce brightly, giving a reverse or R-banding pattern (Fig. 9.7). Thus, the R-bands on chromosomes represent regions that replicate their DNA early in the synthetic (S) phase of the cell cycle. The inactivated, late-replicating X chromosome in females (see later) is also stained differentially (dull) from the active X chromosome following this replication banding protocol (see Fig. 9.7).

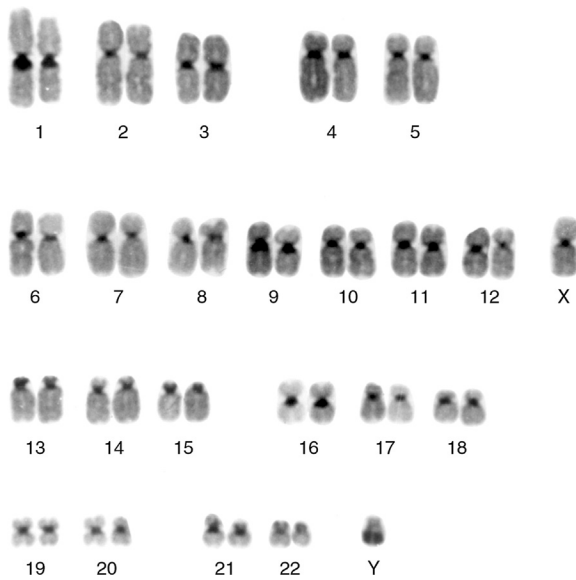


Figure 9.8 C-banded karyotype of a male cell.

The banding pattern of each chromosome is specific and can be shown in the form of a continuous series of bands. A standardized map of banded chromosomes is known as an “idiogram.” Subsequent to the development of banded human karyotypes, a standardized nomenclature for the bands was established by the International Standing Committee on Human Cytogenetic Nomenclature. Updated regularly, this standardized system allows the precise description of chromosome abnormalities [26].

The C-banding method selectively stains the areas located around the centromeres of all chromosomes and on the distal long arm of the Y chromosome [27]. The largest C-bands usually occur on chromosomes 1, 9, 16, and Y in regions that contain highly repetitive, nontranscribed DNA. To elicit C-bands, metaphase chromosome preparations are treated with sodium hydroxide or barium hydroxide followed by Giemsa staining (Fig. 9.8). The size of the C-band on a given chromosome is usually constant in all the cells of an individual but is highly variable from person to person, reflecting variations in the amount of heterochromatic DNA present at the centromeric regions. Such C-band heteromorphisms on chromosomes are transmitted from parent to offspring as simple Mendelian dominant traits. These variations in chromosome morphology are not associated with any known phenotypic effects and are referred to as chromosome polymorphisms. They are, however, useful as

heritable chromosome markers in various clinical and epidemiologic studies of chromosome abnormalities.

Silver NOR (AgNOR) staining uses a silver nitrate solution to selectively stain the sites of transcriptionally active rRNA genes, which are located in the stalk regions on the short arms of human acrocentric chromosomes [28]. Silver staining regions are usually present on 6–8 of the 10 acrocentric chromosomes, 13, 14, 15, 21, and 22 (Fig. 9.6), although they may be seen on as few as 3 or as many as all 10 of these chromosomes. The sizes of the AgNORs are highly variable in the human population, although the size of each AgNOR in the cells of one individual is quite consistent and usually remains unchanged from one generation to the next. AgNOR staining is useful in characterizing rearrangements involving human acrocentric chromosomes. The mechanism of AgNOR staining is based on the oxidation of nucleolar nonhistone proteins with silver nitrate, by which Ag is reduced to black native silver. Interestingly, the acrocentric chromosomes show association of their satellite stalk regions even in metaphase chromosome preparations, reflecting the functional association of these sites in the formation of the nucleolus in the interphase nucleus. This association of the NORs is considered to be a factor responsible for the high incidence of Robertsonian translocations involving the short arms of acrocentric chromosomes.

Sister chromatid exchange (SCE) is an extension of the replication banding technique using BrdU incorporation to produce differential staining of the two sister chromatids of the metaphase chromosome. This requires incorporation of the thymidine (T) analogue BrdU (B) into DNA during two successive rounds of DNA replication. At the end of the first round of DNA replication, the two newly synthesized strands of DNA in the double-stranded helix will contain B, but not the two template strands. At the end of the second round of DNA replication, two new double-stranded helices will be produced, of which one will have B incorporated on both strands (BB) and the other will have B substitution in only one strand of the DNA double helix (TB). When the chromosomes containing singly (TB) and doubly (BB) substituted chromatids are stained with the DNA-binding fluorochrome Hoechst 33258, and exposed to ultraviolet light, they show differential sister chromatid staining, with the bifilarly substituted chromatid exhibiting paler fluorescence [29]. Staining of these B-incorporated chromosomes with Giemsa produces darkly stained (TB) and lightly stained (BB) sister chromatids [30] (Fig. 9.9).

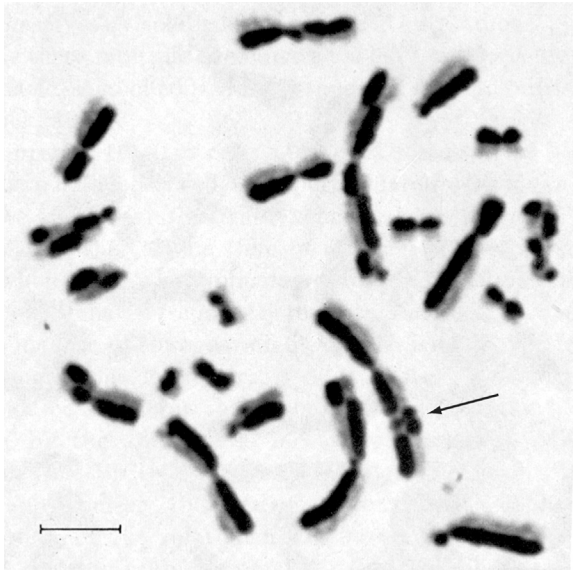


Figure 9.9 Sister chromatid exchanges shown in Chinese hamster ovary cells.

Therefore, exchanges of material between sister chromatids are readily visible at high resolution following this staining protocol. The differential sister chromatid staining observed following the SCE protocol is a remarkable cytologic demonstration of the semiconservative replication of DNA. It also demonstrates that each chromosome is composed of a single very long duplex of DNA. Further, it shows that exchanges between the two sister chromatids take place in somatic cells that could potentially have mutagenic effects. SCE is used to diagnose diseases associated with chromosomal instability in clinical cytogenetics laboratories. For example, SCE analysis is a diagnostic test for Bloom syndrome, a rare autosomal recessive disease caused by mutations in a DNA helicase of the RecQ family that catalyze the unwinding of duplex nucleic acid molecules [31]. It is characterized by growth deficiency, predisposition to neoplasia, and chromosomal instability in somatic cells. The frequency of spontaneous SCEs in cells from patients with Bloom syndrome is markedly increased. SCE analysis is also used to monitor the effects of potentially mutagenic or carcinogenic agents that enhance the rate of SCEs.

9.4.3 Chromosome Banding Reveals Genome Sequence Organization

Quinacrine associates directly with DNA by intercalating between base pairs. Although quinacrine binds

equally well to DNA of any base composition, its fluorescence is enhanced in regions containing uninterrupted runs of AT base pairs, and is quenched in regions with more frequent GC base pairs. In the Q-banding pattern of human chromosomes (see Fig. 9.5), the intensity of fluorescence is generally proportional to the AT richness of the DNA [32]. However, the highly AT-rich satellite DNA that is concentrated at the C-bands of chromosomes 1, 9, and 16 has interspersed GC base pairs and usually fails to show bright Q-banding. That on the Y, in contrast, has no such GC pairs and is intensely fluorescent. Thus Q-banding is related to both base composition and base interspersions, which result in the differential fluorescence or quenching of signals produced by the fluorescent dye. DNA-protein interactions may also be important in the generation of Q-bands.

G-banding is produced most commonly by treatment of chromosomal preparations with the proteolytic enzyme trypsin. Giemsa stains DNA primarily by intercalating between adjacent base pairs in double-stranded regions. G-bands result from the degradation of chromosomal proteins by trypsin, which modifies the interaction of chromosomal DNA with the Giemsa dyes. Since the fixative used in standard chromosome preparation methods, methanol:acetic acid (3:1), removes some of the histone proteins, it is the degradation of the nonhistone proteins that appears to be critical for the production of G-bands. The DNA-protein interactions at the G-band-positive regions apparently render these sites resistant to denaturation by the enzyme.

The commonly used method to generate R-bands is to subject chromosome preparations to moderate heat (~85°C in the presence of high salt) before staining them with Giemsa. The heat pretreatment is thought to selectively denature the more AT-rich DNA sequences, which have a lower thermal stability than GC base pairs, and to result in altered DNA structure on renaturation. Therefore, after chromosomes are exposed to moderate heat, Giemsa stains the unaffected GC-rich double-stranded DNA regions, producing R-banding. R-bands can also be produced by the replication banding technique, which demonstrates that R-band-positive regions contain early-replicating DNA. It also follows that G-band- and Q-band-positive regions contain AT-rich DNA that replicates relatively late in the cell cycle [32].

C-band-positive regions have been found by *in situ* hybridization and DNA sequencing to consist of α -satellite (discussed later) sequences at the

TABLE 9.1 Characteristics of Chromosome Bands

Characteristic	Q- or G-Bands	R-Bands	C-Bands
Location	Chromosome arms	Chromosome arms	Centromeres, distal Yq
Type of DNA sequence	Repetitive, some unique	Unique, some repetitive	Highly repetitive satellite
Base composition	ATrich	GCrich	ATrich, some GCrich
5-Methylcytosine content	Low	Moderate	High
Type of chromatin	Heterochromatin	Euchromatin	Heterochromatin
Replication	Middle to late S phase	Early S phase	Late S phase
Transcription	Low	High	Absent
Gene density	Low	High	Absent
CpG-rich islands	Few	Many	Absent
Repeats	LINE-rich	SINE-rich	—
Acetylated histones	Low	High	Absent

LINE, long interspersed nuclear element; *SINE*, short interspersed nuclear element.

centromeres of human chromosomes and of different families of simple-sequence satellite DNAs at the large pericentromeric C-band blocks on chromosomes 1, 9, and 16 and distal Yq. Analyses of the completed human genome sequence have defined further families of repetitive DNA [33], but these have not yet been associated with functional or structural landmarks of chromosomes. In contrast, studies employing in situ hybridization as well as in silico analyses of the genome sequence have revealed that the human genome also includes highly homologous duplications of DNA ranging in size from 1 to more than 500 kb. These repeats, called segmental duplications, are located mainly in the pericentromeric and subtelomeric regions of chromosomes, although they are also present as interspersed repeats along the length of the chromosome [33,34]. While some of these segmental duplications are known to predispose to genomic deletions and duplications, their significance for chromosomal function is otherwise unknown. A comparison of the characteristics of Q-/G-, R-, and C-bands is presented in Table 9.1.

Also related to simple sequences are chromosomal regions called fragile sites that remain stretched at metaphase after various treatments that limit DNA replication [35]. Fragile sites are classified as rare (inherited) or common (constitutional) and are further subdivided according to the conditions under which they are induced (e.g., folate or aphidicolin sensitive). Several fragile sites have now been cloned and sequenced. These studies have shown that the expression of rare, inherited fragile sites is associated with repeat expansions

[35]. The first folate-sensitive rare fragile site to be characterized was the one associated with the fragile X syndrome (FMR1), which was shown to result from the expansion and methylation of a CGG trinucleotide repeat in the 5'untranslated region of the *FMR1* gene. Other folate-sensitive fragile sites characterized thus far also result from the expansion of trinucleotide repeats [36]. A distamycin-sensitive rare fragile site on chromosome 16 has been shown to involve the expansion of a 33-bp AT-rich minisatellite [36]. In contrast, sequencing of constitutional fragile sites has not revealed any characteristic DNA sequences at these sites [37].

9.4.4 Molecular Cytogenetics

The gap between light microscope resolution of chromosome structure and the gene was bridged by the introduction of several molecular cytogenetic techniques. Fluorescence in situ hybridization (FISH) involves hybridizing a fluorescently labeled single-stranded DNA probe to denatured chromosomal DNA on a microscope slide preparation of metaphase chromosomes and/or interphase nuclei prepared from the patient's sample. After overnight hybridization, the slide is washed and counterstained with a nucleic acid dye (e.g., DAPI), allowing the region where hybridization has occurred to be visualized using a fluorescence microscope [38]. FISH is now widely used for clinical diagnostic purposes. There are different types of FISH probes, including locus-specific probes, centromeric probes (CEPs), and whole-chromosome paint probes. Locus-specific probes are specific for a specific single locus. They are particularly useful for identifying subtle submicroscopic

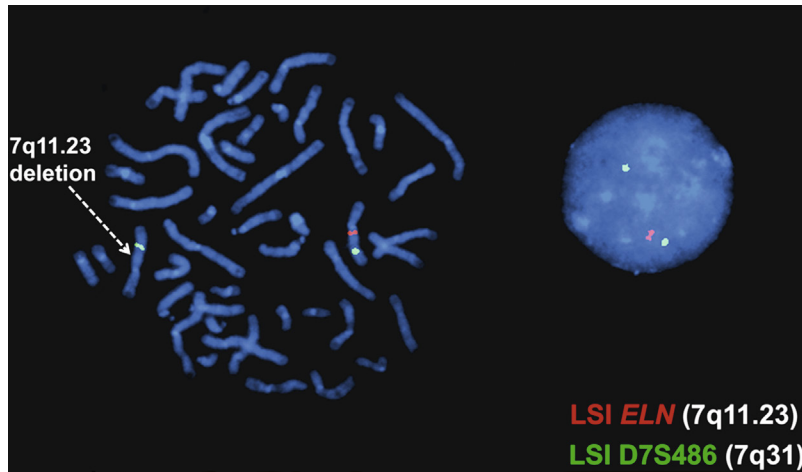


Figure 9.10 Metaphase and interphase fluorescence in situ hybridization analysis in a patient with William syndrome due to deletion on chromosome 7 band q11.23. Note the deletion of the *ELN* gene probe labeled in red.

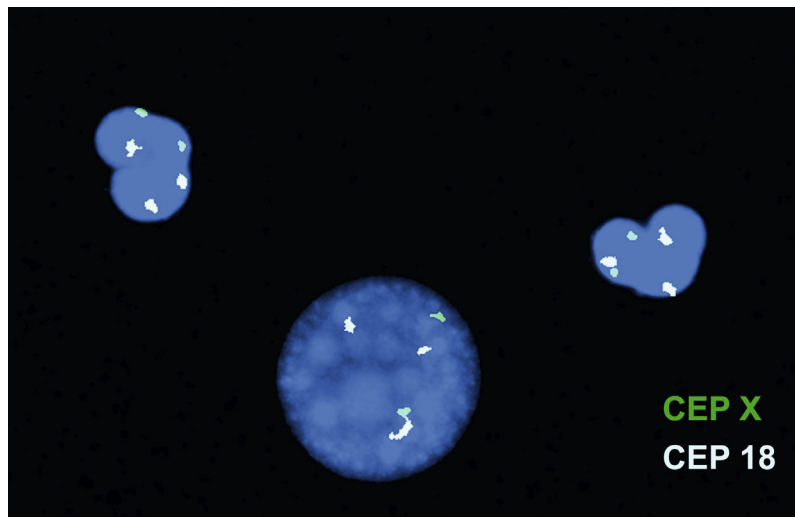


Figure 9.11 Interphase fluorescence in situ hybridization analysis in a patient with trisomy of chromosome 18. Note the three copies of the chromosome 18 centromeric probe labeled in *aqua*.

deletions and duplications (Fig. 9.10). CEPs are specific for unique repetitive DNA sequences (e.g., α -satellite sequences) in the centromere of a specific chromosome. They are suitable for making a rapid diagnosis of one of the common aneuploidy syndromes (trisomies 13, 18, and 21, and sex chromosome aneuploidies) using non-dividing interphase nuclei. This is particularly useful in a prenatal setting using amniotic fluid or chorionic villi samples (Fig. 9.11). Whole-chromosome paint probes consist of a cocktail of probes obtained from different

regions of a particular chromosome. When this cocktail mixture is used in a single hybridization, the entire relevant chromosome fluoresces (is “painted”) (Fig. 9.12). Whole-chromosome paints are useful for characterizing complex chromosomal rearrangements, and for identifying the origin of additional chromosomal material such as small marker or ring chromosomes.

FISH using locus-specific probes has been extremely useful in the detection of “microdeletion syndromes” resulting from deletions of multiple contiguous genes.

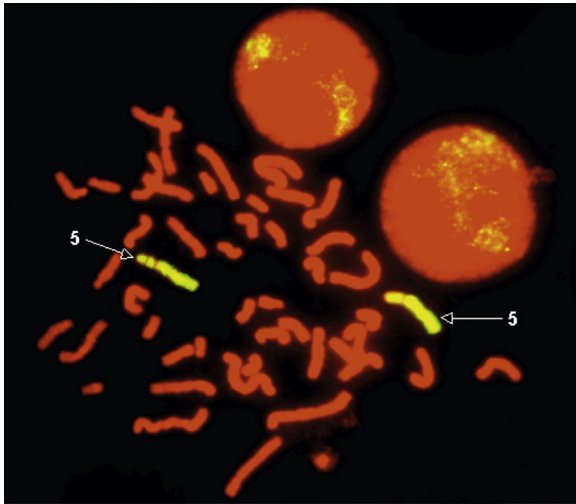


Figure 9.12 Metaphase fluorescence in situ hybridization analysis using a chromosome 5 paint probe.

These are subtle submicroscopic deletions that are below the resolution of the routine G-banded chromosome analysis. Also, two-color and three-color FISH applications are routinely used to diagnose specific deletions, duplications, or other rearrangements, both in metaphase chromosomes and in interphase nuclei. Use of FISH usually requires that the patient either exhibits features consistent with a well-defined syndrome with known chromosomal etiology or demonstrates an abnormal karyotype. This is because single FISH probes reveal rearrangements only of the segments being interrogated, and do not provide information about the rest of the genome. Another limitation of FISH is the number of probes that can be applied in a simultaneous assay. FISH techniques have been developed utilizing pools of whole-chromosome paint probes for every chromosome to provide a multicolor human karyotype in which each pair of homologous chromosomes can be identified on the basis of its unique color when studied using special computer-based image analysis software (spectral karyotyping and multiplex-FISH) [39].

One type of FISH that has the potential to reveal chromosomal imbalances across the genome is comparative genomic hybridization (CGH). In CGH, DNA specimens from the patient and a normal control are differentially labeled with two different fluorescent dyes and hybridized to normal metaphase chromosome spreads. Differences between the fluorescence intensities of the two dyes along the length of any given chromosome will

reveal gains and losses of genomic segments [40]. The limitations of this technology include many of the same limitations of G-banded chromosome analysis. Thus, like G-bands, the resolution of CGH is limited to that of metaphase chromosomes, which is approximately 5 megabases (Mb) for most clinical applications [39].

The latest addition to molecular cytogenetic techniques is chromosomal microarray analysis (CMA), which is also known as cytogenomic array analysis. High-resolution whole-genome coverage CMA platforms have been increasingly used in clinical laboratories. They provide a relatively quick method to scan the entire human genome for copy number variants (CNVs), both gains and losses, with significantly higher resolution and greater clinical abnormality yield than was previously possible. This led to the identification of novel genomic disorders in patients with developmental delay, intellectual disability, autism spectrum disorders, and/or multiple congenital anomalies [44,94]. CMA platforms include two main technologies, namely array CGH and single-nucleotide polymorphism (SNP) array. Array CGH involves cohybridizing a test (patient) DNA sample and a control (reference) DNA sample, each differentially labeled with different colored fluorescent dyes (usually red and green, respectively), to a microarray slide containing thousands of DNA probes (oligos) that cover the entire human genome. The difference in hybridization between the two samples for each probe, as measured by the ratio of fluorescence of the two fluorescent dyes, is used to detect CNVs of genomic regions represented on the array (Fig. 9.13) [41–43]. Another approach to CMA is the high-density SNP arrays. In SNP array analysis, only the test (patient) DNA sample is labeled with a fluorescent dye and hybridized to the array. CNVs are detected by measuring the absolute fluorescent signal intensity of the two SNP alleles of a SNP and spanning several adjacent SNPs, and comparing with the fluorescent signal intensities of multiple normal controls that are separately hybridized and analyzed (in silico analysis). In addition to their ability to detect CNVs, the genotype data detected by SNP array analysis can be used to identify regions of homozygosity (ROHs) that are larger than 5 Mb. These data are plotted in a separate allele analysis track and can be suggestive of either uniparental disomy (UPD; when the homozygosity involves only one chromosome) or regions that are identical by descent (when the homozygosity involves several chromosomes) due to consanguinity. UPD

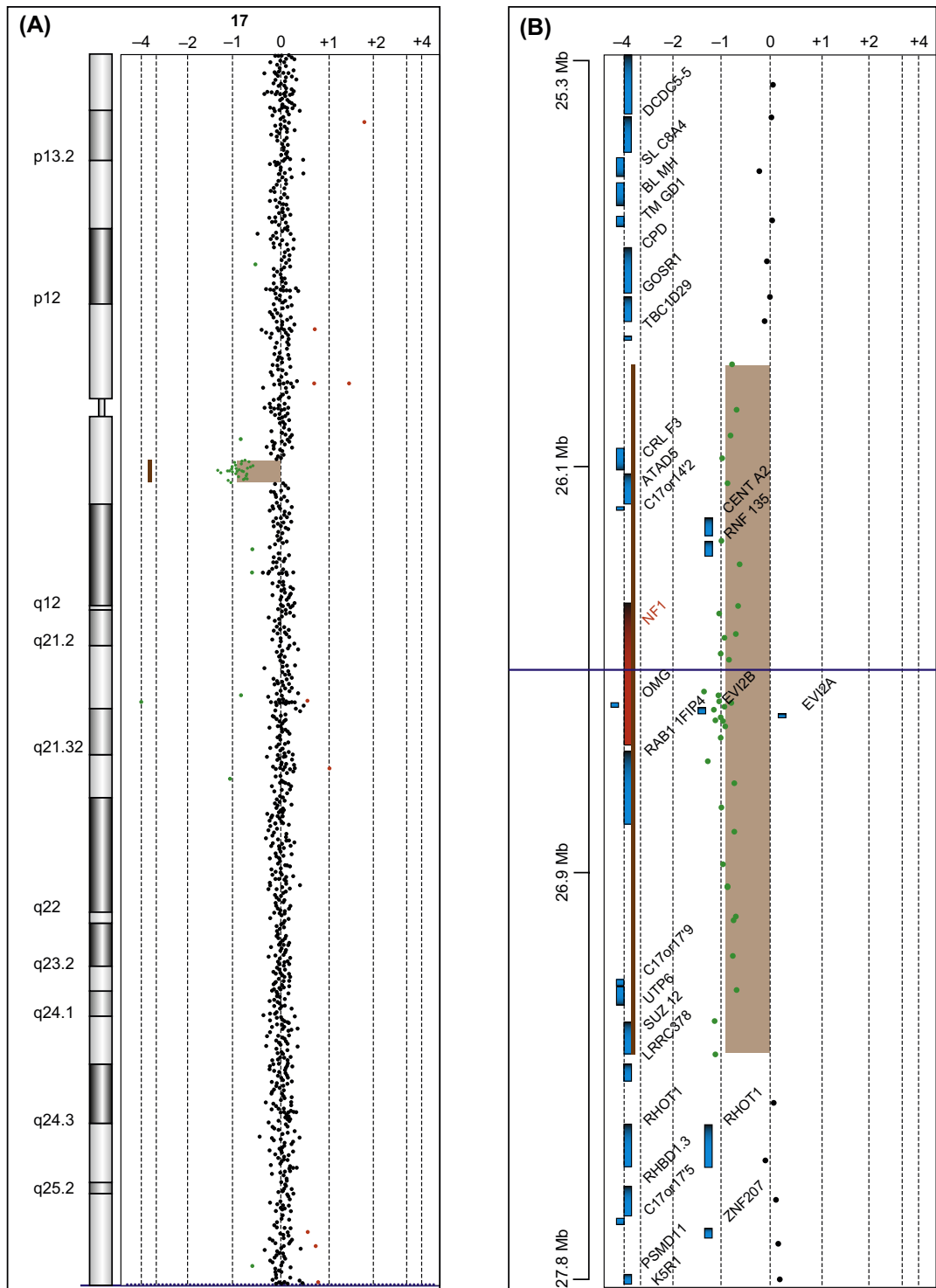


Figure 9.13 Oligoarray comparative genomic hybridization analysis in a patient with *NF1* gene deletion on chromosome 17 band q11.2. (A) Whole chromosome 17. (B) Deleted region. Shaded areas show the deleted region.

in this case requires a separate chromosome-specific confirmatory testing. ROHs can harbor homozygous mutations in autosomal recessive genes, and follow-up sequencing is usually indicated when a recessive condition is suspected. SNP analysis can also allow the detection of polyploidy (triploidy and tetraploidy) in prenatal and postnatal neonatal settings, which is an advantage over the use of array CGH. In recent years, SNP analysis has been added to array CGH platforms.

CMA technologies have two main limitations, namely, their inability to detect balanced chromosomal rearrangements and low-level mosaicism. Several CMA studies have reported finding CNVs at the breakpoints of some *de novo* apparently balanced rearrangements detected by karyotyping [96,97]. In recent years, next-generation sequencing technologies have been rapidly emerging. Whole-genome sequencing (WGS) is a high-resolution methodology that has the potential to eventually replace some cytogenetic techniques. In addition to the detection of sequence variants, WGS is capable of detecting structural chromosomal abnormalities, including CNVs and balanced chromosomal rearrangements. Many groups are evaluating the accuracy of WGS for detecting CNVs of varying sizes and are developing new analysis algorithms based on read depth of coverage to enhance CNV calling capabilities. Assessment of these next-generation sequencing-based technologies compared with CMA-based technologies for CNV detection will provide an opportunity to evaluate which approach can provide the most accurate high-resolution data for routine clinical testing.

9.5 FUNCTIONAL ORGANIZATION OF CHROMOSOMES

Chromatin is classified into euchromatin and heterochromatin. Euchromatin consists of active genes; however, not all genes in euchromatic regions are active at any given time. Therefore, localization in euchromatin is currently thought to be necessary but not sufficient for gene activity. Euchromatin is dispersed in the interphase nucleus and replicates its DNA early in the S phase of the cell cycle. Heterochromatin consists predominantly of inactive genetic material, replicates its DNA late in the S phase, and is condensed in the interphase nucleus. Heterochromatin is further classified into constitutive heterochromatin and facultative heterochromatin. Constitutive heterochromatin consists of highly repetitive

simple-sequence DNA, remains transcriptionally inactive, and is located at specific regions of the chromosomes such as the centromere and the distal long arm of the human Y chromosome. Facultative heterochromatin also remains condensed in the interphase nucleus, replicates its DNA late in the S phase, and is largely transcriptionally inactive; however, it is not inactive permanently, does not consist exclusively of repetitive DNA, and can become transcriptionally active. The inactive X chromosome in the human female is a good example of facultative heterochromatin. However, localization in facultative heterochromatin does not exclude transcription altogether, as several genes on the inactive X chromosome are expressed (see later). As already noted, the R-band-positive regions of human chromosomes have characteristics of euchromatin in that they replicate their DNA in early S phase and have high transcriptional activity due in part to high gene density (see Table 9.1). The G-band-positive regions, on the other hand, are more heterochromatic, as they replicate their DNA in late S phase and are low in transcriptional activity associated with low gene density. Integration of the whole human genome sequence with the cytogenetic map shows a lower density of genes in G-positive bands [45]. The C-band-positive regions consist of constitutive heterochromatin with no known functional genes. The facultative heterochromatin of the inactive X chromosome replicates its DNA in late S phase, and forms the condensed Barr body in the interphase nucleus. Consequently, there is a general relationship of functional properties (time of replication during the S phase and transcriptional status or gene density) with chromosome band classes characterized by differential condensation and staining characteristics [46].

Investigations have provided insights into the molecular organization of two specialized structures on chromosomes, the centromere and the telomere, which are summarized below.

9.5.1 The Centromere

As already noted, each chromosome has a primary constriction, the centromere, where the sister chromatids of a replicated chromosome are held together until the anaphase stage of cell division. A subdomain of the centromere is the kinetochore, a protein–DNA complex that serves as the attachment site for the spindle fibers essential for chromosome movement and segregation during mitosis and meiosis. The structure of the

centromere has been a focus of molecular cytogenetic investigations in recent years. The best characterized eukaryotic centromere is that of the budding yeast *Saccharomyces cerevisiae*. In this organism, a short sequence of about 125 bp specifies the centromere of each of the chromosomes. The nucleotide sequence and organization of this centromere DNA are conserved among the different chromosomes in the budding yeast. The search for a similar specific sequence in the larger and more complex centromeres of higher eukaryotes has not been successful. Rather, the centromeres in these organisms consist of large arrays of repeated α -satellite DNA sequences. In human centromeres, the arrays consist of tandem, head-to-tail repeats of a 171-bp monomer that is further organized into higher order repeats [47]. The centromeric chromatin of human chromosomes spans from 0.1 to 4.0 Mb. The sequence of the basic 171-bp unit is sufficiently divergent among human chromosomes that, with very few exceptions, centromere-specific α -satellite DNA probes can generate fluorescent signals on specific chromosomes in a FISH assay. This is useful from a practical standpoint for identifying and determining the copy number for specific human chromosomes in interphase cells.

Several lines of evidence implicate a critical role for α -satellite DNA in centromere function. Although there are other repeated sequences in the centromeric heterochromatin, α -satellite is the only one localized to the centromeres of all normal human chromosomes. Moreover, studies have shown that human artificial chromosome constructs containing α -satellite DNA are able to form functional centromeres [48]. However, independent evidence from rearranged chromosomes suggests that the presence of α -satellite DNA alone is not sufficient for the formation of an active centromere. Many cases of rearranged human chromosomes containing two centromeric regions have been identified. A true dicentric chromosome with two primary constrictions would be unstable during cell division as spindle fiber attachment occurs independently at the two centromeres, if these are sufficiently far apart. The two centromeres on a single chromatid could then be pulled toward opposite poles of the spindle, breaking the chromosome. However, many dicentric chromosomes with two blocks of α -satellite DNA and C-band regions are stable and show only one primary constriction, indicating that only one of the two centromeres is active. Such stable dicentric chromosomes, referred to as pseudodicentrics, indicate that

the presence of α -satellite DNA alone is not sufficient for the formation of an active centromere. In addition, several human marker chromosomes have been characterized that originate from normal human chromosomes but lack α -satellite DNA sequences. These functional centromeres lacking α -satellite DNA are called neocentromeres [49]. As these chromosomes are mitotically stable, the presence of α -satellite DNA is not an absolute requirement for functional centromeres. Thus, although normal human centromeres are composed of α -satellite DNA, it appears to be neither necessary nor sufficient for centromere formation.

Investigations have identified several proteins associated with centromeres that have contributed to our understanding of centromere structure and function [50,51]. A group of these proteins are constitutively associated with centromeres, while others are associated with centromeres only during a part of the cell cycle and are involved in chromosome movement during cell division. The major constitutive centromere proteins identified are CENP-A, CENP-B, and CENP-C. The localization of these proteins at centromeres has been determined by immunofluorescence microscopy using antibodies specific for these proteins. CENP-A is a 17-kDa histone H3-like protein that participates in producing centromere-specific nucleosomes (in place of histone H3) and altered chromatin structure. CENP-A is detected at all functional centromeres, including the neocentromeres. CENP-B is an 80-kDa protein that binds to a specific 17-bp sequence, the CENP-B box, in α -satellite DNA and is found, as expected, even at the inactive centromere of pseudodicentric chromosomes. CENP-C, a 140-kDa protein, is also found at active centromeres, where it is located in the proteinaceous kinetochore. CENP-C shares homology with a domain of the Mif2 protein of yeast that is essential for normal chromosome segregation. In addition to the CENP-A, -B, and -C proteins that associate with centromeres constitutively, many more that associate transiently during cell division have been identified. An example of the latter class of proteins is CENP-E, a 275-kDa kinesin-related protein that is associated with centromeres and the mitotic spindle during mitosis and plays a role in chromosome movement.

9.5.2 The Telomere

Telomeres are special DNA-protein structures that are present at the ends of linear chromosomes and prevent

fusion of chromosome ends and maintain chromosome integrity. The concept of the telomere was developed from early genetic and cytologic observations that the broken ends of chromosomes are unstable and often fuse with other broken ends. Molecular techniques have now shown that telomeres in eukaryotes exist in a DNA-protein complex consisting of tandem repeats of a simple sequence and a number of proteins. In humans and other vertebrates, the sequence of the basic repeat is 5'-TTAGGG-3' on one strand of the DNA and 5'-CCCTAA-3' on the complementary strand. The G-rich strand runs 5'-3' toward the end of the chromosome, with a short, single-stranded, G-rich overhang [52,53]. The human telomeric sequence typically spans about 2–50 kb and is replicated by its own polymerase, called telomerase. In the absence of telomerase, each round of DNA replication leaves 50–200 bp of DNA unreplicated at the 3' end, as the DNA replication machinery works only in the 5'–3' direction and requires an RNA primer. This would result in the loss of sequences from the ends of chromosomes, ultimately leading to loss of genetic material.

Telomerase is a ribonucleoprotein complex that includes a reverse transcriptase and a short RNA molecule that provides the template for synthesizing the telomeric sequences. By copying the RNA template, telomerase extends the G-rich telomeric DNA strand running 5'–3' toward the distal end of the chromosome. The complementary strand is then synthesized by the cellular DNA replication machinery through lagging-strand synthesis. Telomere-associated proteins regulate telomerase activity so that the length of the telomere repeat tract is maintained at a level required for maintaining functional telomeres [53,54]. Telomerase is present in early embryonic cells and in the majority of immortalized cells, but not in most somatic cells. As a result, somatic cells, but not cancer cells, lose telomeric sequences with each division, leading to dysfunctional telomeres and excessive chromosomal instability. Telomerase activity is therefore considered to be a critical factor contributing to the finite life span of most somatic cells and indefinite growth potential of cancer cells. Studies have shown that telomere sequences can be added to the ends of chromosomes with terminal deletions, thus stabilizing these broken ends. Healing of broken ends can occur through two general pathways, ensuring the acquisition of a new telomeric cap and stabilizing the deleted chromosome. First, direct addition of telomeric sequences onto the

broken end can be achieved through a telomerase-mediated de novo telomere addition [55,56] or a telomerase-independent recombination-based mechanism [57,58]. Second, telomeres can also be retrieved from another location through a mechanism called telomere capture, in which subtelomeres and/or pantelomeres from another chromosome are translocated to the broken end of the deleted chromosome [59,60].

Adjacent to the human terminal (TTAGGG)_n repeat is a complex region of segmentally duplicated DNA tracts generally referred to as subtelomeric repeat DNA or telomere-associated repeats (TARs). This class of low-copy repeat DNA is characterized by very high sequence similarity (>90%) between duplicated tracts and variably sized but often very large duplicated segments. Some of the segmental duplications are unique to TARs, some are shared with a subset of pericentromeric repeat regions, and some are shared with one or several interstitial chromosomal loci. These TARs range in size from 100 to 300 kb, and just proximal to these regions the unique subtelomeric sequences are encountered [61].

9.6 SEX CHROMOSOMES AND SEX DETERMINATION

Sex chromosomes of the human chromosome complement are of special interest, as they determine the sex of the human embryo. Also, the sex chromosome pair, the X and Y, is heteromorphic (different in size and morphology) in humans. The Y chromosome is significantly smaller than the X chromosome and contains a large block of heterochromatin on its q arm comprising noncoding repetitive DNA. This leaves only a short segment of chromosome capable of carrying functional genes. The finding of heteromorphic sex chromosomes in humans, an XX pair in females and an XY pair in males, suggested a chromosomal basis for sex determination in the early 20th century. The dominant role of the Y chromosome in male sex determination became evident only in 1959, when cytogenetic studies showed that individuals with a complete set of autosomes and a single X chromosome developed as females, whereas individuals with two X chromosomes and a Y chromosome developed as males [62,63]. We now know that individuals with as many as four X chromosomes and a Y also develop as males. The number of X chromosomes or its ratio to the number of autosomes is not important for human male sex determination. The Y chromosome thus carries a dominant determinant for testis development.

Unlike autosomal pairs of chromosomes, the heteromorphic X and Y are not completely homologous. There are two regions of complete homology between the X and the Y chromosomes that reside at the distal ends of their short and long arms, covering approximately 2600 and 320 kb of DNA, respectively [64]. The X and Y chromosomes pair and cross over in these regions during prophase I. This appears to be essential for correct segregation of the sex chromosomes. As a result of this crossing over, female offspring of males can inherit DNA sequences from the Y chromosome distal to the point of exchange and vice versa. Thus, genetic markers in this region of pairing and exchange between the X and the Y segregate independent of sexual phenotype, and hence this region is called the pseudoautosomal region.

9.6.1 The Y Chromosome and Sex Determination

Identification of the testis-determining factor (TDF) on the human Y chromosome has been of much interest since the role of the Y in male sex determination was established. Early cytogenetic investigations in individuals with structurally abnormal Y chromosomes showed that the TDF resided on the p arm of the Y chromosome. The isolation and molecular characterization of this gene was made possible by studies of a naturally occurring sex-reversed condition, the XX male. Cytogenetic and molecular investigations of XX males showed that the majority of them resulted from an unequal exchange between X and Y, such that the TDF is transferred from the Y to the X chromosome. By identifying the minimal region of Y necessary for male determination from independent XX males, and searching this region for candidate genes, the *SRY* (sex-determining region on Y) gene was identified. Later studies confirmed that *SRY* is the long-sought TDF [65]. *SRY*, which resides just proximal to the pseudoautosomal region on the p arm of the Y chromosome, encodes a protein of 240 amino acids, which is capable of sequence-specific binding to DNA using a motif known as the HMG box [66,67]. This motif is found in several classes of DNA-binding proteins, including several that are known to be transcription factors. Unlike other transcription factors, the *SRY* protein does not contain any other recognizable motifs, and this has led to the hypothesis that it functions partly as a scaffold protein in chromatin [68]. *SRY* induces the differentiation of Sertoli cells in the developing gonad. Sertoli cells produce anti-Müllerian hormone, which

causes regression of the female internal genitalia; they also induce Leydig cells to secrete the androgens necessary for the development of male internal and external genitalia [65]. Any genetic or environmental factor that prevents testis differentiation in 46,XY embryos leads to the development of a sex-reversed XY female. Molecular dissection of other conditions that result in sex reversal has allowed the identification of some of the other genes involved in the sex-determining pathway [69–71]. Not surprisingly, many of these are autosomal and not sex-linked genes.

9.6.2 The X Chromosome

The heteromorphic nature of the sex chromosome pair in humans immediately raises the question of dosage difference for X-linked genes in the human male and female. The answer to this question was provided by observations on the behavior of the two X chromosomes and the expression patterns of X-linked genes in human and other mammalian females. These findings indicated that there were differences in the functional organization of the two X chromosomes in mammalian females. Early cytologic studies demonstrated that a sex chromatin body (called the Barr body) was present in female interphase cells, but not in male cells. Moreover, the amount of certain X-linked gene products, such as the enzyme glucose-6-phosphate dehydrogenase, was no different in individuals with one, two, or even more X chromosomes. Also, studies on the timing of DNA replication in diploid cells indicated that the DNA in one X chromosome replicated in synchrony with the DNA of the autosomes, while that of any additional X chromosomes replicated late in the S phase. Thus, the number of Barr bodies equals the number of late-replicating X chromosomes.

Based on the observation that female mice heterozygous for X-linked genes show mosaicism for the expression of these genes, Lyon in 1961 proposed the single active X hypothesis [72], which offers an explanation for the gender-specific behavior of the X chromosome and X-linked genes. According to this hypothesis, commonly referred to as the Lyon hypothesis, the somatic cells of all mammals undergo a process of chromosome differentiation early in embryogenesis that leaves a single active X chromosome per cell. All additional X chromosomes are inactivated by a process that renders them heterochromatic and capable of forming a Barr body. Thus, diploid somatic cells of individuals with three

X chromosomes have two Barr bodies, while those of individuals with four X chromosomes have three Barr bodies. The initial choice for inactivation of an X is random in a normal female. However, this differentiation is fixed, so that all the descendants of a cell in which the maternal X was inactivated initially will have the maternal X in the inactive state, while the descendants of a cell in which the paternal X was inactivated will have that X in the inactive state. Every XX individual is thus a genetic mosaic consisting of cells in which the maternal X is active and cells in which the paternal X is active.

The phenomenon of X-chromosome inactivation has been a subject of much interest and investigation in mammalian biology. This interest in X-inactivation derives from the fact that it is a relatively unique epigenetic process of gene regulation at the level of the chromosome. It is epigenetic because the inactivated X chromosome does not undergo any permanent changes in its DNA sequence and can be reactivated, as it is in female germ cells. Genes from the inactive X chromosome can also be reactivated experimentally in cultured cells [73]. Further, attention has been focused on X-inactivation as a means of understanding broader aspects of chromatin, specifically the structure and function of facultative heterochromatin. Finally, the burden of human X-linked diseases and X-chromosome abnormalities has generated an interest in X-inactivation for a better understanding of the pathogenesis and ultimately the treatment of these conditions.

Investigations since the late 1990s have provided insights into the molecular mechanism of X-chromosome inactivation. It is now clear that DNA methylation plays a key role in maintaining the X in the inactive state. Studies of several genes have shown that cytosine residues in cytosine–guanine dinucleotides (CpG) in their 5' promoter regions are methylated when they reside on an inactive X and unmethylated on an active X [74]. The binding of proteins that specifically bind methylated DNA and inhibit transcription could account for the transcriptional silencing of genes on the inactivated X. Studies employing immunofluorescent labeling of human metaphase chromosomes with antibodies specific for acetylated isoforms of nucleosome core histones have shown that the inactive X chromosome is hypoacetylated, linking methylation and histone acetylation in the control of gene expression from the inactive X [75,76]. Examination of the histone H4 acetylation status at the individual gene level has also

shown hypoacetylation in the promoter regions of X-inactivated genes [77]. Hypermethylation of DNA, hypoacetylation of histones, and methylation of histone H3 at lysine 9 (H3–mK9) are features common to all heterochromatin [78]. Initiation of X-inactivation, which must also include counting the number of X's in a cell and the spreading of inactivation along the X, is still not completely understood. However, these early events in X-inactivation are dependent on the X-inactivation center (XIC), a complex specialized control locus located in the proximal q arm of the human X chromosome. The XIC is required for the initiation of X-inactivation and is invariably present on all X chromosomes that undergo inactivation, including those with structural rearrangements. A search for candidate genes mapping to the XIC region led to identification of the *XIST* (X-inactive specific transcript) gene. The *XIST* gene, expressed exclusively from the inactive X and not from the active X, is located at the XIC [79]. The product of *XIST* is a large noncoding RNA molecule that stays associated with the inactive X [80]. Transgenic and knockout experiments indicate that *XIST* is necessary and sufficient for initiating X-inactivation [81–83]. While the precise mechanism(s) of X-chromosome inactivation remain to be revealed, the process is generally described in four stages: recognition of the number of X chromosomes (also called “counting”), initiation early in development, promulgation whereby the initial signal is spread to the rest of the chromosome, and maintenance of the inactivating signal through successive cell divisions [84].

It is now well established that not all genes on the X chromosome are subject to X-inactivation. Early studies showed that the genes for the Xg blood group and for the enzyme steroid sulfatase (deficiency of which causes X-linked ichthyosis) escape X-inactivation [64,84]. More recent studies evaluating an estimated 95% of X-linked genes assayable in cell culture systems show that about 15% of these genes escape inactivation, and an additional 10% show variable levels of expression from the inactive X chromosome [80,85]. The majority of the genes that escape inactivation are located on the p arm of the X chromosome, but they are also present on the q arm and are interspersed with genes that undergo inactivation. As expected, the genes in the pseudoautosomal regions escape X-inactivation; these have homologues on the Y chromosome, and dosage compensation is not a requirement for these genes. However, there are genes on the X that escape X-inactivation for which there is

no functional homologue on the Y, thus resulting in an increased dose in the female. These differences between the X and the Y reflect the evolutionary history of the sex chromosomes. It is thought that the heteromorphic sex chromosomes of today evolved from a homomorphic autosome-like pair with progressive loss of genes from the Y and incorporation of the corresponding genes on the X into the X-inactivation system. The genes that escape inactivation on the X may be essential for normal female development in two doses or they may have simply failed to be incorporated into the X-inactivation system with no adverse consequences. The abnormal development associated with X-chromosome aneuploidy is most readily explained by dosage inequities in these genes that escape inactivation. Identification of the specific genes involved in these diseases is, therefore, of great interest and is a focus of ongoing investigations.

9.7 UNIPARENTAL DISOMY AND IMPRINTING

It has been appreciated for some time that one paternal and one maternal set of chromosomes are required for the normal development of the embryo. In rare cases, a pregnancy arises in which an ovum undergoes some degree of embryonic development by a process of gynogenesis or androgenesis; that is, the cells are solely of maternal origin (gynogenesis) or of paternal origin (androgenesis). Ovarian teratomas appear to be the result of gynogenetic development of ova that have not undergone the second meiotic division. The cells are thus diploid and XX. In contrast, some pregnancies, which terminate in spontaneous abortion, are associated with the presence of a hydatidiform mole, an abnormal development of extraembryonic tissue. Many of these moles are diploid and XX (or rarely, XY), with both sets of chromosomes of paternal origin. They may arise as a result of fertilization of an anucleated ovum by two sperms.

One of the most interesting novel concepts to emerge from these and other experimental studies in the mouse is that of genomic imprinting, which provides an explanation for the abnormal development of gynogenetic and androgenetic embryos. Genomic imprinting refers to a process by which maternal and paternal alleles of specific genes or chromosomal regions are differentially marked during gametogenesis such that they are expressed differently in the embryo [86]. One allele of

the imprinted gene is usually active, while the other is inactive. Thus, the paternal and maternal copies are not functionally equal for all genes, and therefore both a maternal copy and a paternal copy are required for normal development. Like X-chromosome inactivation, genomic imprinting is also an epigenetic phenomenon in that the imprinted gene does not undergo any permanent change and the imprint is reversible. Thus, a female who begins life with a maternally and paternally imprinted allele at a locus will produce gametes that exhibit only maternal imprint even on her own paternal chromosome. Similarly, males produce only gametes with the male-specific imprint. In other words, during gametogenesis the parental imprint is erased and reset in a sex-specific manner. As in X-chromosome inactivation, DNA methylation is a mediator of the maintenance of the imprint in the somatic cells.

Imprinting is known to affect only a small number of genes and chromosomal regions in the human genome. Imprinting thus differs from X-inactivation in that it does not affect a whole or most of a chromosome. Moreover, even within an imprinted chromosomal region, individual genes located within a few hundred kilobases of DNA may show differential imprinting. As a result, one gene may be inactive on the maternal chromosome and active on the paternal chromosome while a neighboring gene exhibits the opposite imprinting, being active on the maternal chromosome and inactive on the paternal chromosome. Imprinting also shows tissue-specific variation for certain genes. Thus, the Angelman syndrome gene, *UBE3A*, on chromosome 15 is expressed from both chromosomes (biallelic expression) in somatic cells but is expressed only from the maternal chromosome in the brain. X-chromosome inactivation in females is different from imprinting in this regard in that it is presumed to be present in all somatic cells.

Although imprinting affects only a few chromosomal regions, imprinted genes contribute to genetic diseases. The phenotypes exhibited by moles and teratomas are the result of failure to receive either the maternal (mole) or the paternal (teratoma) genome. Other phenotypes result from failure to receive specific portions of the maternal or paternal genome, or inappropriately receiving two copies of the same chromosome region from one parent and none from the other parent (UPD) [87]. A fraction of cases of Prader-Willi syndrome (PWS), Angelman syndrome, and Beckwith-Wiedemann syndrome result from such imbalances in the parental

origin-dependent portion of a chromosome region. In the case of PWS, about 70% of the patients have a deletion in the proximal q arm of the paternally inherited chromosome 15. In normal individuals, the PWS critical gene(s) is transcribed only from the paternal homologue. Therefore, with the deletion of the PWS critical region on the paternal 15, PWS patients are completely deficient for the products of these imprinted genes. The remaining 30% of PWS patients have two chromosomes 15 derived from their mother and none from their father. In the absence of a paternal 15, these patients also lack the expression of the PWS critical gene(s). A likely mechanism for the origin of this UPD is the conception of a fetus with trisomy for chromosome 15 with two chromosomes from the mother and one from the father. Trisomy 15 is usually lethal and will lead to miscarriage. However, the loss of a chromosome 15 in an occasional cell during early embryogenesis will allow that cell line to proliferate and result in a viable fetus. If the sole paternal chromosome is the one that is lost in this trisomy rescue, the resulting infant will have maternal UPD and PWS. Alternatively, UPD could arise from the rescue of a monosomic conceptus, by duplication of the single homologue. Maternal and paternal UPDs for many of the human chromosomes have now been identified. Several of these result in a normal phenotype, presumably because the chromosome does not harbor any imprinted gene(s) [87]. However, these individuals may be at risk for being homozygous for recessive genes. The possible role of UPD, a unique form of chromosomal inheritance, in disease states of unknown etiology is being investigated.

9.8 CHROMOSOME ABNORMALITIES

Human cytogenetics has advanced since the late 1970s because of continuing technical advances and the high incidence of chromosome abnormalities in the human population. It is estimated that the frequency of significant chromosome abnormalities among live births is about 1 in 150. It is well documented that about 50% of first-trimester pregnancy losses are due to chromosome abnormalities, mostly numerical anomalies. Thus, chromosome aberrations have a significant impact as causes of pregnancy wastage, congenital malformations, mental retardation, abnormalities of sex differentiation, and behavior problems. Acquired chromosomal changes play a significant role in carcinogenesis and in tumor progression.

Most chromosomal abnormalities exert their phenotypic effects by increasing or decreasing the quantity of genetic material. Chromosomal abnormalities can be divided into numerical and structural abnormalities. Structural changes such as translocations and inversions pose a much more serious familial recurrence risk for chromosome abnormalities. This is due to aberrant segregation of chromosomes during meiosis in clinically normal carriers of these balanced rearrangements.

9.8.1 Numerical Chromosome Abnormalities

The most straightforward of chromosomal abnormalities are alterations of chromosome number. Deviation from the normal diploid complement of 46 chromosomes is referred to as “aneuploidy”; an extra chromosome results in “trisomy,” whereas a missing chromosome results in “monosomy.” Although all the possible chromosomal trisomies have been observed in spontaneous abortions, trisomies 13, 18, and 21 are the only autosomal trisomies to be observed in a nonmosaic state in live-born infants, and are discussed in detail in later volumes. All autosomal monosomies are lethal. The only viable monosomy involves the X chromosome (45, X resulting in Turner syndrome). Abnormalities associated with sex chromosomes are discussed in detail in later volumes. Aneuploidy results from nondisjunction, in which two copies of a chromosome go to the same daughter cell during meiosis or mitosis. Nondisjunction occurs most often in the first meiotic division in the maternal germ-line. In meiosis I nondisjunction, both homologues of a chromosome move to the same pole during anaphase I instead of moving to opposite poles, giving rise to one daughter cell with two copies of the chromosome and the other with none. The latter product is never recovered because of lethality associated with monosomy. In the case of meiosis II nondisjunction, the two sister chromatids of a homologue move to the same pole, again giving rise to one daughter cell with two copies of the chromosome and the other with none. Mitotic nondisjunction results in the presence of an aneuploid and a normal cell line—a condition referred to as “mosaicism.” The causes of nondisjunction are unknown. The only well-documented risk factor is advanced maternal age.

The term “polyploidy,” on the other hand, refers to the presence of a complete extra set of chromosomes; “triploidy” represents three sets with 69 chromosomes, whereas “tetraploidy” represents four sets with 92 chromosomes. Rarely, a triploid fetus will be liveborn, but

in general polyploidy is lethal. In a few instances, however, mosaicism for a diploid and a triploid cell line producing congenital anomalies has been compatible with long-term survival.

9.8.2 Structural Chromosome Abnormalities

Structural chromosomal rearrangements result from chromosome breakage with subsequent reunion in a different configuration. They can be balanced or unbalanced. In balanced rearrangements, the chromosome complement is complete with no loss or gain of genetic material. Consequently, balanced rearrangements are generally harmless, with the exception of rare cases in which one of the breakpoints disrupts an important functional gene. Carriers of balanced rearrangements are often at risk of having children with an unbalanced chromosome complement. When a chromosome rearrangement is unbalanced, the chromosome complement contains an incorrect amount of genetic material, usually with serious clinical effects.

A deletion involves the loss of part of a chromosome and results in monosomy for that segment of the chromosome, whereas duplication represents the doubling of part of a chromosome, resulting in trisomy for that segment. The result is either decrease or increase in gene dosage. In general, duplications appear to be less harmful than deletions. Very large deletions usually are incompatible with survival to term. Deletions or duplications larger than ~5 Mb in size can be visualized under the microscope using G-banded chromosome analysis. Genomic disorders resulting from submicroscopic deletions and duplications (i.e., microdeletions and microduplications) with a size <5 Mb have been identified with the help of molecular cytogenetic techniques. In these syndromes, groups of contiguous genes are either deleted or duplicated, resulting in a defined set of congenital anomalies.

The use of CMA to analyze the genomes of normal humans led to the discovery of extensive genomic benign CNVs with the majority <500 kb in size [88,89]. Benign CNVs have been proposed to be a major factor responsible for human diversity [90]. Through genomic rearrangement of rearrangement-prone regions as a result of the genomic architecture, pathogenic CNVs, on the other hand, can cause genomic disorders due to the loss or gain of a dosage-sensitive gene(s) resulting in a clinical phenotype [91]. These pathogenic CNVs include recurrent microdeletions and microduplications

with common size and breakpoint clustering, which are flanked by segmental duplications (also called low-copy repeats) and mediated by nonallelic homologous recombination, as well as nonrecurrent deletions and duplications with different sizes and variable breakpoints for each CNV, which vary in size from a few hundred kilobases to a few megabases, and are mediated by other molecular mechanisms [92,93]. Well-known genomic disorders can be phenotypically heterogeneous and variable due to incomplete penetrance or variable expression. Clinical variability could also be explained in part by other genetic or environmental determinants, modifying factors of other genes, multigenic inheritance, imprinting, and unmasking of recessive genes. For many years, genomic disorders due to microdeletions and microduplications that are clinically recognizable by their typical constellation of clinical features were tested for by FISH analysis. The advances in CMA technologies since 2008 have allowed their widespread use as a clinical diagnostic modality in a wide variety of human genetic diseases. High-resolution CMA allows the detection of pathogenic CNVs in nearly 17%–19% of patients with developmental delay and/or intellectual disability who had a normal G-banded karyotype [94]. Because of its high diagnostic yield, CMA was recommended in 2010 by the American College of Medical Genetics and Genomics as the preferred first-tier clinical diagnostic test for individuals with developmental delay, intellectual disability, and/or multiple congenital anomalies [94,95].

Translocations involve the exchange of genetic material between chromosomes. In a balanced reciprocal translocation the exchange is equal, with no loss or gain of genetic material, though it is possible for a gene to be disrupted at one of the breakpoints. More often, the carrier of a balanced translocation is free of clinical signs or symptoms but is at risk for having offspring with unbalanced chromosomes. The phenotype usually is a complex mixture as a result of the loss or gain of at least two chromosome segments and therefore can be difficult to predict. One specific type of translocation that is relatively common is the “Robertsonian translocation.” This results from a fusion of two acrocentric chromosomes at the centromere. Carriers of a Robertsonian translocation have 45 chromosomes and are clinically unaffected. The most common clinically significant outcome is trisomy 21, in which a carrier for a Robertsonian translocation involving chromosome 21 produces a gamete with both the translocation chromosome and a normal 21, resulting in trisomy 21 after fertilization.

Inversions occur when there are two breaks in a chromosome and the intervening material flips 180 degrees. Inversions that span the centromere are referred to as “pericentric,” whereas those that do not are called “paracentric.” Inversions generally do not result in added or lost genetic material, and therefore usually are viewed as neutral changes. Disruption of a gene at one of the breakpoints, however, could change the function of that gene. Also, alteration of the gene order at the borders of the inversion could affect the function of blocks of genes that are coordinately regulated (“position effect”). If a crossover occurs in the inverted segment of a pericentric inversion during meiosis, two recombinant chromosomes result, one with duplication of one end and deletion of the other end, and the other having the opposite arrangement. Such a crossover event in a paracentric inversion results in dicentric or acentric chromosomes that tend to be unstable.

An insertion occurs when a segment of one chromosome becomes inserted into another chromosome. Because these changes require three chromosomal breakpoints, they are relatively rare. Abnormal segregation in a balanced insertion carrier can produce offspring with either duplication or deletion of the inserted segment, as well as balanced carriers and normal offspring.

A “marker” chromosome is a rearranged chromosome whose genetic origin is unknown based on its G-banded chromosome morphology. Usually they are present in addition to the normal chromosome complement and are thus called supernumerary marker chromosomes. Two-thirds of de novo marker chromosomes can be associated with an abnormal outcome, whereas inherited ones can be passed from generation to generation without apparent clinical effects. Larger markers with more genetically active material are more likely to be of clinical significance. FISH and CMA have proved very helpful in the precise identification of the genetic origin of supernumerary marker chromosomes. Ring chromosomes are formed when a chromosome undergoes two breaks and the broken ends reunite in a ring structure. Rings encounter difficulties in mitosis and are unstable, resulting in some cells that lose the ring and are therefore monosomic for the chromosome, and others that have multiple copies of the ring. An “isochromosome” is a chromosome in which one arm is missing and the other duplicated in a mirror-image fashion. The most commonly encountered

isochromosome is that which consists of two long arms of the X chromosome. This accounts for ~15% of all cases of Turner syndrome [98].

9.9 CONCLUDING REMARKS

In this chapter, we have outlined the structural, functional, and behavioral aspects of human chromosomes and their relationship to disease states. Although investigations have provided insights into several aspects of chromosome structure, the details of the higher order structure of chromosomes are not well understood at the molecular level in full detail.

We have begun to understand the DNA sequences that are associated with chromosomal landmarks, such as centromeres, telomeres, chromosome bands, and fragile site; however, we still do not understand the role of such sequences in producing the associated functional correlates. For example, which sequences are critical for centromere function, and how do dicentric chromosomes decide which will function? Most enigmatic, what is the origin of chromosome bands and what is the molecular organization of a band border?

The availability of the finished human genome sequence and CMA has allowed the detection of genomic CNVs on a global scale. It is now appreciated that the underlying genomic architecture plays a crucial role in the origin of these genomic rearrangements in rearrangement-prone regions. Segmental duplications have arisen in the primate genome, driving the process of chromosome evolution. In addition to creating a dynamic, evolvable genome, these segmental duplications result in instability, genomic rearrangement, and disease. We have begun to understand the organization of segmental duplications, which predisposes the chromosomes that carry them to germline genomic rearrangements such as deletions, duplications, and inversions; however, we do not understand the mechanisms that initially led to the formation of segmental duplications, nor the sequences or structures responsible for their continued instability.

Many new insights have come from understanding the structure and function of the human X chromosome and genomic imprinting; however, we do not know to what extent the remainder of the genome may contain imprinted or partially imprinted genes whose parental origin in part determines tissue-specific expression. Might such “epigenetic” phenomena provide another mechanism for both normal human variation and disease susceptibility?

Much remains to be learned about the molecular aspects of chromosome structure, function, and behavior. It is anticipated that the human genome sequence and its functional characterization will provide the tools with which to approach these problems and define a new frontier for the role of chromosomes in human disease.

REFERENCES

- [1] Ford CE, Hamerton JL. The chromosomes of man. *Nature* 1956;178:1020–3.
- [2] Tjio JH, Levan A. The chromosome number of man. *Hereditas* 1956;42:1–6.
- [3] Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 1999;98:285–94.
- [4] Finch JT, Klug A. Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci USA* 1976;73:1897–901.
- [5] Woodcock CL, Frado LL, Rattner JB. The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J Cell Biol* 1984;99:42–52.
- [6] Bassett A, Cooper S, Wu C, Travers A. The folding and unfolding of eukaryotic chromatin. *Curr Opin Genet Dev* 2009;19:159–65.
- [7] Dorigo B, Schalch T, Kulangara A, Duda S, Schroeder RR, Richmond TJ. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* 2004;306:1571–3.
- [8] Schalch T, Duda S, Sargent DF, Richmond TJ. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* 2005;436:138–41.
- [9] Grigoryev SA, Arya G, Correll S, Woodcock CL, Schlick T. Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc Natl Acad Sci USA* 2009;106:13317–22.
- [10] Hart CM, Laemmli UK. Facilitation of chromatin dynamics by SARs. *Curr Opin Genet Dev* 1998;8: 519–25.
- [11] Hirano T. At the heart of the chromosome: SMC proteins in action. *Nat Rev Mol Cell Biol* 2006;7:311–22.
- [12] Maeshima K, Hihara S, Eltsov M. Chromatin structure: does the 30-nm fibre exist in vivo? *Curr Opin Cell Biol* 2010;22:291–7.
- [13] Maeshima K, Eltsov M. Packaging the genome: the structure of mitotic chromosomes. *J Biochem* 2008;143:145–53.
- [14] Nasmyth K. A prize for proliferation. *Cell* 2001;107:689–701.
- [15] Nurse P. The incredible life and times of biological cells. *Science* 2000;289:1711–6.
- [16] Nasmyth K. Segregating sister genomes: the molecular biology of chromosome separation. *Science* 2002;297:559–65.
- [17] Crow JF. The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci USA* 1997;94:8380–6.
- [18] Hassold T, Hunt P. Maternal age and chromosomally abnormal pregnancies: what we know and what we wish we knew. *Curr Opin Pediatr* 2009;21:703–8.
- [19] Hassold T, Sherman S. Down syndrome: genetic recombination and the origin of the extra chromosome 21. *Clin Genet* 2000;57:95–100.
- [20] Lamb NE, Hassold TJ. Nondisjunction—a view from ringside. *N Engl J Med* 2004;351:1931–4.
- [21] Moorhead PS, Nowell PC, Mellman WJ, et al. Chromosome preparations of leukocytes cultured from peripheral blood. *Exp Cell Res* 1960;20:613–6.
- [22] Caspersson T, Zech L, Johansson C, Modest EJ. Identification of human chromosomes by DNA-binding fluorescent reagents. *Chromosoma* 1970;30:215–27.
- [23] Seabright M. Rapid banding technique for human chromosomes. *Lancet* 1971;2:971–2.
- [24] Yunis JJ. High resolution of human chromosomes. *Science* 1976;191:1268–70.
- [25] Dutrillaux B, Laurent C, Couturier J, Lejeune J. Coloration par l'acridine orange de chromosomes préalablement traités par le 5-bromodeoxyuridine (BUDR). *CR Acad Sci D Sci Naturelles (Paris)* 1973;276(24):3179–81.
- [26] Shaffer LG, Slovak ML, Campbell LJ, editors. An international system for human cytogenetic nomenclature (ISCN). Basel: S. Karger; 2009.
- [27] Arrighi FE, Hsu TC. Localization of heterochromatin in human chromosomes. *Cytogenetics* 1971;10:81–6.
- [28] Goodpasture C, Bloom SE, Hsu TC, Arrighi FE. Human nucleolus organizers: the satellites or the stalks? *Am J Hum Genet* 1976;28:559–66.
- [29] Latt SA. Microfluorometric detection of deoxyribonucleic acid replication in human metaphase chromosomes. *Proc Natl Acad Sci USA* 1973;70:3395–9.
- [30] Korenberg JR, Friedlander EF. Giemsa technique for the detection of sister chromatid exchanges. *Chromosoma* 1974;48:355–60.
- [31] van Brabant AJ, Stan R, Ellis NA. DNA helicases, genomic instability and human genetic disease. *Ann Rev Genomics Hum Genet* 2000;1:409–59.
- [32] Korenberg JR, Engels WR. Base ratio, DNA content, and quinacrine-brightness of human chromosomes. *Proc Natl Acad Sci USA* 1978;75:3382–6.
- [33] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.

- [34] Bailey JA, Eichler EE. Genome-wide detection and analysis of recent segmental duplications within the mammalian organisms. *Cold Spring Harb Symp Quant Biol* 2003;68:115–24.
- [35] Sutherland GR, Baker E, Richards RI. Fragile sites still breaking. *Trends Genet* 1998;14:501–6.
- [36] Sutherland GR. Rare fragile sites. *Cytogenet Genome Res* 2003;100:77–84.
- [37] Arlt MF, Casper AM, Glover TW. Common fragile sites. *Cytogenet Genome Res* 2003;100:92–100.
- [38] Pinkel D, Straume T, Gray JW. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc Natl Acad Sci USA* 1986;83:2934–8.
- [39] Liehr T, Starke H, Weise A, Lehrer H, Claussen U. Multicolor FISH probe sets and their applications. *Histol Histopathol* 2004;19:229–37.
- [40] Levy B, Dunn TM, Kaffe S, Kardon N, Hirschhorn K. Clinical applications of comparative genomic hybridization. *Genet Med* 1998;1:4–12.
- [41] Shaffer LG, Bejjani BA. Medical applications of array CGH and the transformation of clinical cytogenetics. *Cytogenet Genome Res* 2006;115:303–9.
- [42] Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* 2003;13:2291–305.
- [43] Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* 2006;34:445–50.
- [44] Edelmann L, Hirschhorn K. Clinical utility of array CGH for the detection of chromosomal imbalances associated with mental retardation and multiple congenital anomalies. *Ann NY Acad Sci* 2009;1151:157–66.
- [45] Furey TS, Haussler D. Integration of the cytogenetic map with the draft human genome sequence. *Hum Mol Genet* 2003;12:1037–44.
- [46] Holmquist GP. Chromosome bands, their chromatin flavors and their functional features. *Am J Hum Genet* 1992;51:17–37.
- [47] Willard HF. Centromeres: the missing link in the development of human artificial chromosomes. *Curr Opin Genet Dev* 1998;8:219–25.
- [48] Grimes B, Cooke H. Engineering mammalian chromosomes. *Hum Mol Genet* 1998;7:1635–40.
- [49] Liehr T, Claussen U, Starke H. Small supernumerary marker chromosomes (sSMC) in humans. *Cytogenet Genome Res* 2004;107:55–67.
- [50] Craig JM, Earnshaw WC, Vagnarelli P. Mammalian centromeres: DNA sequence, protein composition, and role in cell cycle progression. *Exp Cell Res* 1999;246:249–62.
- [51] Fukagawa T. Assembly of kinetochores in vertebrate cells. *Exp Cell Res* 2004;296:21–7.
- [52] Chan SRWL, Blackburn EH. Telomeres and telomerase. *Philos Trans R Soc Lond B Biol Sci* 2004;359:109–21.
- [53] Osterhage JL, Friedman KL. Chromosome end maintenance by telomerase. *J Biol Chem* 2009;284:16061–5.
- [54] Blasco MA, Gasses SM, Lingner J. Telomeres and telomerase. *Genes Dev* 1999;13:2353–9.
- [55] Flint J, Craddock CF, Villegas A, et al. Healing of broken human chromosomes by the addition of telomeric repeats. *Am J Hum Genet* 1994;55:505–12.
- [56] Varley H, Di S, Scherer SW, Royle NJ. Characterization of terminal deletions at 7q32 and 22q13.3 healed by de novo telomere addition. *Am J Hum Genet* 2000;67:610–22.
- [57] Varley H, Pickett HA, Foxon JL, Reddel RR, Royle NJ. Molecular characterization of inter-telomere and intra-telomere mutations in human ALT cells. *Nat Genet* 2002;30:301–5.
- [58] Neumann AA, Reddel RR. Telomere maintenance and cancer—look, no telomerase. *Nat. Rev. Cancer* 2002;2:879–84.
- [59] Fortin F, Beaulieu Bergeron M, Fetni R, Lemieux N. Frequency of chromosome healing and interstitial telomeres in 40 cases of constitutional abnormalities. *Cytogenet Genome Res* 2009;125:176–85.
- [60] Bosco G, Haber JE. Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. *Genetics* 1998;150:1037–47.
- [61] Ambrosini A, Paul S, Hu S, Riethman H. Human subtelomeric duplcon structure and organization. *Genome Biol* 2007;8:R151.
- [62] Ford CE, Jones KW, Polani PE, et al. A sex chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* 1959;1:711–3.
- [63] Jacobs PA, Strong JA. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* 1959;183:302–3.
- [64] Rappold GA. The pseudoautosomal regions of the human sex chromosomes. *Hum Genet* 1993;92:315–24.
- [65] Goodfellow PN, Ferguson-Smith MA, Hawkins JR, Camerino G. SRY and primary sex-reversal syndromes. In: Valle D, Beaudet A, Vogelstein B, Kinzler K, Antonarakis S, Ballabio S, editors. *The online metabolic and molecular bases of inherited diseases*. McGraw-Hill; 2007. [Chapter 62].
- [66] Harley VR, Jackson DI, Hextall PJ, Hawkins JR, Berkovitz GD, Sockanathan S, Lovell-Badge R, Goodfellow PN. DNA binding activity of recombinant SRY from normal males and XY females. *Science* 1992;255:453–6.
- [67] Nasrin N, Buggs C, Kong XF, Carnazza J, Goebel M, Alexander-Bridges M. DNA-binding properties of the product of the testis-determining gene and a related protein. *Nature* 1991;354:317–20.

- [68] Pontiggia A, Rimini R, Harley VR, Goodfellow PN, Lovell-Badge R, Bianchi ME. Sex-reversing mutations affect the architecture of SRY-DNA complexes. *EMBO J* 1994;13:6115–24.
- [69] McLaughlin DT, Donahoe PK. Sex determination and differentiation. *N Engl J Med* 2004;350:367–78.
- [70] Roberts LM, Shen J, Ingraham HA. New solutions to an ancient riddle: defining the differences between Adam and Eve. *Am J Hum Genet* 1999;65:933–42.
- [71] Swain A, Lovell-Badge R. Mammalian sex determination: a molecular drama. *Genes Dev* 1999;13:755–67.
- [72] Lyon MF. Epigenetic inheritance in mammals. *Trends Genet* 1993;9:123–8.
- [73] Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X-inactivation by DNA methylation. *Science* 1981;211:393–6.
- [74] Riggs AD, Pfeifer GP. X-chromosome inactivation and cell memory. *Trends Genet* 1992;8:169–74.
- [75] Jeppesen P, Turner BM. The inactive X chromosome in female mammals is distinguished by lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* 1994;74:281–9.
- [76] Keohane AM, Lavender JS, O'Neill LP, Turner BM. Histone acetylation and X inactivation. *Dev Genet* 1998;22:65–73.
- [77] Gilbert SL, Sharp PA. Promoter-specific hypoacetylation of X-inactivated genes. *Proc Natl Acad Sci USA* 1999;96:13825–30.
- [78] Elgin SCR, Grewal SIS. Heterochromatin: silence is golden. *Curr Biol* 2003;13:R895–8.
- [79] Brown CJ, Ballabio A, Rupert JL, et al. A gene from the region of the human X-chromosome inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* 1991;349:38–44.
- [80] Valley CM, Willard HF. Genomic and epigenomic approaches to the study of X chromosome inactivation. *Curr Opin Genet Dev* 2006;16:240–5.
- [81] Lee JT, Strauss WM, Dausman JA, Jaenisch R. A 450kb transgene displays properties of the mammalian X-inactivation center. *Cell* 1996;86:83–94.
- [82] Lee JT, Lu N, Han Y. Genetic analysis of the mouse X inactivation center defines an 80-kb multifunction domain. *Proc Natl Acad Sci USA* 1999;96:3836–41.
- [83] Lee JT, Jaenisch R. Long-range cis effects of ectopic X-inactivation centres on a mouse autosome. *Nature* 1997;386:275–9.
- [84] Willard HF. The sex chromosomes and X chromosome inactivation. In: Valle D, Beaudet A, Vogelstein B, Kinzler K, Antonarakis S, Ballabio S, editors. *The online metabolic and molecular bases of inherited diseases*. McGraw-Hill; 2007. [Chapter 61].
- [85] Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 2005;434:400–4.
- [86] Bartolomei MS, Tilghman SM. Genomic imprinting in mammals. *Annu Rev Genet* 1997;31:493–525.
- [87] Ledbetter DH, Engel E. Uniparental disomy in humans: development of an imprinting map and its implications for prenatal diagnosis. *Hum Mol Genet* 1995;4:1757–64.
- [88] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–51.
- [89] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. Large-scale copy number polymorphism in the human genome. *Science* 2004;305:525–8.
- [90] Lupski JR. Genome structural variation and sporadic disease traits. *Nat Genet* 2006;38:974–6.
- [91] Stankiewicz P, Beaudet AL. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr Opin Genet Dev* 2007;17:182–92.
- [92] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010;61:437–55.
- [93] Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics* 2008;1(4).
- [94] Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010;86:749–64.
- [95] Manning M, Hudgins L. Professional Practice and Guidelines Committee. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genet Med* 2010;12:742–5.
- [96] Baptista J, Mercer C, Prigmore E, Gribble SM, Carter NP, Maloney V, Thomas NS, Jacobs PA, Crolla JA. Breakpoint mapping and array CGH in translocations: comparison of a phenotypically normal and an abnormal cohort. *Am J Hum Genet* 2008;82:927–36.
- [97] Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL, et al. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *Am J Hum Genet* 2008;82:712–22.
- [98] Gardener RJM, Sutherland GR, editors. *Chromosome abnormalities and genetic counseling*. Oxford and New York: Oxford University Press; 2004.

Mitochondrial Biology and Medicine

*Douglas C. Wallace^{1,2}, Marie T. Lott¹,
Vincent Procaccio³*

¹Center for Mitochondrial and Epigenomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, United States

²Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

³Biochemistry and Genetics Department, MitoVasc Institute, UMR CNRS 6015 – INSERM U1083, CHU Angers, Angers, France

10.1 INTRODUCTION

Western medicine is organized around anatomy, yet life is the interplay between structure (anatomy), energy (vital force), and information. Consequently, the anatomical paradigm of Western medicine has largely overlooked the central role of bioenergetics in health and disease. This may be a critical factor in our inability to understand and develop effective therapies for the common metabolic and degenerative diseases and aging. While bioenergetics has been neglected by Western medicine, it is central to Eastern medicine with the concept of Qi, which can be loosely translated as “vital force.”

The dichotomy between anatomy and energy for our cells harkens back to the origin of the eukaryotic cell about 2.5 billion years ago. In a unique single event, two co-equal micro-organisms formed a symbiosis that set the stage for all multicellular plants and animals. The original microorganisms were an archaeobacterium that gave rise to the eukaryotic cell nucleus and cytosol and an oxidative eubacterium that gave rise to the cytoplasmic mitochondria. While the original archaeobacterium and eubacterium had similar-size genomes, most of the eubacterial mitochondrial genes were transferred to the archaeobacterial nuclear genome in association

with the proliferation of the mitochondria within the cytoplasm. Because most of a bacterium's energy is used in replicating its DNA, transcribing its RNA, and translating its proteins, by transferring the mitochondrial genes into the nucleus, the number of gene copies for each gene could be reduced from hundreds to thousands down to two, with hundreds of -fold savings of energy. The excess mitochondrial energy could then be used to sustain a much larger nuclear genome with extra genes allocated for multicellularity and organogenesis.

By this process, the residual mitochondrial DNA (mtDNA) of multicellular animals and humans has been reduced to about 13 polypeptide genes plus the rRNA and tRNA genes for their translation on mitochondria-specific ribosomes. The mitochondrial ribosomes retain several features of bacterial translation including chloramphenicol and aminoglycoside antibiotic sensitivity and polypeptide initiation with a formyl-methionine. While few in number, the mtDNA polypeptide genes are essential components of the mitochondrial energy-generating process, oxidative phosphorylation (OXPHOS). In essence, the mtDNA codes for the wiring diagram of the cellular power plant while the nuclear DNA (nDNA) contains the blueprints for building the power plants.

Interest in mitochondrial medicine has been increasing rapidly during the past decade, with the current annual number of mitochondria-related biomedical papers exceeding that of genomics papers [1]. Furthermore, the massive efforts to sequence nDNAs to identify the common genetic variants that cause the common metabolic and degenerative disease have been disappointing. This suggests that the classic Mendelian paradigm of genetics is inadequate for a coherent understanding of human genetics. Many of the seemingly puzzling features of the genetics of common diseases such as variable penetrance, delayed onset and progressive course, and multiorgan involvement are naturally explained by mitochondrial genetics. Thus, by combining Mendelian genetics with mitochondrial genetics we can arrive at a synthetic paradigm that can explain many of the novel features of clinical genetics.

The first report that mitochondrial dysfunction could be associated with a clinical phenotype came with the report of a woman with hypermetabolism, abnormal muscle mitochondria (mitochondrial myopathy), and uncoupled mitochondrial OXPHOS [2]. Subsequent studies revealed a variety of clinical phenotypes associated with mitochondrial myopathy and OXPHOS dysfunction resulting in a proliferation of clinical descriptors such as chronic progressive external ophthalmoplegia (CPEO), mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS), and myoclonic epilepsy and ragged red fiber (MERRF) disease [3,4]. However, family studies showed considerable variability in phenotypes leading to the debate as to whether “mitochondrial diseases” should be split into subphenotypes or lumped into larger categories (“splitters” versus “lumpers”). This ambiguity was resolved by the demonstration that mtDNA mutations could cause disease [5–8] and the realization that cells can have thousands of copies of the mtDNA. The high mtDNA ploidy could then accommodate different percentages of mutant resulting in variable biochemical defects and diverse phenotypes [9]. It is now clear that pathogenic mtDNA mutations are not rare, but common. The prevalence of pathogenic mtDNA mutation cases has been estimated at 1:4300 [10–13], but greater than 1:200 newborn cord bloods harbor one of the 10 most common pathogenic mtDNA mutations [14].

Current estimates of the prevalence of mitochondrial disease are based on “primary mitochondrial diseases,” those that are caused by nDNA or mtDNA

mitochondrial gene mutations that are severe enough to cause a clinically relevant disease by themselves. The phenotypes and mtDNA and nDNA mutations associated with primary mitochondrial diseases were reported in our previous chapter in this series [15], with a current listing of the nDNA and mtDNA genes implicated provided in MITOMAP [16].

However, even in the early days of clinical mitochondrial genetics, it was clear that nDNA gene mutations could alter mtDNA structures, which alter mitochondrial functions and which generate mitochondrial disease. It is now clear that mitochondrial bioenergetic dysfunction, often resulting from faulty nDNA–mtDNA interactions, is prevalent and associated with the etiology of a wide range of metabolic and degenerative diseases and aging. Thus, the characterization of mitochondrial medical genetics during the past 30 years is restructuring the way we understand human genetics and reorienting the way we investigate the etiology of common diseases.

10.2 MITOCHONDRIAL BIOCHEMISTRY

Each cell harbors hundreds to thousands of double-membrane mitochondria. The outer mitochondrial membrane is perhaps the remnant of the phagocytic vesicle and the inner membrane that of the eubacterial plasma membrane. The matrix is the bacterial cytoplasm, and the intermembrane space separates the two organisms.

The highly invaginated inner membrane harbors the enzymes of OXPHOS on in-foldings called cristae. The cristae are closed at the junction with the intermembrane space to create “cristae lumens” [17]. The mitochondria inner membrane has a unique composition that includes cardiolipin, and the mitochondrion has an independent bacteria-like lipid biosynthesis system that produces the essential lipoic acid [18]. The mitochondria are the metabolic hub of the cell. Central metabolic pathways include the tricarboxylic cycle (TCA, Krebs cycle), fatty acid β -oxidation, amino acid metabolism, and cholesterol and heme synthesis. Pyruvate from glycolysis enters the mitochondrion via the pyruvate carrier [19] and is then cleaved by pyruvate dehydrogenase (PDH) to generate acetyl-CoA and reduced nicotinamide dinucleotide (NADH). Acetyl-CoA is also generated by β -oxidation, and the mitochondrial acetyl-CoA is condensed with oxaloacetate to generate citrate. Citrate can

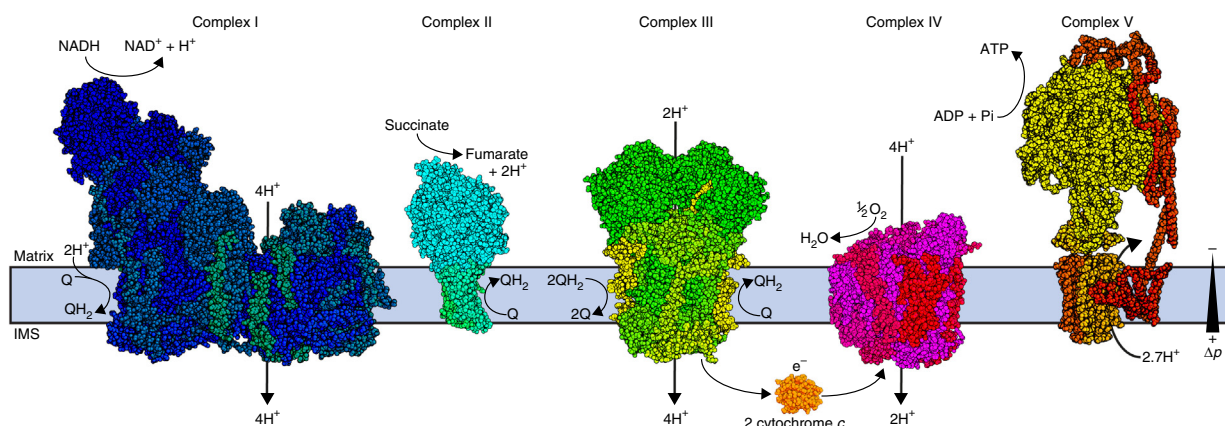


Figure 10.1 Overview of the OXPHOS complexes I–V showing the molecular structures of the five OXPHOS complexes [36].

be exported into the cytosol where it is cleaved by ATP-citrate-lyase to acetyl-CoA and oxaloacetate. Cytosolic acetyl-CoA can be used for fatty acid synthesis or as substrates for protein and histone acetylation [20]. Within the mitochondrion, citrate is metabolized to isocitrate, aconitate, α -ketoglutarate, succinyl-CoA, succinate, malate, and oxaloacetate via the TCA cycle. Isocitrate dehydrogenase (IDH) has three isoforms—two in the mitochondrion, an NADP⁺-linked IDH2 and an NAD⁺-linked IDH3, and one in the cytosol, an NADP⁺-linked form. IDH1, IDH1, and IDH2 are important in cancer genetics [21,22]. PDH, α -ketoglutarate dehydrogenase, and branched chain-keto acid dehydrogenase (BCKDH) are multiple polypeptide complexes that share common subunits and cofactors, including lipoic acid, and are modulated by Ca²⁺ [23]. The integrity of the mitochondrial inner membrane requires special transport systems to move organic molecules in and out of the mitochondrion. This is accomplished in part by the 53 members of the solute carrier family 25 (SLC25), also called the mitochondrial carrier family, which transport carboxylates, amino acids, nucleotides, and cofactors across the inner mitochondrial membrane [24–26]. Mitochondria have also been proposed to contain an adenyllyl cyclase [27] and a nitric oxide (NO) synthase [28].

Each mitochondrion is a capacitor composed of a proton gradient across the mitochondria inner membrane generated by the electron transport chain (ETC). The ETC burns dietary calories (reducing equivalents) with oxygen in a stepwise process that transmits electrons from reduced to oxidized starting with NADH and succinate. NADH is oxidized by complex I (NADH:CoQ

oxidoreductase or NADH dehydrogenase) and succinate is oxidized by complex II (succinate:CoQ oxidoreductase, or succinate dehydrogenase). Complexes I and II then transfer the electrons to the lipid carrier, coenzyme Q₁₀ (CoQ), which transfers them to complex III. Complex III transfers the electrons to cytochrome c and cytochrome c transfers the electrons to complex IV (cytochrome c oxidase or COX). Complex IV combines four electrons with O₂ to generate 2H₂O. As the electrons traverse complex I, III, and IV, the energy released is used to pump protons across the mitochondrial inner membrane, four through complex I and one each through complexes III and IV (Fig. 10.1A).

Each of the respiratory complexes is composed of multiple polypeptides. The molecular structures of complex I [29,30], complex II [31], complex III [32], and complex IV [33,34] have been established. Complex I is composed of 45 polypeptides, complex II of four, complex III of 11, and complex IV of 13 (Fig. 10.1). Each complex also contains various prosthetic groups that conduct electrons. Complex I harbors an FMN site to collect electrons from NADH, eight iron-sulfur (FeS) centers, and a CoQ binding site. Complex II harbors an FAD, an FeS, a cytochrome b, and a CoQ binding site. Complex III has the heme-based cytochromes b and c₁, the Rieske FeS center, and two CoQ binding sites and the cytochrome c contains heme c. Complex IV encompasses cytochromes a + a₃, two Cu centers, and the oxygen reaction center [35]. The respiratory chain is also assembled into super-complexes, the physiological function of which remains to be determined [36–39].

The proton gradient ($\Delta P = \Delta \Psi + \Delta \mu^{H^+}$) is used to energize multiple mitochondrial functions, the best known is the generation of ATP from ADP + Pi by complex V (proton-translocating ATP synthase or ATP synthase). Complex V is composed of 18 polypeptides. The ETC enzymes and the ATP synthase are arrayed within the cristae lumen membranes with the ETC charging the proton gradient within the cristae lumens and the ATP synthase utilizing the proton gradient to drive its spinning ring of “c” subunits within the inner membrane. Each c subunit has a negatively charged carboxyl group that picks up a proton from the cristae lumen via a half channel in the abutting membrane bound and static ATP6 protein. The c ring then rotates 360 degrees within the plane of the inner membrane until it comes back to the ATP6 subunit. The ATP6 protein has a second half proton channel open to the matrix through which the proton is released. The c-ring wheel has a γ subunit axial that protrudes inside the F1 ($3\alpha:3\beta$) barrel. The barrel is fixed to the ATP6 inner membrane protein and is also static [40]. The wheel spins at 300 Hz [41], and the spinning axial contacts the three β subunits sequentially, causing conformational changes to condense ADP + Pi to make ATP. The ATP synthase is an offset dimer [42], which is aggregated around the edges of the cristae lumens.

ATP generated in the matrix is exported to the intermembrane space by the adenine nucleotide translocators (ANTs) which belonging to the mitochondrial carrier protein family. There are four ANT isoforms in humans, ANT1 being expressed in the heart and muscle, with ANT1 mutations having been identified in human CPEO and mitochondrial myopathy and cardiomyopathy [43–45]. The intermembrane space communicates with the cytosol via the voltage-dependent anion channels (VDAC, porin).

Besides ATP synthesis, the proton gradient is used for multiple other processes. The best characterized of these is the uptake of cytosolic Ca^{2+} through the mitochondrial Ca^{2+} uniporter (MCU) complex [46,47]. The cristae lumens are closed at the intermembrane space by the MICOS + Opa1 complex [48,49]. Opa1 can be cleaved by OMA1 resulting in the opening of the cristae and the release of cytochrome c and protons into the cytoplasm initiating apoptosis. This is associated with the activation of the inner membrane mitochondrial permeability transition pore (mtPTP) located within the inner membrane and resulting in depolarization of

the proton gradient. This combination of events initiates the intrinsic pathway of apoptosis. The structure of the mtPTP, which interacts with the Ca^{2+} -activated cyclophilin D (cypD), is actively debated with current contenders being the ATP synthase dimers, ATP synthase c-ring [41,50–54], or the SPG7 hexamer [55].

The mitochondria are highly dynamic organelles undergoing fission via activated *Drp1* and fusion mediated by *Opa1*, *Mfn1*, and *Mfn2* [56]. The mitochondria replicate within the cytosol and the excess mitochondria are removed by mitophagy [57]. Hence, the mitochondria are a colony of bacteria maintained in a metastable state via a balance between proliferation and degradation.

10.3 MITOCHONDRIAL GENETICS

The cellular mitochondrial genome is composed of hundreds to thousands of copies of the mtDNA plus between 1 and 2000 nDNA coded mitochondrial genes. While the Mendelian rules for nDNA gene inheritance were well established by the mid-twentieth century, the principles of human mtDNA genetics were elucidated much more recently.

10.3.1 Genetics of mtDNA Genes

The first evidence that the human mtDNAs could code for inheritable traits was obtained by showing that chloramphenicol resistance could be transferred from cell to cell via fusion of cytoplasmic fragments in the absence of the nucleus (transmitochondrial cybrids) [58]. Subsequently, somatic cells were shown to harbor many copies of the mtDNA [59], which can encompass mixtures of different proportions of mutant and normal mtDNAs (heteroplasmy). During cellular replication and cytokinesis, the proportion of mutant and normal mtDNAs can segregate (replicative segregation) and the proportion of mutant mtDNAs determines the degree of the cellular and tissue bioenergetics defect. Each cell, tissue, organ, and individual has a minimal mitochondrial bioenergetic capacity required for normal function. As the proportion of mutant mtDNAs increases, cellular energetics declines until it crosses this minimum bioenergetic threshold and results in phenotypic manifestations (threshold effect) [60–63]. Finally, the human mtDNA is exclusively maternally inherited [64,65].

The mtDNA codes for 13 of the most important polypeptides of OXPHOS: the *ND1*, *ND2*, *ND3*, *ND4*, *ND4L*,

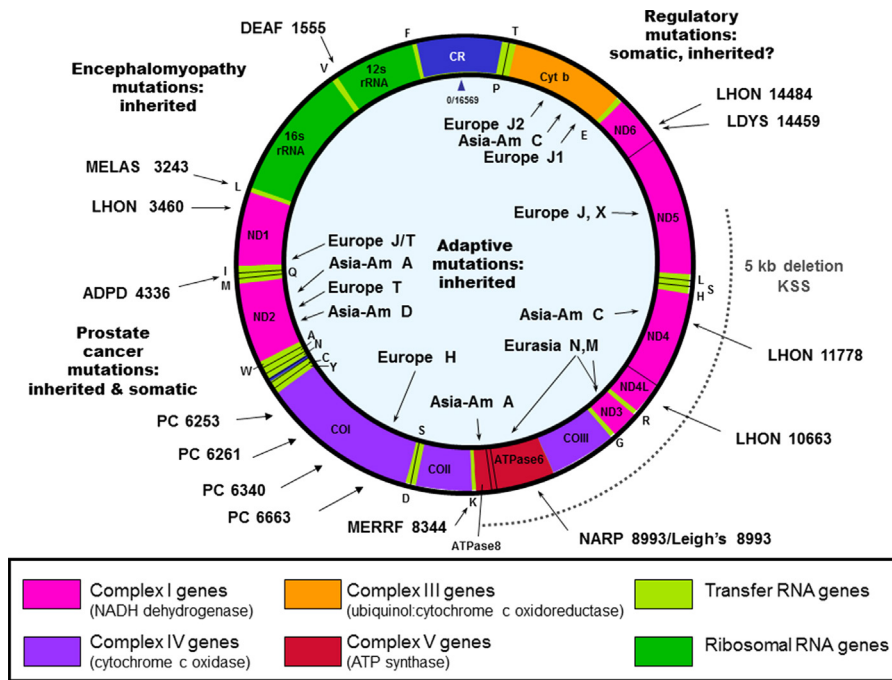


Figure 10.2 The Human mtDNA Map. The gene locations are shown between the concentric lines. Examples of common mtDNA pathogenic mutations are presented on the outside of the circle while the positions of haplogroup specific markers are located within the circle. Letters stand for the amino acid tRNAs. CR, control region. Clinical phenotypes are *Deaf*, deafness; *LDYS*, Leber and dystonia; *LHON*, Leber hereditary optic neuropathy; *MELAS*, mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; *MERRF*, myoclonic epilepsy and ragged red fibers; *NARP*, neurogenic muscle weakness, ataxia, and retinitis pigmentosa; *PC*, prostate cancer. (Figure reproduced from MITOMAP. A Human Mitochondrial Genome Database, 2018. <http://www.mitomap.org>.)

ND5, and *ND6* genes of complexes I; the *cytochrome b* gene of complex III; the *COI*, *COII*, and *COIII* genes of complex IV; and the *ATP6* and *ATP8* genes of complex V. In addition, the mtDNA codes for the 22 tRNAs and two rRNAs for mitochondrial protein synthesis and contains an approximately 1000-nucleotide “control region” that regulates mtDNA transcription and replication [15] (Fig. 10.2).

In addition to the major OXPHOS genes, recent studies have revealed additional polypeptide open reading frames embedded within the mtDNA rRNA genes. Two of these, Humanin and MOTS-c, are thought to generate mRNAs that are exported from the mitochondrion where they are translated on cytosolic ribosomes to generate diffusible peptide hormones [66,67]. The mtDNA may code for additional functional elements such as regulatory RNAs that have yet to be delineated.

Because the mtDNA codes for key OXPHOS polypeptides, genetic alterations in the mtDNA will affect

energy metabolism. However, the central role of the mitochondrial membrane potential and mitochondrial intermediary metabolism means that the physiological effects of mtDNA variation can impinge on virtually every cellular and tissue function.

Because it is exclusively maternally inherited, there is virtually no physical interaction between maternal and paternal mtDNAs. Hence, there is little if any recombination and the mtDNA sequence can change only by the sequential accumulation of mutations along radiating maternal lineages. Thus, mtDNA sequence variants remain in total linkage disequilibrium so the functional effects of individual mtDNA nucleotide variants cannot be analyzed in isolation but must be considered within the context of all of the other variants within that mtDNA haplotype.

The mtDNA has a very high mutation rate [68–70]. For a mutant mtDNA to have a phenotypic effect, it must become enriched within the cell from a single mutant

mtDNA to a significant proportion of the cellular mtDNAs so that its biochemical effect overshadows that of the residual nonmutant mtDNAs. The phenotypic effect of an mtDNA mutation will depend on the nature and severity of the gene alteration, the percentage heteroplasmy, nDNA-modifying factors, the tissue physiology, and environmental influences. Hence, for mtDNA mutations, the one gene–one polypeptide–one phenotype model of Beadle and Tatum does not apply [9].

In addition to germline mutations, the mtDNA can accumulate mutations during development and in somatic tissues with age [71]. The developmental mutants can become enriched in specific tissues by replicative segregation and result in seemingly spontaneous disease [72]. The accumulation of somatic mtDNA mutations in postmitotic cells can slowly erode mitochondrial energetics resulting in tissue decline, potentially explaining the delayed onset and progressive course of adult common diseases and the molecular basis of aging.

The factors that result in the enrichment of a mtDNA mutation in a somatic cell are still unknown. In CPEO, caused by single mtDNA deletions, the deleted mtDNA becomes clonally and regionally enriched along the skeletal muscle fibers [73]. Also, individual mtDNA mutations accumulate in individual neurons in neurodegenerative diseases [74,75] and in cardiomyocytes during aging [76]. One possible mechanism for the preferential accumulation of the mutant mtDNAs is that an individual mtDNA mutation that affects OXPHOS could change the local redox state. The altered redox state could signal the nucleus to upregulate mitochondrial biogenesis to increase oxidation of the excess reducing equivalents. The increased replication and turnover of mtDNAs within a cell could, with a certain probability, favor a mutant mtDNA, which would become progressively enriched and lead to a functional mitochondrial defect.

Enrichment of mutant mtDNAs within the mtDNA female germline has a different mechanism. The mammalian oocyte contains several hundred thousand mtDNAs. After fertilization and up to the blastocyst stage, the mtDNAs do not actively replicate but become distributed into the blastocyst cells. Hence, the resulting primordial germ cells contain only a very few mtDNA, estimates ranging from a couple to a couple hundred. Subsequent mtDNA replication in the derived oögonia leads to proto-oocytes with reexpanded mtDNA

populations of several thousand mtDNAs. The contraction and expansion of the intracellular mtDNA populations cause rapid genetic drift of heteroplasmic mtDNAs generating proto-oocytes enriched for either the mutant or normal mtDNAs [77].

The proto-oocytes and/or oocytes with the most severe mtDNA mutations can then be selectively eliminated before or soon after fertilization. This is possible because, unlike anatomical alterations that require developmental elaboration of structures before they can be acted on by selection, mitochondrial physiological alterations are expressed at the single-cell level. Hence, cells with highly deleterious mtDNA mutations can be detected and eliminated within the ovary [78–80]. This permits the mtDNAs of mammalian species to have a high mutation rate without the accumulation of large numbers of deleterious mutations (excessive genetic load). Through this system bioenergetic variation is continuously introduced into the population, thus providing a powerful tool for animal adaptation to changing environments [81].

10.3.2 Genetics of nDNA Mitochondrial Genes

The 1 to 2000 nDNA genes of the mitochondrial genome encompass all of the polypeptide genes for mitochondrial replication, transcription, and translation; mitochondrial intermediate metabolism; mitochondrial dynamics and mitophagy; and regulation [82]. These proteins must be synthesized on cytoplasmic ribosomes and the polypeptides imported into the mitochondrion. Several import systems have been identified to target cytosolic ribosome synthesized polypeptides into the mitochondrion. The best characterized systems are the transport through the outer mitochondrial membrane (TOM) complex, the transport through the inner mitochondrial membrane (TIM22 and TIM23) complexes, and the Mia40–Erv1 intermembrane space complex. The TOM–TIM23 complexes mediate the import of polypeptides having an N-terminal mitochondrial targeting peptide, the TOM–TIM22 complexes integrate the carrier proteins including the ANTs into the inner membrane, and the Mia40–Erv1 complex facilitates the import for disulfide containing proteins [83].

Concurrent with elaboration of the rules of mtDNA genetics and their relevance to disease in the 1980s, efforts were underway to clone and characterize essential nDNA OXPHOS genes such as the ANTs and the β subunit of complex V [69,70,84–86]. As knowledge of

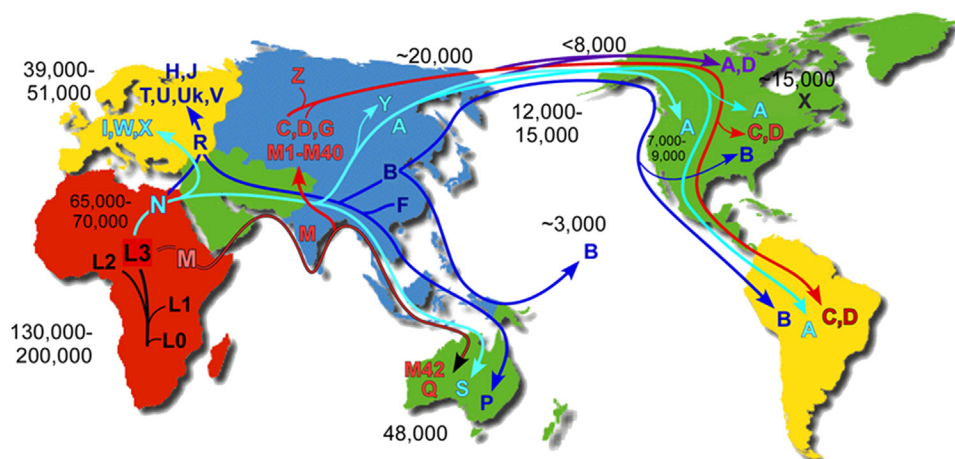


Figure 10.3 Regional radiation of human mtDNAs from their origin in Africa and colonization of Eurasia and the Americas implies that environmental selection constrained regional mtDNA variation. All African mtDNAs are subsumed under macrohaplogroup L and coalesce to a single origin about 130,000–200,000 YBP. African haplogroup L0 is the most ancient mtDNA lineage found in the Koi-San peoples, L1 and L2 in Pygmy populations. The M and N mtDNA lineages emerged from Sub-Saharan African L3 in northeastern Africa, and only derivatives of M and N mtDNAs successfully left Africa, giving rise to macrohaplogroups M and N. N haplogroups radiated into European and Asian indigenous populations, while M haplogroups were confined to Asia. Haplogroups A, C, and D became enriched in northeastern Siberia and were positioned to migrate across the Bering land bridge 20,000 YBP to found Native Americans. Additional Eurasian migrations brought to the Americas haplogroups B and X. Finally, haplogroup B colonized the Pacific Islands. (Figure reproduced from MITOMAP. A Human Mitochondrial Genome Database, 2018. <http://www.mitomap.org>.)

the nDNA mitochondrial genes accumulated, it became possible to identify nDNA mutations associated with mitochondrial disease [87] and then diseases resulting from faulty interactions of mtDNA and nDNA mitochondrial gene variants [88].

10.4 mtDNA AND HUMAN ORIGINS

The human mtDNA sequence is highly polymorphic, and a survey of mtDNA variation among indigenous populations revealed that different populations have population-specific mtDNA variants [89,90]. Because of its high mutation rate and maternal transmission, it followed that by characterizing the mtDNA sequence variation of indigenous populations around the world and incorporating the mtDNA changes into a sequential mutational tree, it would be possible to determine the genetic relationship of all human populations through the maternal lineage. The result is the reconstruction of the origins and ancient migrations of women [90–93].

The detailed characterization of regional population mtDNAs led to the discovery of regional clusters of

mtDNA haplotypes, designated haplogroups. These, in turn, were shown to be founded by one or more functional mtDNA mutations that altered energy metabolism in ways that were adaptive for the regional environment and thus were enriched by natural selection. This created regionally localized clusters of related haplotypes, haplogroups [94–96] (Fig. 10.3).

The global survey of mtDNA variation revealed that the greatest mtDNA variation was found in Africans [91] leading to the conclusion that human mtDNAs originated in Africa approximately 150,000–200,000 years before the present (YBP) [91,97,98]. Because all of the African haplogroups were derived from a common origin, they have been clustered together into macrohaplogroup L [99]. Of all of the African mtDNA variations, only two mtDNA lineages, which arose in Ethiopia [100], left Africa about 65,000 YBP to colonize the rest of the world founding the Eurasian lineages macrohaplogroups M and N [90,94]. The N lineage moved directly north and westward to found all of the European mtDNA haplogroups (H, I, J, Uk, T, U, V, W, and X [101]) and eastward to establish Asian haplogroups

such as A, B, F, etc. Macrohaplogroup M migrated out of Africa through India and Southeast Asia to Australia. Subsequently, various derivative lineages from macrohaplogroup M moved northward to form multiple Central Asian haplogroups, including C, D, E, G, and multiple additional numbered M haplogroups [90].

All Native American mtDNAs were derived from only four central Asian haplogroups: A, B, C, and D [102] plus X [103]. Native American lineages A, C, and D became enriched in eastern Siberia between 30,000 and 40,000 YBP and consequently were in position to cross the Bering land bridge into the Americas, when it became available about 20,000 YBP. Subsequent migrations brought haplogroups B and X to the Americas [90].

The facts that only two mtDNA lineages, M and N, left Africa to colonize the rest of the world and that only three Asian mtDNAs, A, C, and D, migrated north to colonize northeastern Siberia strongly support an adaptive selection hypothesis [81]. Moreover, haplogroups founded by functional mtDNA variants are distributed throughout the human mtDNA phylogeny, consistent with their regional enrichment as women migrated through alternative environments. For example, at the macrohaplogroup level, the out-of-Africa macrohaplogroup N was founded and is defined by two amino acid variants: *ND3* nucleotide (nt) 10389G>A (A114T) and *ATP6* nt 8701G>A (A59T). These variants have been associated with altered mitochondrial membrane potential and Ca^{2+} regulation [104] suggesting that they may have changed the OXPHOS coupling efficiency. This would have increased mitochondrial heat production permitting macrohaplogroup N individuals to move into the colder temperate zone. By contrast, individuals with M mtDNAs, which lack such functional variants, remained in the tropics.

As macrohaplogroup N radiated within Europe, haplogroups J and U arose. Haplogroup J was founded by the reversion of the N-defining *ND3* 10,389G>A variant and the acquisition of a new *ND5* nt 13,708G>A (A458T) variant. Further, haplogroup J radiation gave rise to subhaplogroup J1c, which harbors a cytochrome b variant at 14,798T>C (F18L) and subhaplogroup J2 with a cytochrome b variant at 15,257G>A (D171N). Haplogroup U was founded by the tRNA^{Leu(CUN)} 12,308A>G variant, and its further radiation gave rise to subhaplogroup Uk, which encompasses the *ATP6* 9055G>A (A177T) and cytochrome b 14,798T>C (F18L) variants [95,96]. Because of the sequential accumulation of functional

variants, the physiological effects of mtDNA variants are manifest in the context of the whole mtDNA haplotype and derivative mtDNAs of the regional haplogroups.

10.5 mtDNA CODED MITOCHONDRIAL DISEASES

There are three classes of clinically relevant mtDNA variation: maternally inherited deleterious mutations, ancient adaptive/maladaptive polymorphisms, and developmental and somatic tissue mutations.

10.5.1 Maternally Inherited mtDNA Diseases

Deleterious female germline mtDNA mutations can be either heteroplasmic or homoplasmic depending on how long it has been since the mutation arose, its nature, and its severity. Maternally inherited disease mutations have been reported in polypeptide genes, protein synthesis genes, and regulatory elements (Table 10.1).

Classic examples of maternally inherited mtDNA polypeptide missense mutations are LHON [7,105] and NARP [106]. The common LHON mutations are generally mild and become associated with clinical manifestations when they approach homoplasmy. Hence, LHON gives a more stereotypic phenotypic presentation. By contrast, NARP is more deleterious and can induce clinical manifestations when heteroplasmic. As a result, it can manifest as a range of phenotypes due to varying heteroplasmy levels.

LHON is one of the most common mtDNA diseases with an estimated frequency of greater than 1:7000 [107]. Most LHON mutations occur in the mtDNA complex I genes, with the three most common being the *ND4* gene mutation at nt 11,778G>A (R340H) [7], the *ND1* 3460G>A (A52T) mutation [108], and the *ND6* 14,484T>C (M64V) mutation [109]. While the common LHON mutations are generally mild and homoplasmic, more severe complex I mutations such as *ND6* 14,459G>A (A72V) [110] and *ND6* 14,600G>A (P25L) [111] can be manifest as optic atrophy when heteroplasmic but basal ganglia degeneration, dystonia, and Leigh syndrome when approaching homoplasmy.

The most common NARP mutations occur in the *ATP6* gene at nt 8993T>G (L156R) [106] or 8993T>C (L156P) [112,113]. When homoplasmic, the 8993T>G mutation results in a 70% inhibition of ADP stimulated respiration rate [114]. This mutation has little phenotypic consequence at low-level heteroplasmy but can

TABLE 10.1 Confirmed Mitochondrial DNA Mutations

PANEL A: CODING REGION MUTATIONS						
Locus	Disease Presentations	Mutation	Amino Acid Change	Homoplasmy Reported	Heteroplasmy Reported	References
<i>MT-ND1</i>	LHON-MELAS overlap	m.3376G>A	E>K	+	+	a
<i>MT-ND1</i>	LHON	m.3460G>A	A>T	+	+	b
<i>MT-ND1</i>	LHON	m.3635G>A	S>N	+	–	c
<i>MT-ND1</i>	MELAS/LS/LDYT	m.3697G>A	G>S	+	+	d
<i>MT-ND1</i>	LHON	m.3700G>A	A>T	+	–	e
<i>MT-ND1</i>	LHON	m.3733G>A	E>K	+	+	f
<i>MT-ND1</i>	Progressive encephalomyopathy/LS/optic atrophy	m.3890G>A	R>Q	–	+	g
<i>MT-ND1</i>	EXIT + myalgia/others	m.3902inversionACCTTGC	DLA>GKV	–	+	h
<i>MT-ND1</i>	LHON	m.4171C>A	L>M	+	+	i
<i>MT-CO1</i>	SNHL	m.7445A>G	Ter>Ter	+	+	j
<i>MT-ATP8</i> <i>MT-ATP6</i>	Infantile cardiomyopathy	m.8528T>C	ATP8:W>R; ATP6:M (start)>T	+	+	k
<i>MT-ATP6</i>	NARP/LS/others	m.8993T>C	L>P	–	+	l
<i>MT-ATP6</i>	NARP/LS/MILS/other	m.8993T>G	L>R	–	+	m
<i>MT-ATP6</i>	Ataxia syndromes	m.9035T>C	L>P	+	+	n
<i>MT-ATP6</i>	FBSN/LS/spastic paraplegia	m.9176T>C	L>P	+	+	o
<i>MT-ATP6</i>	LS/ataxia syndromes	m.9176T>G	L>R	–	+	p
<i>MT-ATP6</i>	NARP-like disease	m.9185T>C	L>P	+	+	q
<i>MT-ATP6</i>	Encephalopathy/seizures/lactic acidemia	m.9205TA>del	Ter>M	+	–	r
<i>MT-ND3</i>	LS	m.10158T>C	S>P	+	+	s
<i>MT-ND3</i>	LS/Leigh-like disease/ESOC	m.10191T>C	S>P	–	+	t
<i>MT-ND3</i>	LS/dystonia/stroke/LDYT	m.10197G>A	A>T	+	+	u
<i>MT-ND4</i>	LS/encephalopathy	m.11777C>A	R>S	–	+	v
<i>MT-ND4</i>	LHON/progressive dystonia	m.11778G>A	R>H	+	+	w
<i>MT-ND5</i>	LS	m.12706T>C	F>L	–	+	x
<i>MT-ND5</i>	Optic neuropathy/LS/MERRF-MELAS	m.13042G>A	A>T	–	+	y
<i>MT-ND5</i>	LHON/Leigh-like neurodegeneration	m.13051G>A	G>S	+	–	z
<i>MT-ND5</i>	LS/MELAS/LHON-MELAS overlap syndrome	m.13513G>A	D>N	–	+	aa
<i>MT-ND5</i>	LS/MELAS	m.13514A>G	D>G	–	+	bb
<i>MT-ND6</i>	LDYT/LS	m.14459G>A	A>V	+	+	cc
<i>MT-ND6</i>	LHON	m.14482C>A	M>I	+	+	dd
<i>MT-ND6</i>	LHON	m.14482C>G	M>I	+	+	ee
<i>MT-ND6</i>	LHON	m.14484T>C	M>V	+	+	ff
<i>MT-ND6</i>	Dystonia/LS/ataxia/ptosis/epilepsy	m.14487T>C	M>V	–	+	gg

Continued

TABLE 10.1 Confirmed Mitochondrial DNA Mutations—cont'd

PANEL A: CODING REGION MUTATIONS						
Locus	Disease Presentations	Mutation	Amino Acid Change	Homoplasmy Reported	Heteroplasmy Reported	References
<i>MT-ND6</i>	LHON	m.14495A>G	L>S	–	+	hh
<i>MT-ND6</i>	LHON	m.14568C>T	G>S	+	–	ii
<i>MT-CYB</i>	EXIT/septo-optic dysplasia	m.14849T>C	S>P	–	+	jj
<i>MT-CYB</i>	MELAS	m.14864T>C	C>R	–	+	kk
PANEL B: tRNA AND rRNA MUTATIONS						
Locus	Disease Presentations	Mutation	RNA	Homoplasmy Reported	Heteroplasmy Reported	References
<i>MT-TF</i>	MELAS/ MM + EXIT	m.583G>A	tRNA Phe	–	+	ll
<i>MT-RNR1</i>	DEAF	m.1494C>T	12S rRNA	+	–	mm
<i>MT-RNR1</i>	DEAF	m.1555A>G	12S rRNA	+	–	nn
<i>MT-TV</i>	AMDF	m.1606G>A	tRNA Val	–	+	oo
<i>MT-TV</i>	LS/HCM/ MELAS	m.1644G>A	tRNA Val	–	+	pp
<i>MT-TL1</i>	MELAS/LS/ DMDF/MIDD/ SNHL/CPEO/ MM/others	m.3243A>G	tRNA Leu (UUR)	–	+	qq
<i>MT-TL1</i>	MM/MELAS/ SNHL/CPEO	m.3243A>T	tRNA Leu (UUR)	–	+	rr
<i>MT-TL1</i>	MELAS	m.3256C>T	tRNA Leu (UUR)	–	+	ss
<i>MT-TL1</i>	MELAS/ myopathy	m.3258T>C	tRNA Leu (UUR)	–	+	tt
<i>MT-TL1</i>	MMC/MELAS	m.3260A>G	tRNA Leu (UUR)	–	+	uu
<i>MT-TL1</i>	MELAS/DM	m.3271T>C	tRNA Leu (UUR)	–	+	vv
<i>MT-TL1</i>	PEM	m.3271T>del	tRNA Leu (UUR)	–	+	ww
<i>MT-TL1</i>	Myopathy	m.3280A>G	tRNA Leu (UUR)	–	+	xx
<i>MT-TL1</i>	MELAS/ myopathy/ deafness + cognitive impairment	m.3291T>C	tRNA Leu (UUR)	–	+	yy

TABLE 10.1 Confirmed Mitochondrial DNA Mutations—cont'd

PANEL B: tRNA AND rRNA MUTATIONS						
Locus	Disease Presentations	Mutation	RNA	Homoplasmy Reported	Heteroplasmy Reported	References
<i>MT-TL1</i>	MM	m.3302A>G	tRNA Leu (UUR)	—	+	zz
<i>MT-TL1</i>	MMC	m.3303C>T	tRNA Leu (UUR)	+	+	aaa
<i>MT-TI</i>	CPEO/MS	m.4298G>A	tRNA Ile	—	+	bbb
<i>MT-TI</i>	MICM	m.4300A>G	tRNA Ile	+	+	ccc
<i>MT-TI</i>	CPEO	m.4308G>A	tRNA Ile	—	+	ddd
<i>MT-TQ</i>	Encephalopathy/MELAS	m.4332G>A	tRNA Gln	—	+	eee
<i>MT-TW</i>	LS	m.5537A>AT	tRNA Trp	—	+	fff
<i>MT-TA</i>	Myopathy	m.5650G>A	tRNA Ala	—	+	ggg
<i>MT-TN</i>	CPEO + ptosis + proximal myopathy	m.5690A>G	tRNA Asn	—	+	hhh
<i>MT-TN</i>	CPEO/MM	m.5703G>A	tRNA Asn	—	+	iii
<i>MT-TS1</i>	PEM/AMDF/motor neuron disease-like	m.7471C>CC	tRNA Ser (UCN)	+	+	jjj
<i>MT-TS1</i>	MM/EXIT	m.7497G>A	tRNA Ser (UCN)	+	+	kkk
<i>MT-TS1</i>	SNHL	m.7510T>C	tRNA Ser (UCN)	—	+	lll
<i>MT-TS1</i>	SNHL	m.7511T>C	tRNA Ser (UCN)	+	+	mmm
<i>MT-TK</i>	MERRF	m.8344A>G	tRNA Lys	—	+	nnn
<i>MT-TK</i>	MERRF	m.8356T>C	tRNA Lys	—	+	ooo
<i>MT-TK</i>	MICM + DEAF/MERRF/autism/LS/ataxia + lipomas	m.8363G>A	tRNA Lys	—	+	ppp
<i>MT-TG</i>	PEM	m.10010T>C	tRNA Gly	—	+	qqq
<i>MT-TH</i>	MERRF-MELAS/encephalopathy	m.12147G>A	tRNA His	—	+	rrr
<i>MT-TL2</i>	CPEO	m.12276G>A	tRNA Leu (CUN)	—	+	sss
<i>MT-TL2</i>	CPEO/KSS	m.12315G>A	tRNA Leu (CUN)	—	+	ttt
<i>MT-TL2</i>	CPEO	m.12316G>A	tRNA Leu (CUN)	—	+	uuu

Continued

TABLE 10.1 Confirmed Mitochondrial DNA Mutations—cont'd

PANEL B: tRNA AND rRNA MUTATIONS						
Locus	Disease Presentations	Mutation	RNA	Homoplasmy Reported	Heteroplasmy Reported	References
<i>MT-TE</i>	Reversible COX deficiency myopathy	m.14674T>C	tRNA Glu	+	—	vvv
<i>MT-TE</i>	MM + DMDF/encephalo-myopathy/dementia + diabetes + ophthalmoplegia	m.14709T>C	tRNA Glu	+	+	www

See <http://www.mitomap.org/MITOMAP/MutationsCodingControl> and <http://www.mitomap.org/MITOMAP/MutationsRNA> for additional reports and phenotypes. *ADPD*, Alzheimer disease and Parkinson disease; *AMDF*, ataxia, myopathy, and deafness; *COX*, cytochrome c oxidase; *CPEO*, chronic progressive ophthalmoplegia; *DEAF/SNHL*, deafness/sensorineural hearing loss; *DEMCHO*, dementia and chorea; *DM*, diabetes mellitus; *DMDF*, diabetes mellitus and deafness; *ESOC*, epilepsy, strokes, optic atrophy, and cognitive decline; *EXIT*, exercise intolerance; *FBSN*, familial bilateral striatal necrosis; *FSGS*, focal segmental glomerulosclerosis; *GER*, gastrointestinal reflux; *HCM*, hypertrophic cardiomyopathy; *LDYT*, LHON + dystonia; *LHON*, Leber hereditary optic neuropathy; *LS*, Leigh syndrome; *MELAS*, mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes; *MERRF*, myoclonic epilepsy and ragged red fiber disease; *MICM*, maternally inherited cardiomyopathy; *MIDD*, maternally inherited diabetes and deafness; *MILS*, maternally inherited Leigh syndrome; *MM*, mitochondrial myopathy; *MMC*, mitochondrial myopathy and cardiomyopathy; *MNGIE*, mitochondrial neurogastrointestinal encephalopathy; *MS*, multiple sclerosis; *NARP*, neurogenic muscle weakness, ataxia, and retinitis pigmentosa; *PEM*, progressive encephalomyopathy; *SNHL*, sensorineural hearing loss, +, reported; —, not reported.

^aBlakely EL, de Silva R, King A, Schwarzer V, Harrower T, Dawidek G, Turnbull DM, Taylor RW. LHON/MELAS overlap syndrome associated with a mitochondrial MTND1 gene mutation. *Eur J Hum Genet* 2005;13:623–27.

^bHowell N, Bindoff LA, McCullough DA, Kubacka I, Poulton J, Mackey D, Taylor L, Turnbull DM. Leber hereditary optic neuropathy: identification of the same mitochondrial ND1 mutation in six pedigrees. *Am J Hum Genet* 1991a;49:939–50; Huoponen K, Vilkki J, Aula P, Nikoskelainen EK, Savontaus ML. A new mtDNA mutation associated with Leber hereditary optic neuroretinopathy. *Am J Hum Genet* 1991;48:1147–53.

^cBrown MD, Zhadanov S, Allen JC, Hosseini S, Newman NJ, Atamonov VV, Mikhailovskaya IE, Sukernik RI, Wallace DC. Novel mtDNA mutations and oxidative phosphorylation dysfunction in Russian LHON families. *Hum Genet* 2001;109:33–9.

^dKirby DM, McFarland R, Ohtake A, Dunning C, Ryan MT, Wilson C, Ketteridge D, Turnbull DM, Thorburn DR, Taylor RW. Mutations of the mitochondrial ND1 gene as a cause of MELAS. *J Med Genet* 2004;41:784–9.

^eAchilli A, Iommarini L, Olivieri A, Pala M, Kashani BH, Reynier P, La Morgia C, Valentino ML, Liguori R, Pizza F, Barboni P, Sadun F, De Negri A, Zeviani M, Dollfus H, Moulignier A, Ducos G, Orssaud C, Bonneau D, Procaccio V, Leo-Kottler B, Fauser S, Wissinger B, Amati-Bonneau P, Torroni A, Carelli V. Rare primary mitochondrial DNA mutations and synergistic variants in Leber's Hereditary Optic Neuropathy. *PLoS One* 2012;7:e42242, Valentino ML, Barboni P, Ghelli A, Bucci L, Rengo C, Achilli A, Torroni A, Liguori R, Lodi R, Barbiroli B, Dotti M, Federico A, Baruzzi A, Carelli V. The ND1 gene of complex I is a mutational hot spot for Leber's hereditary optic neuropathy. *Biochem Biophys Res Commun* 2002b;295:342–7.

^fAchilli A, Iommarini L, Olivieri A, Pala M, Kashani BH, Reynier P, La Morgia C, Valentino ML, Liguori R, Pizza F, Barboni P, Sadun F, De Negri A, Zeviani M, Dollfus H, Moulignier A, Ducos G, Orssaud C, Bonneau D, Procaccio V, Leo-Kottler B, Fauser S, Wissinger B, Amati-Bonneau P, Torroni A, Carelli V. Rare primary mitochondrial DNA mutations and synergistic variants in Leber's Hereditary Optic Neuropathy. *PLoS One* 2012;7:e42242, Valentino ML, Barboni P, Ghelli A, Bucci L, Rengo C, Achilli A, Torroni A, Liguori R, Lodi R, Barbiroli B, Dotti M, Federico A, Baruzzi A, Carelli V. The ND1 gene of complex I is a mutational hot spot for Leber's hereditary optic neuropathy. *Ann Neurol* 2004;56:631–41.

^gCaporali L, Ghelli AM, Iommarini L, Maresca A, Valentino ML, La Morgia C, Liguori R, Zanna C, Barboni P, De Nardo V, Martinuzzi A, Rizzo G, Tonon C, Lodi R, Calvaruso MA, Cappelletti M, Porcelli AM, Achilli A, Pala M, Torroni A, Carelli V. Cybrid studies establish the causal link between the mtDNA m.3890G>A/MT-ND1 mutation and optic atrophy with bilateral brainstem lesions. *Biochim Biophys Acta* 2013;1832:445–52.

^hBlakely EL, Rennie KJ, Jones L, Elstner M, Chrzanowska-Lightowlers ZM, White CB, Shield JP, Pilz DT, Turnbull DM, Poulton J, Taylor RW. Sporadic intragenic inversion of the mitochondrial DNA MTND1 gene causing fatal infantile lactic acidosis. *Pediatric*

- Research 2006;59:440–4. Musumeci O, Andreu AL, Shanske S, Bresolin N, Comi GP, Rothstein R, Schon EA, DiMauro S. Intragenic inversion of mtDNA: a new type of pathogenic mutation in a patient with mitochondrial myopathy. *Am J Hum Genet* 2000;66:1900–4.
- ^kKim JY, Hwang JM, Park SS. Mitochondrial DNA C4171A/ND1 is a novel primary causative mutation of Leber's hereditary optic neuropathy with a good prognosis. *Ann Neurol* 2002;51:630–4.
- ^lReid FM, Vernham GA, Jacobs HT. A novel mitochondrial point mutation in a maternal pedigree with sensorineural deafness. *Human Mutat* 1994;3:243–7.
- ^kWare SM, El-Hassan N, Kahler SG, Zhang Q, Ma YW, Miller E, Wong B, Spicer RL, Craigen WJ, Kozel BA, Grange DK, Wong LJ. Infantile cardiomyopathy caused by a mutation in the overlapping region of mitochondrial ATPase six and eight genes. *J Med Genet* 2009;46:308–14.
- ^lDe Vries DD, Van Engelen BG, Gabreels FJ, Ruitenbeek W, Van Oost BA. A second missense mutation in the mitochondrial ATPase six gene in Leigh's syndrome. *Ann Neurol* 1993;34:410–2.
- ^mHarding, AE Holt IJ, Sweeney MG, Brockington M, Davis MB. Prenatal diagnosis of mitochondrial DNA8993 T-G disease. *Am J Hum Genet* 1992;50:629–33. Holt IJ, Harding AE, Petty RK, Morgan-Hughes JA. A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *Am J Hum Genet* 1990;46:428–33.
- ⁿPfeffer, G Blakely, EL Alston, CL Hassani A, Boggild M, Horvath R, Samuels DC, Taylor RW, Chinnery PF. Adult-onset spinocerebellar ataxia syndromes due to MTATP6 mutations *J Neurol Neurosurg Psychiatry* 2012;83:883–6, Sikorska M, Sandhu JK, Simon DK, Pathiraja V, Sodja C, Li Y, Ribocco-Lutkiewicz M, Lanthier P, Borowy-Borowski H, Upton A, Raha S, Pulst SM. Tarnopolsky MA Identification of ataxia-associated mtDNA mutations (m.4452T>C and m.9035T>C) and evaluation of their pathogenicity in trans-mitochondrial cybrids. *Muscle Nerve* 2009;40:381–94.
- ^oThyagarajan D, Shanske S, Vazquez-Memije M, De Vivo D, DiMauro S. A novel mitochondrial ATPase six point mutation in familial bilateral striatal necrosis. *Ann Neurol* 1995;38:468–72, Verny C, Guegen N, Desquiret V, Chevrolier A, Prundean A, Dubas F, Cassereau J, Ferre M, Amati-Bonneau P, Bonneau D, Reynier P, Procaccio V. Hereditary spastic paraplegia-like disorder due to a mitochondrial ATP6 gene point mutation. *Mitochondrion* 2011;11:70–5.
- ^pCarrozzo R, Murray J, Santorelli FM, Capaldi RA. The T9176G mutation of human mtDNA gives a fully assembled but inactive ATP synthase when modeled in *Escherichia coli*. *FEBS Letters* 2000;486:297–9.
- ^qMoslemi AR, Darin N, Tulinius M, Oldfors A, Holme E. Two new mutations in the MTATP6 gene associated with Leigh syndrome. *Neuropediatrics* 2005;36:314–8.
- ^rTemperley RJ, Seneca SH, Tonska K, Bartnik E, Bindoff LA, Lightowlers RN, Chrzanowska-Lightowlers ZM. Investigation of a pathogenic mtDNA microdeletion reveals a translation-dependent deadenylation decay pathway in human mitochondria. *Hum Mol Genet* 2003;12:2341–8.
- ^sCrimi M, Papadimitriou A, Galbiati S, Palamidou P, Fortunato F, Bordini A, Papandreou U, Papadimitriou D, Hadjigeorgiou GM, Drogari E, Bresolin N, Comi GP. A new mitochondrial DNA mutation in ND3 gene causing severe Leigh Syndrome with early lethality. *Pediatric Research* 2004;55:842–6, Lebon S, Chol M, Benit P, Mugnier C, Chretien D, Giurgea I, Kern I, Girardin E, Hertz-Pannier L, de Lonlay P, Rotig A, Rustin P, Munnich A. Recurrent de novo mitochondrial DNA mutations in respiratory chain deficiency. *J Med Genet* 2003;40:896–9, McFarland R, Kirby DM, Fowler KJ, Ohtake A, Ryan MT, Amor DJ, Fletcher JM, Dixon JW, Collins FA, Turnbull DM, Taylor RW, Thorburn DR. De novo mutations in the mitochondrial ND3 gene as a cause of infantile mitochondrial encephalopathy and complex I deficiency. *Ann Neurol* 2004;55:58–64.
- ^tTaylor RW, Singh-Kler R, Hayes CM, Smith PE, Turnbull DM. Progressive mitochondrial disease resulting from a novel missense mutation in the mitochondrial DNA ND3 gene. *Ann Neurol* 2001;50:104–7.
- ^uKirby DM, McFarland R, Ohtake A, Dunning C, Ryan MT, Wilson C, Ketteridge D, Turnbull DM, Thorburn DR, Taylor RW. Mutations of the mitochondrial ND1 gene as a cause of MELAS. *Journal of Medical Genetics* 2004;41:784–9, Sarzi E, Brown M, Lebon S, Chretien D, Munnich A, Rotig A, Procaccio V. A novel recurrent mitochondrial DNA mutation in ND3 gene is associated with isolated complex I deficiency causing Leigh syndrome and dystonia. *American Journal of Medical Genetics* 2007;143A:33–41.
- ^vDeschauer M, Bamberg C, Claus D, Zierz S, Turnbull DM, Taylor RW. Late-onset encephalopathy associated with a C11777A mutation of mitochondrial DNA. *Neurology* 2003;60:1357–9, Komaki H, Akanuma J, Iwata H, Takahashi T, Mashima Y, Nonaka I, Goto Y. A novel mtDNA C11777A mutation in Leigh syndrome. *Mitochondrion* 2003;2:293–304.
- ^wWallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AM, Elsas LJ, Nikoskelainen EK. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* 1988a;242:1427–30.
- ^xTaylor RW, Morris AA, Hutchinson M, Turnbull DM. Leigh disease associated with a novel mitochondrial DNA ND5 mutation. *Eur J Hum Genet* 2002;10:141–4.
- ^yNaini AB, Lu J, Kaufmann P, Bernstein RA, Mancuso M, Bonilla E, Hirano M, DiMauro S. Novel mitochondrial DNA ND5 mutation in a patient with clinical features of MELAS and MERRF. *Arch Neurol* 2005;62:473–6, Valentino ML, Barboni P, Rengo C, Achilli A, Torroni A, Lodi R, Tonon C, Barbiroli B, Fortuna F, Montagna P, Baruzzi A, Carelli V. The 13,042G→A/ND5 mutation in mtDNA is pathogenic and can be associated also with a prevalent ocular phenotype. *J Med Genet* 2006;43:e38.
- ^zDombi E, Diot A, Morten K, Carver J, Lodge T, Fratter C, Ng YS, Liao C, Muir R, Blakely EL, Hargreaves I, Al-Dosary M, Sarkar G, Hickman SJ, Downes SM, Jayawant S, Yu-Wai-Man P, Taylor RW, Poulton J. The m.13,051G>A mitochondrial DNA mutation results in variable neurology and activated mitophagy. *Neurology* 2016;86:1921–23, Howell N, Oostra RJ, Bolhuis PA, Spruijt L, Clarke LA,

- Mackey DA, Preston G, Herrnsstadt C. Sequence analysis of the mitochondrial genomes from Dutch pedigrees with Leber hereditary optic neuropathy. *Am J Hum Genet* 2003;72:1460–9.
- ^{aa}Santorelli FM, Tanji K, Kulikova R, Shanske S, Vilarinho L, Hays AP, DiMauro S. Identification of a novel mutation in the mtDNA ND5 gene associated with MELAS. *Biochem Biophys Res Commun* 1997a;238:326–328.
- ^{bb}Corona P, Antozzi C, Carrara F, D'Incerti L, Lamantea E, Tiranti V, Zeviani M. A novel mtDNA mutation in the ND5 subunit of complex I in two MELAS patients. *Ann Neurol* 2001;49:106–10.
- ^{cc}Jun AS, Brown MD, Wallace DC. A mitochondrial DNA mutation at np 14,459 of the ND6 gene associated with maternally inherited Leber's hereditary optic neuropathy and dystonia. *Proc Natl Acad Sci USA* 1994;91:6206–10, Kirby DM, Kahler SG, Freckmann ML, Reddihough D, Thorburn DR. Leigh disease caused by the mitochondrial DNA G14459A mutation in unrelated families. *Ann Neurol* 2000;48:102–4.
- ^{dd}Achilli A, Iommarini L, Olivieri A, Pala M, Kashani BH, Reynier P, La Morgia C, Valentino ML, Liguori R, Pizzo F, Barboni P, Sadun F, De Negri A, Zeviani M, Dollfus H, Moulignier A, Ducos G, Orssaud C, Bonneau D, Procaccio V, Leo-Kottler B, Fauser S, Wissinger B, Amati-Bonneau P, Torroni A, Carelli V. Rare primary mitochondrial DNA mutations and synergistic variants in Leber's Hereditary Optic Neuropathy. *PLoS One* 2012;7:e42242, Valentino ML, Avoni P, Barboni P, Pallotti F, Rengo C, Torroni A, Bellan M, Baruzzi A, Carelli V. Mitochondrial DNA nucleotide changes C14482G and C14482A in the ND6 gene are pathogenic for Leber's hereditary optic neuropathy. *Ann Neurol* 2002;51:774–8.
- ^{ee}Howell N, Bogolin C, Jamieson R, Marenda DR, Mackey DA. mtDNA mutations that cause optic neuropathy: how do we know? *Am J Hum Genet* 1998;62:196–202.
- ^{ff}Brown MD, Voljavec AS, Lott MT, MacDonald I, Wallace DC. Leber's hereditary optic neuropathy: a model for mitochondrial neurodegenerative diseases. *FASEB J* 1992;6:2791–9, Howell N, Kubacka I, Xu M, McCullough DA. Leber hereditary optic neuropathy: involvement of the mitochondrial ND1 gene and evidence for an intragenic suppressor mutation. *Am J Hum Genet* 1991b;48:935–42, Johns DR, Neufeld MJ, Park RD. An ND-6 mitochondrial DNA mutation associated with Leber hereditary optic neuropathy. *Biochem Biophys Res Commun* 1992;187:1551–7.
- ^{gg}Solano A, Roig M, Vives-Bauza C, Hernandez-Pena J, Garcia-Arumi E, Playan A, Lopez-Perez MJ, Andreu AL, Montoya J. Bilateral striatal necrosis associated with a novel mutation in the mitochondrial ND6 gene. *Ann Neurol* 2003;54:527–30, Ugalde C, Triepels RH, Coenen MJ, van den Heuvel LP, Smeets R, Uusimaa J, Briones P, Campistol J, Majamaa K, Smeitink JA, Nijtmans LG. Impaired complex I assembly in a Leigh syndrome patient with a novel missense mutation in the ND6 gene. *Ann Neurol* 2003;54:665–9.
- ^{hh}Chinnery PF, Brown DT, Andrews RM, Singh-Kler R, Riordan-Eva P, Lindley J, Applegarth DA, Turnbull DM, Howell N. The mitochondrial ND6 gene is a hot spot for mutations that cause Leber's hereditary optic neuropathy. *Brain* 2001;124:209–18.
- ⁱⁱFauser S, Leo-Kottler B, Besch D, Luberichs J. Confirmation of the 14,568 mutation in the mitochondrial ND6 gene as causative in Leber's hereditary optic neuropathy. *Ophthalmic Genetics* 2002a;23:191–7, Wissinger B, Besch D, Baumann B, Fauser S, Christ-Adler M, Jurklics B, Zrenner E, Leo-Kottler B. Mutation analysis of the ND6 gene in patients with Leber's hereditary optic neuropathy. *Biochem Biophys Res Commun* 1997;234:511–5.
- ^{jj}Schuelke M, Krude H, Finckh B, Mayatepek E, Janssen A, Schmeltz M, Trefz F, Trijbels F, Smeitink J. Septo-optic dysplasia associated with a new mitochondrial cytochrome b mutation. *Ann Neurol* 2002;51:388–92.
- ^{kk}Emmanuele V, Sotiriou E, Rios PG, Ganesh J, Ichord R, Foley AR, Akman HO, DiMauro S. A novel mutation in the mitochondrial DNA cytochrome b gene (MTCYB) in a patient with mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes syndrome. *J Child Neurol* 2013;28:236–42.
- ^{ll}Darin N, Kollberg G, Moslemi AR, Tulinius M, Holme E, Gronlund MA, Andersson S, Oldfors A. Mitochondrial myopathy with exercise intolerance and retinal dystrophy in a sporadic patient with a G583A mutation in the mt tRNA(phe) gene. *Neuromusc Disord* 2006;16:504–6, Hanna MG, Nelson IP, Morgan-Hughes JA, Wood NW. MELAS: a new disease associated mitochondrial DNA mutation and evidence for further genetic heterogeneity. *J Neurol Neurosurg Psychiatry* 1998;65:512–7.
- ^{mm}Zhao H, Li R, Wang Q, Yan Q, Deng JH, Han D, Bai Y, Young WY, Guan MX. Maternally inherited aminoglycoside-induced and nonsyndromic deafness is associated with the novel C1494T mutation in the mitochondrial 12S rRNA gene in a large Chinese family. *Am J Hum Genet* 2004;74:139–152.
- ⁿⁿFischel-Ghodsian N, Prezant TR, Bu X, Oztas S. Mitochondrial ribosomal RNA gene mutation in a patient with sporadic aminoglycoside ototoxicity. *Am J Otolaryngol* 1993;14:399–403, Hutchin T, Haworth I, Higashi K, Fischel-Ghodsian N, Stoneking M, Saha N, Arnos C, Cortopassi G. A molecular basis for human hypersensitivity to aminoglycoside antibiotics. *Nucleic Acids Res* 1993;21:4174–9, Prezant TR, Agopian JV, Bohlman MC, Bu X, Oztas S, Qiu WQ, Arnos KS, Cortopassi GA, Jaber L, Rotter JL, Shohat M, Fischel-Ghodsian N. Mitochondrial ribosomal RNA mutation associated with both antibiotic-induced and non-syndromic deafness. *Nature Genet* 1993;4:289–94.
- ^{oo}Tiranti V, D'Agruma L, Pareyson D, Mora M, Carrara F, Zelante L, Gasparini P, Zeviani M. A novel mutation in the mitochondrial tRNA(Val) gene associated with a complex neurological presentation. *Ann Neurol* 1998;43:98–101.
- ^{pp}Fraidakis MJ, Jardi C, Allouche S, Nelson I, Aure K, Slama A, Lemiere I, Thenint JP, Hamon JB, Zagnoli F, Heron D, Sedel F, Lombes A. Phenotypic diversity associated with the MT-TV gene m.1644G>A mutation, a matter of quantity. *Mitochondrion* 2014;15:34–9, Menotti F, Brega A, Diegoli M, Grasso M, Modena MG, Arbustini E. A novel mtDNA point mutation in tRNA(Val) is associated with hypertrophic cardiomyopathy and MELAS. *Ital Heart J* 2004;5:460–5.

- ⁹⁹Goto Y, Nonaka I, Horai S. A mutation in the tRNA^{Leu(UUR)} gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature* 1990;348:651–3, Manouvrier S, Rotig A, Hannebique G, Gheerbrandt JD, Royer-Legrain G, Munnich A, Parent M, Grunfeld JP, Largilliere C, Lombes A, Bonnefont JP. Point mutation of the mitochondrial tRNA^{Leu} gene (A 3243 G) in maternally inherited hypertrophic cardiomyopathy, diabetes mellitus, renal failure, and sensorineural deafness. *J Med Genet* 1995;32:654–6, Massin P, Guillausseau PJ, Vialettes B, Paquis V, Orsini F, Grimaldi AD, Gaudric A. Macular pattern dystrophy associated with a mutation of mitochondrial DNA. *Am J Ophthalmol* 1995;120:247–8, van den Ouweland JM, Lemkes HHP, Ruitenbeek W, Sandkuijl LA, deVijlder MF, Struyvenberg PAA, van de Kamp JJP, Maassen JA. Mutation in mitochondrial tRNA^{Leu(UUR)} gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. *Nat Genet* 1992;1:368–71.
- ¹⁰⁰Shaag A, Saada A, Steinberg A, Navon P, Elpeleg ON. Mitochondrial encephalomyopathy associated with a novel mutation in the mitochondrial tRNA(Leu)(UUR) gene (A3243T). *Biochem Biophys Res Commun* 1997;233:637–9.
- ¹⁰¹Moraes CT, Ciacci F, Bonilla E, Jansen C, Hirano M, Rao N, Lovelace RE, Rowland LP, Schon EA, DiMauro S. Two novel pathogenic mitochondrial DNA mutations affecting organelle number and protein synthesis. Is the tRNA^{Leu(UUR)} gene an etiologic hot spot? *J Clin Invest* 1993;92:2906–15, Sato W, Hayasaka K, Shoji Y, Takahashi T, Takada G, Saito M, Fukawa O, Wachi E. A mitochondrial tRNA(Leu)(UUR) mutation at 3256 associated with mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS). *Biochem Mol Biol Int (Sydney)* 1994;33:1055–61.
- ¹⁰²Sternberg D, Chatzoglou E, Laforet P, Fayet G, Jardel C, Blondy P, Fardeau M, Amselem S, Eymard B, Lombes A. Mitochondrial DNA transfer RNA gene sequence variations in patients with mitochondrial disorders. *Brain* 2001;124:984–94.
- ¹⁰³Sweeney MG, Brockington M, Weston MJ, Morgan-Hughes JA, Harding AE. Mitochondrial DNA transfer RNA mutation Leu(UUR) A-G 3260: a second family with myopathy and cardiomyopathy. *Q J Med* 1993;86:435–8, Zeviani M, Gellera C, Antozzi C, Rimoldi M, Morandi L, Villani F, Tiranti V, DiDonato S. Maternally inherited myopathy and cardiomyopathy: association with mutation in mitochondrial DNA tRNA^{Leu(UUR)}. *Lancet* 1991;338:143–7.
- ¹⁰⁴Goto Y, Nonaka I, Horai S. A new mtDNA mutation associated with mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes (MELAS). *Biochim Biophys Acta* 1991;1097:238–40, Hayashi J, Ohta S, Takai D, Miyabayashi S, Sakuta R, Goto Y, Nonaka I. Accumulation of mtDNA with a mutation at position 3271 in tRNA^{Leu(UUR)} gene introduced from a MELAS patient to HeLa cells lacking mtDNA results in progressive inhibition of mitochondrial respiratory function. *Biochem Biophys Res Commun* 1993;197:1049–55, Sakuta R, Goto Y, Horai S, Nonaka I. Mitochondrial DNA mutations at nucleotide positions 3243 and 3271 in mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes: a comparative study. *J Neurol Sci* 1993;115:158–60, Tsukuda K, Suzuki Y, Kameoka K, Osawa N, Goto Y, Katagiri H, Asano T, Yazaki Y, Oka Y. Screening of patients with maternally transmitted diabetes for mitochondrial gene mutations in the tRNA[Leu(UUR)] region. *Diabet Med* 1997;14:1032–7.
- ¹⁰⁵Shoffner JM, Bialer MG, Pavlakis SG, Lott MT, Kaufman A, Dixon J, Teichberg S, Wallace DC. Mitochondrial encephalomyopathy associated with a single nucleotide pair deletion in the mitochondrial tRNA^{Leu(UUR)} gene. *Neurology* 1995;45:286–92.
- ¹⁰⁶Sternberg D, Chatzoglou E, Laforet P, Fayet G, Jardel C, Blondy P, Fardeau M, Amselem S, Eymard B, Lombes A. Mitochondrial DNA transfer RNA gene sequence variations in patients with mitochondrial disorders. *Brain* 2001;124:984–94.
- ¹⁰⁷Goto Y. Clinical features of MELAS and mitochondrial DNA mutations. *Muscle Nerve* 1995;3:S107–12, Goto Y, Tsugane K, Tanabe Y, Nonaka I, Horai S. A new point mutation at nucleotide pair 3291 of the tRNA^{Leu(UUR)} gene in a patient with mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS). *Biochem Biophys Res Commun* 1994;202:1624–30.
- ¹⁰⁸Bindoff LA, Howell N, Poulton J, McCullough DA, Morten KJ, Lightowlers RN, Turnbull DM, Weber K. Abnormal RNA processing associated with a novel tRNA mutation in mitochondrial DNA. A potential disease mechanism. *J Biol Chem* 1993;268:19559–64, Maniura-Weber K, Helm M, Engemann K, Eckertz S, Mollers M, Schauen M, Hayrapetyan A, von Kleist-Retzow JC, Lightowlers RN, Bindoff LA, Wiesner RJ. Molecular dysfunction associated with the human mitochondrial 3302A>G mutation in the MTTL1 (mt-tRNA^{Leu(UUR)}) gene. *Nucleic Acids Res* 2006;34:6404–15, Shoffner JM, Krawiecki N, Cabell MF, Torroni A, Wallace DC. A novel tRNA^{Leu(UUR)} mutation in childhood mitochondrial myopathy. Poster 949. *Am J Hum Genet* 1993;53:287.
- ¹⁰⁹Silvestri G, Santorelli FM, Shanske S, Whitley CB, Schimmenti LA, Smith SA, DiMauro S. A new mtDNA mutation in the tRNA^{Leu(UUR)} gene associated with maternally inherited cardiomyopathy. *Human Mutat* 1994;3:37–43.
- ¹¹⁰Taylor RW, Chinnery PF, Bates MJ, Jackson MJ, Johnson MA, Andrews RM, Turnbull DM. A novel mitochondrial DNA point mutation in the tRNA(Ile) gene: studies in a patient presenting with chronic progressive external ophthalmoplegia and multiple sclerosis. *Biochem Biophys Res Commun* 1998;243:47–51.
- ¹¹¹Casali C, Santorelli FM, D'Amati G, Bernucci P, DeBiase L, DiMauro S. A novel mtDNA point mutation in maternally inherited cardiomyopathy. *Biochem Biophys Res Commun* 1995;213:588–93.
- ¹¹²Schaller A, Desetty R, Hahn D, Jackson CB, Nuoffer JM, Gallati S, Levinger L. Impairment of mitochondrial tRNA(Ile) processing by a novel mutation associated with chronic progressive external ophthalmoplegia. *Mitochondrion* 2011;11:488–96, Souilem S, Chebel S, Mancuso M, Petrozzi L, Siciliano G, Frihayed M, Hentati F, Amouri R. A novel mitochondrial tRNA(Ile) point mutation associated with chronic progressive external ophthalmoplegia and hyperCKemia. *J Neurol Sci* 2011;300:187–90.
- ¹¹³Bataillard M, Chatzoglou E, Rumbach L, Sternberg D, Tournade A, Laforet P, Jardel C, Maisonneuve T, Lombes A. Atypical MELAS syndrome associated with a new mitochondrial tRNA glutamine point mutation. *Neurology* 2001;56:405–7.
- ¹¹⁴Santorelli FM, Tanji K, Sano M, Shanske S, El-Shahawi M, Kranz-Eble P, DiMauro S, De Vivo DC. Maternally inherited encephalopathy associated with a single-base insertion in the mitochondrial tRNA^{Trp} gene. *Ann Neurol* 1997b;42:256–60.

- ⁹⁹⁹McFarland R, Swalwell H, Blakely EL, He L, Groen EJ, Turnbull DM, Bushby KM, Taylor RW. The m.5650G>A mitochondrial tRNA(Ala) mutation is pathogenic and causes a phenotype of pure myopathy. *Neuromusc Disord* 2008;18:63–7.
- ^{hhh}Blakely EL, Yarham JW, Alston CL, Craig K, Poulton J, Brierley C, Park S-M, Dean A, Xuereb JH, Anderson KN, Compston A, Allen C, Sharif S, Enevoldson P, Wilson M, Hammans SR, Turnbull DM, McFarland R, Taylor RW. Pathogenic mitochondrial tRNA point mutations: nine novel mutations affirm their importance as a cause of mitochondrial disease. *Hum Mutat* 2013;34:1260–8.
- ⁱⁱⁱHao H, Moraes CT. A disease-associated G5703A mutation in human mitochondrial DNA causes a conformational change and a marked decrease in steady-state levels of mitochondrial tRNA(Asn). *Mol Cell Biol* 1997;17:6831–7, Moraes CT, Ciacci F, Bonilla E, Jansen C, Hirano M, Rao N, Lovelace RE, Rowland LP, Schon EA, DiMauro S. Two novel pathogenic mitochondrial DNA mutations affecting organelle number and protein synthesis. Is the tRNA^{Leu}(UUR) gene an etiologic hot spot? *J Clin Invest* 1993;92:2906–15, Vives-Bauza C, Del Toro M, Solano A, Montoya J, Andreu AL, Roig M. Genotype-phenotype correlation in the 5703G>A mutation in the tRNA(ASN) gene of mitochondrial DNA. *J Inher Metab Dis* 2003;26:507–8.
- ⁱⁱⁱJaksch M, Hofmann S, Kleinle S, Liechti-Gallati S, Pongratz DE, Muller-Hocker J, Jedeke KB, Meitinger T, Gerbitz KD. A systematic mutation screen of 10 nuclear and 25 mitochondrial candidate genes in 21 patients with cytochrome c oxidase (COX) deficiency shows tRNA(Ser)(UCN) mutations in a subgroup with syndromal encephalopathy. *J Med Genet* 1998a;35:895–900, Jaksch M, Klopstock T, Kurlermann G, Dorner M, Hofmann S, Kleinle S, Hegemann S, Weissert M, Muller-Hocker J, Pongratz D, Gerbitz KD. Progressive myoclonus epilepsy and mitochondrial myopathy associated with mutations in the tRNA(Ser(UCN)) gene. *Ann Neurol* 1998b;44:635–40, Schuelke M, Bakker M, Stoltenburg G, Sperner J, von Moers A. Epilepsia partialis continua associated with a homoplasmic mitochondrial tRNA(Ser(UCN)) mutation. *Ann Neurol* 1998;44:700–704, Tiranti V, Chariot P, Carella F, Toscano A, Soliveri P, Giralda P, Carrara F, Fratta GM, Reid FM, Mariotti C, Zeviani M. Maternally inherited hearing loss, ataxia and myoclonus associated with a novel point mutation in mitochondrial tRNA^{Ser(UCN)} gene. *Hum Mol Genet* 1995;4:1421–7.
- ^{kkk}Grafakou O, Hol FA, Otfried Schwab K, Siers MH, ter Laak H, Trijbels F, Ensenauer R, Boelen C, Smeitink J. Exercise intolerance, muscle pain and lactic acidemia associated with a 7497G>A mutation in the tRNA^{Ser(UCN)} gene. *J Inher Metab Dis* 2003;26:593–600, Jaksch M, Klopstock T, Kurlermann G, Dorner M, Hofmann S, Kleinle S, Hegemann S, Weissert M, Muller-Hocker J, Pongratz D, Gerbitz KD. Progressive myoclonus epilepsy and mitochondrial myopathy associated with mutations in the tRNA(Ser(UCN)) gene. *Ann Neurol* 1998b;44:635–40, Mollers M, Maniura-Weber K, Kiseljakovic E, Bust M, Hayrapetyan A, Jaksch M, Helm M, Wiesner RJ, von Kleist-Retzow JC. A new mechanism for mtDNA pathogenesis: impairment of post-transcriptional maturation leads to severe depletion of mitochondrial tRNA^{Ser(UCN)} caused by T7512C and G7497A point mutations. *Nucleic Acids Res* 2005;33:5647–58.
- ^{lll}Hutchin TP, Parker MJ, Young ID, Davis AC, Pulleyn LJ, Deeble J, Lench NJ, Markham AF, Mueller RF. A novel mutation in the mitochondrial tRNA(Ser(UCN)) gene in a family with non-syndromic sensorineural hearing impairment. *J Med Genet* 2000;37:692–4.
- ^{mmm}Sue CM, Tanji K, Hadjigeorgiou G, Andreu AL, Nishino I, Krishna S, Bruno C, Hirano M, Shanske S, Bonilla E, Fischel-Ghodsian N, DiMauro S, Friedman R. Maternally inherited hearing loss in a large kindred with a novel T7511C mutation in the mitochondrial DNA tRNA(Ser(UCN)) gene. *Neurology* 1999;52:1905–8.
- ⁿⁿⁿShoffner JM, Lott MT, Lezza AM, Seibel P, Ballinger SW, Wallace DC. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA^{Lys} mutation. *Cell* 1990;61:931–7, Wallace DC, Zheng X, Lott MT, Shoffner JM, Hodge JA, Kelley RI, Epstein CM, Hopkins LC, 1988b. Familial mitochondrial encephalomyopathy (MERRF): Genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease. *Cell* 1990;55:601–10.
- ^{ooo}Masucci JP, Davidson M, Koga Y, Schon EA, King MP. In vitro analysis of mutations causing myoclonus epilepsy with ragged-red fibers in the mitochondrial tRNA^{Lys} gene: two genotypes produce similar phenotypes. *Mol Cell Biol* 1995;15:2872–81, Silvestri G, Moraes CT, Shanske S, Oh SJ, DiMauro S. A new mtDNA mutation in the tRNA^{Lys} gene associated with myoclonic epilepsy and ragged red fibers (MERRF). *Am J Hum Genet* 1992;51:1213–7, Zeviani M, Muntoni F, Savarese N, Serra G, Tiranti V, Carrara F, Mariotti C, DiDonato S. A MERRF/MELAS overlap syndrome associated with a new point mutation in the mitochondrial DNA tRNA^{Lys} gene. *Eur J Hum Genet* 1993;1:80–7.
- ^{ppp}Ozawa M, Nishino I, Horai S, Nonaka I, Goto YI. Myoclonus epilepsy associated with ragged-red fibers: a G-to-A mutation at nucleotide pair 8363 in mitochondrial tRNA(Lys) in two families. *Muscle Nerve* 1997;20:271–8, Santorelli FM, Mak SC, El-Schahawi M, Casali C, Shanske S, Baram TZ, Madrid RE, DiMauro S. Maternally inherited cardiomyopathy and hearing loss associated with a novel mutation in the mitochondrial tRNA^{Lys} gene (G8363A). *Am J Hum Genet* 1996;58:933–9.
- ^{qqq}Bidooki SK, Johnson MA, Chrzanowska-Lightowlers Z, Bindoff LA, Lightowlers RN. Intracellular mitochondrial triplasm in a patient with two heteroplasmic base changes. *Am J Hum Genet* 1997;60:1430–8.
- ^{rrr}Melone MA, Tessa A, Petrini S, Lus G, Sampaolo S, di Fede G, Santorelli FM, Cotrufo R. Revelation of a new mitochondrial DNA mutation(G12147A) in a MELAS/MERFF phenotype. *Arch Neurol* 2004;61:269–72, Taylor RW, Schaefer AM, McDonnell MT, Petty RK, Thomas AM, Blakely EL, Hayes CM, McFarland R, Turnbull DM. Catastrophic presentation of mitochondrial disease due to a mutation in the tRNA(His) gene. *Neurology* 2004;62:1420–3.
- ^{sss}Cardaioli E, Da Pozzo P, Gallus GN, Malandrini A, Gambelli S, Gaudiano C, Malfatti E, Viscomi C, Zicari E, Berti G, Serni G, Dotti MT, Federico A. A novel heteroplasmic tRNA(Ser(UCN)) mtDNA point mutation associated with progressive external ophthalmoplegia and hearing loss. *Neuromusc Disord* 2007;17:681–3.

- ^{ttt}Karadimas CL, Salviati L, Sacconi S, Chronopoulou P, Shanske S, Bonilla E, De Vivo DC, DiMauro S. Mitochondrial myopathy and ophthalmoplegia in a sporadic patient with the G12315A mutation in mitochondrial DNA. *Neuromusc Disord* 2002;12:865–8.
- ^{uuu}Cardaioli E, Da Pozzo P, Malfatti E, Gallus GN, Rubegni A, Malandrini A, Gaudiano C, Guidi L, Serni G, Berti G, Dotti MT, Federico A. Chronic progressive external ophthalmoplegia: a new heteroplasmic tRNA(Leu(CUN)) mutation of mitochondrial DNA. *J Neurol Sci* 2008;272:106–9.
- ^{vvv}Horvath R, Kemp JP, Tuppen HA, Hudson G, Oldfors A, Marie SK, Moslemi AR, Servidei S, Holme E, Shanske S, Kollberg G, Jayakar P, Pyle A, Marks HM, Holinski-Feder E, Scavina M, Walter MC, Coku J, Gunther-Scholz A, Smith PM, McFarland R, Chrzanowska-Lightowlers ZM, Lightowlers RN, Hirano M, Lochmuller H, Taylor RW, Chinnery PF, Tulinius M, DiMauro S. Molecular basis of infantile reversible cytochrome c oxidase deficiency myopathy. *Brain* 2009;132:3165–74. Mimaki M, Hatakeyama H, Komaki H, Yokoyama M, Arai H, Kirino Y, Suzuki T, Nishino I, Nonaka I, Goto Y. Reversible infantile respiratory chain deficiency: a clinical and molecular study. *Ann Neurol* 2010;68:845–54.
- ^{www}Hanna MG, Nelson I, Sweeney MG, Cooper JM, Watkins PJ, Morgan-Hughes JA, Harding AE. Congenital encephalomyopathy and adult-onset myopathy and diabetes mellitus: different phenotypic associations of a new heteroplasmic mtDNA tRNA glutamic acid mutation. *Am J Hum Genet* 1995;56:1026–33. Hao H, Bonilla E, Manfredi G, DiMauro S, Moraes CT. Segregation patterns of a novel mutation in the mitochondrial tRNA glutamic acid gene associated with myopathy and diabetes mellitus. *Am J Hum Genet* 1995;56:1017–25.

present with retinitis pigmentosa at 75% mutant, olivopontocerebellar atrophy at about 85% mutant, and Leigh syndrome at 95%–100% mutant [115–117].

Classic examples of mtDNA protein synthesis mutations are the tRNA^{Lys} nt 8344A>G mutation associated with MERRF syndrome [6,8] and the tRNA^{Leu(UUR)} nt 3243A>G mutation associated with MELAS [118]. Many tRNA mutations changes in the heteroplasmy levels have profound effects on the phenotypic presentation.

For the tRNA^{Lys} 8344A>G mutation, individuals with low heteroplasmy levels may manifest with sensory neural hearing loss and mitochondrial myopathy. However, at high heteroplasmy levels, individuals can present with debilitating myoclonus, cardiomyopathy, and dementia [6,8]. The phenotypic variability of the tRNA^{Leu(UUR)} 3243A>G mutation is even more striking. At about 10%–30% heteroplasmy, the 3243G mutation allele can present as type 1 and type 2 diabetes or autism; at about 50%–80% heteroplasmy, with migraines and neuromuscular diseases including MELAS; and at 90%–100% heteroplasmy, with perinatal lethal disease. These marked differences in cellular and clinical phenotype correlate with changes in mitochondrial and cellular physiology and structure. But the abrupt changes in phenotypic manifestations is the result of mitochondrial signaling to the nucleus, presumably through mitochondrial high-energy intermediates, which results in abrupt changes in nuclear gene transcription profiles. These phase shifts in nuclear gene expression correlate the changes in the clinical manifestations, presumably

mediated by alterations in cellular signal transduction and epigenomic signaling [119,120].

Changes in functional elements in the mtDNA control region can also result in maternally transmitted clinical phenotypes. One well-characterized phenotype is cyclic vomiting and migraine headaches [121].

10.5.2 Ancient Adaptive mtDNA Variants

Ancient adaptive mtDNA variants can modulate the penetrance of maternally inherited mutations, such as the milder LHON *ND4* 11778G>A (R340H), *ND6* 14484T>C (M64V), and *ND4L* 10663T>C (V65A) [122] mutations. In Europeans, mutations that arise on mtDNA haplogroup J have a significantly increased penetrance [122–125]. This increased penetrance is associated with lower respiration in cybrids in which the 11778G>A (R340H) mutation is on the J background versus on the haplogroup H background [126–128]. mtDNA haplogroups have been associated with increased risk for a wide range of common clinical manifestations [129]. Because these associations increase risk rather than being sufficient in themselves to cause the disease, they will be discussed under the common “complex” disease section.

10.5.3 Developmental and Somatic mtDNA Mutations

Finally, developmental and somatic tissue mutations can result in “spontaneous” diseases. De novo mtDNA deletions can arise early in development and result in varying severity phenotypes depending on the distribution

and heteroplasmy levels of the deletion. Interestingly, as long as the deletion removes a tRNA gene, the actual size and position of the deletion are less important to the phenotype. Deletions that are at lower heteroplasmy levels present with CPEO [5] but at higher deletion levels with the earlier-onset Kearns-Sayre syndrome (KSS) [130]. In both CPEO and KSS, the deletion is not found in blood presumably because the deletion segregates during mitotic replication of the bone marrow stem cells. Those cells that lose the deletion have a replicative advantage and replace the stem cells that retain the deleted mtDNA [131]. Deletions that are widely distributed throughout the body and at high heteroplasmy levels result in the Pearson marrow-pancreas syndrome. In this syndrome the blood cells have high levels of deletion associated with early childhood transfusion-dependent pancytopenia [132,133]. Presumably in Pearson syndrome the mtDNA deletion level is sufficiently high that few bone marrow stem cells segregate to normal mtDNAs. Occasionally, however, Pearson patients do revert back to more normal bone marrow, relieving the pancytopenia, but these patients progress to KSS.

Most cases of single deletion are spontaneous, but occasionally deletions can be maternally inherited. These maternally inherited rearrangement mutations appear to be mtDNA duplications, which are less lethal than deletions. Within postmitotic tissues, the duplications undergo rearrangement to generate deletions, which then accumulate in the postmitotic tissues to produce the phenotype [134].

Random mtDNA mutations also accumulated in adult stem and postmitotic cells with age. These somatic mtDNA mutations can segregate to higher heteroplasmy progressively eroding mitochondrial function until the bioenergetic capacity of the cell or tissue falls below the expression threshold. Such somatic cell mutations can exacerbate inherited partial bioenergetics defects resulting in the delayed onset and progressive course of diseases and for normal individuals may constitute the aging clock [9,135–137].

10.5.4 The Range of mtDNA Disease Phenotypes in MITOMAP and MITOMASTER

A list of the current mtDNA diseases and representative mutations is presented in Table 10.1. Since the late 1980s, diseases associated with mtDNA mutations have been descriptive of the phenotype, so the range of clinical presentations can be deduced from this table. A

detailed description of the mtDNA variants associated with these diseases is available in our previous chapter in this series [15], and a complete accounting of human mtDNA variation is available through our information service: MITOMAP and MITOMASTER [16]. MITOMAP currently encompasses 46,092 full-length mtDNA sequences, representing all global populations, with 13,662 nucleotide variants, including those that typify the mtDNA haplogroups. The clinical databases encompass 686 mtDNA mutations, 353 in protein coding regions, and 333 in tRNA and rRNA genes. MITOMASTER provides a comprehensive set of analytical tools for the analysis of mtDNA variation including the capacity to import mtDNA sequences to be interrogated, information on mtDNA variants, and tools to interpret the potential pathophysiological significance of mtDNA variants. In the year ending December 31, 2017, the MITOMAP portal was visited 118,467 times. The distribution of access is provided in Fig. 10.4.

10.6 nDNA CODED MITOCHONDRIAL DISEASES

Bioenergetic disease can result from mutations in any one of the hundreds of nDNA genes that code for mitochondrial proteins. Mutations in several hundred nDNA gene loci have already been reported to cause mitochondrial bioenergetic dysfunction [15,87]. Pathogenic mutations have been identified in OXPHOS structural and assembly factors, intermediate metabolism, replication, translation, and regulator genes. When both copies of a chromosomal gene are mutated, severe OXPHOS defects can occur and cause devastating pediatric diseases, the most commonly recognized phenotype being Leigh syndrome [15,87]. However, phenotypes can be highly variable. For example mutations in the complex I *NDUFS2* originally associated with Leigh syndrome can also present with LHON-like hereditary optic neuropathy [138], and mutations in 10 mitochondrial carrier proteins have been associated with a variety of clinical phenotypes [24–26].

Multiple clinical manifestations can result from mutations in nDNA genes. OXPHOS defects can result from mutations in the mitochondrial translation proteins including aminoacyl tRNA synthetases and related mitochondrial translation polypeptides [139,140]. Moreover, protein synthesis defects can result from faulty interactions between mtDNA and nDNA variants.



Figure 10.4 Global access of MITOMAP during 2017.

For example, the mtDNA 12 rRNA 1555A>G variant alone predisposes to aminoglycoside-induced sensorineural hearing loss, but in conjunction with the A10S variant in the TRMU (methyaminomethyl-2-thiouridylate-methyltransferase) modification gene, it can result in inherited deafness [141].

Mutations in the nDNA-coded mtDNA biogenesis genes can cause degenerative diseases by destabilizing mtDNA biogenesis, resulting in multiple deletions and/or mtDNA depletion. Mutations in the mtDNA polymerase γ (*POLG*) are the most common nDNA mutations that cause mtDNA multiple deletions or depletions and can present with a wide variety of phenotypes ranging from CPEO to Alper syndrome [142–144]. Other important mtDNA stability mutations include the Twinkle helicase [145], mitochondrial deoxyguanosine kinase and thymidine kinase 2 [146,147], cytosolic thymidine phosphorylase [148], and the heart-muscle adenine nucleotide (ADP/ATP) translocator (*ANT1*) [43–45], to name a few [15,149].

A representative list of the nDNA mitochondrial genes that have been found to be mutant in OXPHOS-deficient patients is provided in Tables 10.2 and 10.3. Table 10.2 lists the 33 structural OXPHOS genes found to be mutant in patients. This includes 20 complex I genes, all four complex II genes, two complex III genes, four

complex IV genes, and three complex V genes. Table 10.3 lists the plethora of additional nDNA mitochondrial genes found to be mutant in patients including 29 OXPHOS complex assembly genes, 37 mitochondrial protein synthesis genes, 16 mitochondrial maintenance genes, 11 iron homeostasis genes, 10 CoQ synthesis genes, and miscellaneous additional genes encoding for proteins involved in other mitochondrial functions. These lists are available online through MITOMAP [16]. The number of clinically relevant nDNA mitochondrial genes is increasing rapidly.

10.7 MITOCHONDRIAL ETIOLOGY OF COMPLEX DISEASES

The extraordinary complexity of mtDNA and nDNA mitochondrial genetics and of mitochondrial–nuclear interactions and the commonality between primary mitochondrial disease phenotypes and the phenotypes of the common diseases strongly implicate mitochondrial dysfunction in the etiology of the common “complex” diseases. Mitochondrial dysfunction selectively impairs the organs with the highest bioenergetic demand—the brain, heart, muscle, kidney, endocrine tissues, and liver—and these are the same organs that are affected by the common diseases. Mitochondrial

TABLE 10.2 Structural Nuclear Genes

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Complex I	<i>NDUFS1</i>	157655	IP fraction	2q33-q34	AR	LS	a
	<i>NDUFS2</i>	602985	IP fraction	1q23	AR	Encephalopathy, cardiomyopathy	b
	<i>NDUFS3</i>	603846	IP fraction	11p11.11	AR	LS	c
	<i>NDUFS4</i>	602694	IP fraction	5q11.1	AR	LS	d
	<i>NDUFS6</i>	603848	IP fraction	5pter-p15.33	AR	Fatal infantile lactic acidosis	e
	<i>NDUFS7</i>	601825	HP fraction	19p13.3	AR	LS	f
	<i>NDUFS8</i>	602141	HP fraction	11q13	AR	LS	g
	<i>NDUFB3</i>	603839	HP fraction	2q31.3	AR	Fatal infantile lactic acidosis	h
	<i>NDUFB9</i>	601445	HP fraction	8q24.13	AR	Hypotonia, lactic acidosis	i
	<i>NDUFB10</i>	603843	HP fraction	16p13.3	AR	Lactic acidosis, cardiomyopathy	j
	<i>NDUFB11</i>	300403	HP fraction	Xp11.3	X	Intrauterine growth restriction, lactic acidosis	k
	<i>NDUFV1</i>	161015	FP fraction	11q13	AR	LS	l
	<i>NDUFV2</i>	600532	FP fraction	18p11	AR	Cardiomyopathy, hypotonia, encephalopathy	m
	<i>NDUFA1</i>	300078	HP fraction	Xq24	X	LSProgressive neurodegenerative disorder	n
	<i>NDUFA2</i>	602137	HP fraction	5q31.2	AR	LS	o
	<i>NDUFA9</i>	603834	HP fraction	12p13.32	AR	LS	p
	<i>NDUFA10</i>	603835	HP fraction	2q37.3	AR	LS	q
	<i>NDUFA11</i>	612638	IP fraction	19p13.3	AR	Fatal infantile lactic acidosis encephalocardiomyopathy	r
Complex II	<i>NDUFA12</i>	609653	HP fraction	12q22	AR	LS	s
	<i>NDUFA13</i>	609435	HP fraction	19p13.11	AR	Encephalopathy, optic atrophy	t
	<i>SDH-A</i>	600857	FP subunit	5p15	AR	LS	u
	<i>SDH-B</i>	185470	IP subunit	1p36.1-p35	AD	Phaeochromocytoma and paraganglioma	v
Complex III	<i>SDH-C</i>	602413	Membrane subunit	1q21	AD	Autosomal dominant paraganglioma type 3	w
	<i>SDH-D</i>	602690	Membrane subunit	11q23	AD	Autosomal dominant paraganglioma type 1, pheochromocytoma	x
	<i>UQCRL</i>	191330	Electron transfer	8q22	AR	Hypoglycemia, lactic acidosis	y
	<i>UQCRCQ</i>	612080	Electron transfer	5q31.1	AR	Severe neurological phenotype	z

TABLE 10.2 Structural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Complex IV	<i>COX6A1</i>	602072	Cytochrome oxidase activity	12q24.31	AR	Charcot-Marie-Tooth disease	aa
	<i>COX6B1</i>	124089	Cytochrome oxidase activity and assembly	19q13.1	AR	Encephalomyopathy	bb
	<i>COX7B</i>	300885	Cytochrome oxidase activity	Xq21.1	X	Microphthalmia with linear skin lesions	cc
	<i>COX8A</i>	123870	Cytochrome oxidase activity	11q13.1	AR	LS	dd
Complex V	<i>ATP5E</i>	606153	ATPase activity	20q13.3	AR	Lactic acidosis, mental retardation, peripheral neuropathy	ee
	<i>ATP5A1</i>	164360	ATPase activity	18q21.1	AR	Neonatal encephalopathy	ff
	<i>ATP8A2</i>	605870	ATPase activity	13q12.13	AR	Cerebellar ataxia, mental retardation	gg

See <https://mitomap.org/MITOMAP/NuclearGenesStructural> for additional reports and phenotypes. *AD*, autosomal dominant; *AR*, autosomal recessive; *FP*, flavoprotein; *HP*, hydrophobic; *IP*, iron-protein; *LS*, Leigh syndrome; *X*, X-linked.

^aBenit P, Chretien D, Kadhon N, de Lonlay-Debeney P, Cormier-Daire V, Cabral A, Peudenier S, Rustin P, Munnich A, Rotig A. Large-scale deletion and point mutations of the nuclear *NDUFV1* and *NDUFS1* genes in mitochondrial complex I deficiency. *Am J Hum Genet* 2001;68:1344–52.

^bLoeffen J, Elpeleg O, Smeitink J, Smeets R, Stockler-Ipsiroglu S, Mandel H, Sengers R, Trijbels F, van den Heuvel L. Mutations in the complex I *NDUFS2* gene of patients with cardiomyopathy and encephalomyopathy. *Ann Neurol* 2001;49:195–201.

^cBenit P, Slama A, Cartault F, Giurgea I, Chretien D, Lebon S, Marsac C, Munnich A, Rotig A, Rustin P. Mutant *NDUFS3* subunit of mitochondrial complex I causes Leigh syndrome. *J Med Genet* 2004;41:14–17.

^dvan den Heuvel L, Ruitenbeek W, Smeets R, Gelman-Kohan Z, Elpeleg O, Loeffen J, Trijbels F, Mariman E, de Bruijn D, Smeitink J. Demonstration of a new pathogenic mutation in human complex I deficiency: a 5-bp duplication in the nuclear gene encoding the 18-kD (AQDQ) subunit. *Am J Hum Genet* 1998;62:262–8.

^eKirby DM, Salemi R, Sugiana C, Ohtake A, Parry L, Bell KM, Kirk EP, Boneh A, Taylor RW, Dahl HH, Ryan MT, Thorburn DR. *NDUFS6* mutations are a novel cause of lethal neonatal mitochondrial complex I deficiency. *J Clin Invest* 2004;114:837–45. Spiegel R, Shaag A, Mandel H, Reich D, Penyakov M, Hujeirat Y, Saada A, Elpeleg O, Shalev SA. Mutated *NDUFS6* is the cause of fatal neonatal lactic acidemia in Caucasus Jews. *Eur J Hum Genet* 2009;17:1200–3.

^fSmeitink J, van den Heuvel L. Human mitochondrial complex I in health and disease. *Am J Hum Genet* 1999;64:1505–10.

^gLoeffen J, Smeitink J, Triepels R, Smeets R, Schuelke M, Sengers R, Trijbels F, Hamel B, Mullaart R, van den Heuvel L. The first nuclear-encoded complex I mutation in a patient with Leigh Syndrome. *Am J Hum Genet* 1998;63:1598–1608. Procaccio V, Wallace DC. Late-onset Leigh syndrome in a patient with mitochondrial complex I *NDUFS8* mutations. *Neurology* 2004;62:1899–1901.

^hCalvo SE, Compton AG, Hershman SG, Lim SC, Lieber DS, Tucker EJ, Laskowski A, Garone C, Liu S, Jaffe DB, Christodoulou J, Fletcher JM, Bruno DL, Goldblatt J, Dimauro S, Thorburn DR, Mootha VK. Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci Transl Med* 2012;4:118ra110.

ⁱHaack TB, Madignier F, Herzer M, Lamantea E, Danhauser K, Invernizzi F, Koch J, Freitag M, Drost R, Hillierl, Haberberger B, Mayr JA, Ahting U, Tiranti V, Rotig A, Iuso A, Horvath R, Tesarova M, Baric I, Uziel G, Rolinski B, Sperl W, Meitinger T, Zeviani M, Freisinger P, Prokisch H. Mutation screening of 75 candidate genes in 152 complex I deficiency cases identifies pathogenic variants in 16 genes including *NDUFB9*. *J Med Genet* 2012;49:83–9.

^jFriederich MW, Erdogan AJ, Coughlin CR, 2nd Elos, MT Jiang, H O'Rourke, CP Lovell, MA Wartchow E, Gowan K, Chatfield KC, Chick WS, Spector EB, Van Hove JLK, Riemer J. Mutations in the accessory subunit *NDUFB10* result in isolated complex I deficiency and illustrate the critical role of intermembrane space import for complex I holoenzyme assembly. *Hum Mol Genet* 2017;26:702–16.

- ^kKohda M, Tokuzawa Y, Kishita Y, Nyuzuki H, Moriyama Y, Mizuno Y, Hirata T, Yatsuka Y, Yamashita-Sugahara Y, Nakachi Y, Kato H, Okuda A, Tamaru S, Borna NN, Banshoya K, Aigaki T, Sato-Miyata Y, Ohnuma K, Suzuki T, Nagao A, Maehata H, Matsuda F, Higasa K, Nagasaki M, Yasuda J, Yamamoto M, Fushimi T, Shimura M, Kaiho-Ichimoto K, Harashima H, Yamazaki T, Mori M, Murayama K, Ohtake A, Okazaki Y. A comprehensive genomic analysis reveals the Genetic Landscape of Mitochondrial Respiratory Chain Complex Deficiencies. *mitochondrial respiratory chain complex deficiencies*. *PLoS Genet* 2016;12:e1005679.
- ^lSmeitink J, van den Heuvel L. Human mitochondrial complex I in health and disease. *Am J Hum Genet* 1999;64:1505–10.
- ^mBenit P, Beugnot R, Chretien D, Giurgea I, De Lonlay-Debeney P, Issartel JP, Corral-Debrinski M, Kerscher S, Rustin P, Rotig A, Munnich A. Mutant NDUFV2 subunit of mitochondrial complex I causes early onset hypertrophic cardiomyopathy and encephalopathy. *Hum Mutat* 2003;21:582–6.
- ⁿFernandez-Moreira D, Ugalde C, Smeets R, Rodenburg RJ, Lopez-Laso E, Ruiz-Falco ML, Briones P, Martin MA, Smeitink JA, Arenas J. X-linked NDUFA1 gene mutations associated with mitochondrial encephalomyopathy. *Ann Neurol* 2007;61:73–83, Potluri P, Davila A, Ruiz-Pesini E, Mishmar D, O'Hearn S, Hancock S, Simon MC, Scheffler I, Wallace DC, Procaccio V. A novel NDUFA1 mutation leads to a progressive mitochondrial complex I-specific neurodegenerative disease. *Mol Genet Metab* 2009;96:189–95.
- ^oHoefs SJ, Dieteren CE, Distelmaier F, Janssen RJ, Epplen A, Swarts HG, Forkink M, Rodenburg RJ, Nijtmans LG, Willems PH, Smeitink JA, van den Heuvel LP. NDUFA2 complex I mutation leads to Leigh disease. *Am J Hum Genet* 2008;82:1306–15.
- ^pvan den Bosch BJ, Gerards M, Sluiter W, Stegmann AP, Jongen EL, Hellebrekers DM, Oegema R, Lambrichts EH, Prokisch H, Danhauser K, Schoonderwoerd K, de Coe IF, Smeets HJ. Defective NDUFA9 as a novel cause of neonatally fatal complex I disease. *J Med Genet* 2012;49:10–15.
- ^qHoefs SJ, van Spronsen FJ, Lenssen EW, Nijtmans LG, Rodenburg RJ, Smeitink JA, van den Heuvel LP. NDUFA10 mutations cause complex I deficiency in a patient with Leigh disease. *Eur J Hum Genet* 2011;19:270–4.
- ^rBerger I, Hershkovitz E, Shaag A, Edvardson S, Saada A, Elpeleg O. Mitochondrial complex I deficiency caused by a deleterious NDUFA11 mutation. *Ann Neurol* 2008;63:405–8.
- ^sOstergaard E, Rodenburg RJ, van den Brand M, Thomsen LL, Duno M, Batbayli M, Wibrand F, Nijtmans L. Respiratory chain complex I deficiency due to NDUFA12 mutations as a new cause of Leigh syndrome. *J Med Genet* 2011;48:737–40.
- ^tAngebault C, Charif M, Guegen N, Piro-Megy C, Mousson de Camaret B, Procaccio V, Guichet PO, Hebrard M, Manes G, Leboucq N, Rivier F, Hamel CP, Lenaers G, Roubertie A. Mutation in NDUFA13/GRIM19 leads to early onset hypotonia, dyskinesia and sensorial deficiencies, and mitochondrial complex I instability. *Hum Mol Genet* 2015;24:3948–55.
- ^uBourgeron T, Rustin P, Chretien D, Birch-Machin M, Bourgeois M, Viegas-Pequignot E, Munnich A, Rotig A. Mutation of a nuclear succinate dehydrogenase gene results in mitochondrial respiratory chain deficiency. *Nat Genet* 1995;11:144–9.
- ^vAstuti D, Latif F, Dallol A, Dahia PL, Douglas F, George E, Skoldberg F, Husebye ES, Eng C, Maher ER. Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. *Am J Hum Genet* 2001;69:49–54.
- ^wNiemann S, Muller U. Mutations in SDHC cause autosomal dominant paraganglioma, type 3. *Nat Genet* 2000;26:268–70.
- ^xBaysal BE, Ferrell RE, Willett-Brozick JE, Lawrence EC, Myssiorek D, Bosch A, van der Mey A, Taschner PE, Rubinstein WS, Myers EN, Richard CW, Cornelisse CJ, Devilee P, Devlin B. Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. *Science* 2000;287:848–51.
- ^yHaut S, Brivet M, Touati G, Rustin P, Lebon S, Garcia-Cazorla A, Saudubray JM, Boutron A, Legrand A, Slama A. A deletion in the human QP-C gene causes a complex III deficiency resulting in hypoglycaemia and lactic acidosis. *Hum Genet* 2003;113:118–122.
- ^zBarel O, Shorer Z, Flusser H, Ofir R, Narkis G, Finer G, Shalev H, Nasasra A, Saada A, Birk OS. Mitochondrial complex III deficiency associated with a homozygous mutation in UQCRCQ. *Am J Hum Genet* 2008;82:1211–6.
- ^{aa}Tamiya G, Makino S, Hayashi M, Abe A, Numakura C, Ueki M, Tanaka A, Ito C, Toshimori K, Ogawa N, Terashima T, Maegawa H, Yanagisawa D, Tooyama I, Tada M, Onodera O, Hayasaka K. A mutation of COX6A1 causes a recessive axonal or mixed form of Charcot-Marie-Tooth disease. *Am J Hum Genet* 2014;95:294–300.
- ^{bb}Massa V, Fernandez-Vizcarra E, Alshahwan S, Bakhsh E, Goffrini P, Ferrero I, Mereghetti P, D'Adamo P, Gasparini P, Zeviani M. Severe infantile encephalomyopathy caused by a mutation in COX6B1, a nucleus-encoded subunit of cytochrome c oxidase. *Am J Hum Genet* 2008;82:1281–9.
- ^{cc}Indrieri A, van Rahden VA, Tiranti V, Morleo M, Iaconis D, Tammara R, D'Amato I, Conte I, Maystadt I, Demuth S, Zvulunov A, Kutsche K, Zeviani M, Franco B. Mutations in COX7B cause microphthalmia with linear skin lesions, an unconventional mitochondrial disease. *Am J Hum Genet* 2012;91:942–9.
- ^{dd}Hallmann K, Kudin AP, Zsurka G, Kornblum C, Reimann J, Stuve B, Waltz S, Hattingen E, Thiele H, Nurnberg P, Rub C, Voos W, Kopatz J, Neumann H, Kunz WS. Loss of the smallest subunit of cytochrome c oxidase, COX8A, causes Leigh-like syndrome and epilepsy. *Brain* 2016;139:338–45.
- ^{ee}Mayr JA, Havlickova V, Zimmermann F, Magler I, Kaplanova V, Jesina P, Pecinova A, Nuskova H, Koch J, Sperl W, Houstek J. Mitochondrial ATP synthase deficiency due to a mutation in the ATP5E gene for the F1 epsilon subunit. *Hum Mol Genet* 2010;19:3430–9.
- ^{ff}Jonckheere AI, Renkema GH, Bras M, van den Heuvel LP, Hoischen A, Gilissen C, Nabuurs SB, Huynen MA, de Vries MC, Smeitink JA, Rodenburg RJ. A complex V ATP5A1 defect causes fatal neonatal mitochondrial encephalopathy. *Brain* 2013;136:1544–54.
- ^{gg}Onat OE, Gulsuner S, Bilguvar K, Nazli Basak A, Topaloglu H, Tan M, Tan U, Gunel M, Ozcelik T. Missense mutation in the ATPase, aminophospholipid transporter protein ATP8A2 is associated with cerebellar atrophy and quadrupedal locomotion. *Eur J Hum Genet* 2013;21:281–5.

TABLE 10.3 Nonstructural Nuclear Genes

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Assembly							
Complex I	<i>NDUFAF1</i> (CIA30)	606934	Assembly	15q13.3	AR	Cardioencephalomyopathy	a
	<i>NDUFAF2</i> (B17.2L)	609653	Assembly	5q12.1	AR	Early onset progressive encephalopathy	b
	<i>NDUFAF3</i>	612911	Assembly	3p21.31	AR	Neonatal encephalopathy	c
	<i>NDUFAF4</i> (HRPAP2)	611776	Assembly	6q16.1	AR	Infantile encephalopathy	d
	<i>NDUFAF5</i> (C20orf7)	612360	Assembly	20p12.1	AR	LS	e
	<i>NDUFAF6</i>	612392	Assembly	8q22.1	AR	LS	f
	<i>NUBPL</i>	613621	Assembly	14q12	AR	Encephalomyopathy	g
	<i>FOXRED1</i>	613622	Assembly	11q24.2	AR	LS	h
Complex II	<i>ACAD9</i>	611103	Assembly and activity	3q26	AR	Hypertrophic cardiomyopathy encephalopathy	i
	<i>SDHAF1</i>	612848	Assembly	19q12-q13.2	AR	Leukoencephalopathy	j
	<i>SDHAF2</i>	613019	Assembly	11q12.2	AD	Autosomal dominant paraganglioma type 2	k
Complex III	<i>BCS1L</i>	603647	Assembly	2q33	AR	Encephalopathy, hepatic failure and tubulopathy, LS, GRACILE syndrome, Bjornstad syndrome	l
	<i>UQCC2</i>	614461	Assembly	6p21.31	AR	Lactic acidosis and renal tubular dysfunction	m
	<i>UQCC3</i>	616097	Assembly	11q12.3	AR	Lactic acidosis, hypoglycemia, hypotonia	n
Complex IV	<i>SURF1</i>	185620	Assembly	9q34	AR	LS	o
	<i>SCO1</i>	603644	Copper transport	17p13-p12	AR	Neonatal hepatic failure and encephalopathy	p
	<i>SCO2</i>	604272	Copper transport	22q13	AR	Neonatal cardioencephalomyopathy	q
	<i>COX10</i>	602125	Heme A farnesyl-transferase	17p12-p11.2	AR	Neonatal tubulopathy and encephalopathy, LS, cardiomyopathy	r
	<i>COX14</i> (C12orf62)	614478	COX assembly	12q13.12	AR	Neonatal lactic acidosis	s

Continued

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Complex V	<i>COX15</i>	603646	Heme A synthesis	10q24	AR	Early-onset hypertrophic cardiomyopathy, LS	t
	<i>COX20</i>	614698	Assembly	1q44	AR	Ataxia, muscle hypotonia	u
	<i>COA3</i>	614775	Assembly	17q21.2	AR	Neuropathy, exercise intolerance	v
	<i>COA5</i>	613920	Assembly	2q11.2	AR	Cardioencephalomyopathy	w
	<i>COA6</i>	614772	Assembly	1q42.2	AR	Cardioencephalomyopathy	x
	<i>LRPPRC</i>	220111	Assembly	2p21-p16	AR	French-Canadian LS	y
	<i>FASTKD2</i>	612322	Role in apoptosis	2q33.3	AR	Encephalomyopathy	z
	<i>TACO1</i>	612958	Translational activator of COX1	17q22-q24.2	AR	LS	aa
	<i>ATPAF2</i>	608918	Assembly	17p11.2	AR	Early-onset encephalopathy, lactic acidosis	bb
	<i>TMEM70</i>	604273	Assembly	8q21.11	AR	Neonatal encephalopathy, cardiomyopathy	cc
MtDNA maintenance	<i>POLG</i> (PEOA1)	174763	Polymerase gamma mtDNA replication	15q25	AD-AR	Alpers syndrome, AD-PEO and AR-PEO, male infertility, SANDO* syndrome, SCAE*	dd
	<i>POLG2</i> (PEOA4)	610131	Catalytic subunit of DNA polymerase gamma	17q23-q24	AD	AD-PEO	ee
	<i>ANT1</i> (PEOA2)	609283	Adenine nucleotide translocator isoform 1	4q35	AD-AR	AD-PEO, multiple mtDNA deletions	ff
	<i>MPV17</i>	137960	Regulation of mtDNA copy number	2p23-p21	AR	Hepatocerebral MDDS	gg
	<i>OPA1</i>	165500	Dynamin-related protein	3q28-q29	AD	AD-optic atrophy	hh
	<i>MFN2</i>	609260	Mitofusin Mitochondrial fusion	1p36-p35	AD	Multiple deletions Charcot-Marie-Tooth disease-2A2 (CMT2A2)	ii
						Multiple deletions	

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Mitochondrial import	<i>C10ORF2</i> (PEOA3)	609286	Twinkle helicase	10q24	AD	AD-PEO, SANDO syndrome	jj
	<i>TYMP</i> (ECGF1)	603041	Thymidine phosphorylase	22q13.32-qter	AR	MNGIE, mtDNA depletion	kk
	<i>DGUOK</i>	601465	Deoxyguanosine kinase Mitochondrial dNTP pool maintenance	2p13	AR	Hepatocerebral mtDNA depletion syndrome	ll
	<i>RRM2B</i> (PEOA5)	604712	Ribonucleotide reductase M2 B dNTP pool	8q23.1	AR	Encephalomyopathic renal tubulopathy MNGIE, AD-PEO	mm
	<i>SUCLA2</i>	603921	Succinate-CoA ligase, ADP-forming, beta Subunit	13q12.2-q13	AR	Encephalomyopathy with methylmalonic aciduria	nn
	<i>SUCLG1</i>	611224	Succinate-CoA ligase, alpha subunit	2p11.2	AR	Encephalomyopathy with methylmalonic aciduria	oo
	<i>TK2</i>	188250	Thymidine kinase Mitochondrial dNTP pool maintenance	16q22	AR	Myopathic mtDNA depletion	pp
	<i>TFAM</i>	600438	mitochondrial transcription factor A	10q21.1	AR	Encephalomyopathy mtDNA depletion	qq
	<i>FBXL4</i>	605654	mtDNA maintenance	6q16.1-q16.2	AR	Encephalomyopathy and myopathy mtDNA depletion	rr
	<i>MGME1</i>	615084	mtDNA maintenance	20p11.23	AR	CPEO and myopathy mtDNA depletion	ss
	<i>DDP</i>	304700	Protein import	Xq22	X-linked	Deafness-dystonia or Mohr-Tranebjaerg syndrome	tt
	<i>DNAJC19</i>	608977	Protein import	3q26.3	AR	Cardiomyopathy, ataxia	uu

Continued

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Mitochondrial protein synthesis	AARS2	612035	Alanyl-tRNA synthetase	6p21.1	AR	Cardiomyopathy; leukoencephalopathy	vv
	CARS2	612800	Cysteinyl-tRNA synthetase	13q34	AR	Myoclonic epilepsy	vww
	DARS2	611105	Aspartyl-tRNA synthetase	1q25.1	AR	Leukoencephalopathy and lactic acidosis	xx
	EARS2	612799	Glutamyl tRNA synthetase	16p12.2	AR	Leukoencephalopathy	yy
	FARS2	611592	Phenylalanyl-tRNA synthetase	6p25.1	AR	Alpers syndrome, spastic paraplegia	zz
	GARS	600287	Glycyl-tRNA synthetase	7p14.3	AD	Charcot-Marie-Tooth disease	aaa
	HARS2	600783	Histidyl-tRNA synthetase	5q31.3	AR	Perrault syndrome	bbb
	IARS2	612801	Isoleucyl tRNA-Synthetase	1q41	AR	Cataract, deafness, neuropathy/Leigh syndrome	ccc
	KARS	601421	Lysyl-tRNA synthetase	16q23.1	AR	CMT disease/Deafness	ddd
	LARS	615438	Leucine-tRNA synthetase	5q32	AR	Hepatopathy	eee
	LARS2	604544	Leucyl-tRNA synthetase	3p21.31	AR	Perrault syndrome	fff
	NARS2	612803	Asparaginyl-tRNA synthetase	11q14.1	AR	Alpers syndrome/nonsyndromic deafness and Leigh syndrome	ggg
	PARS2	612036	Prolyl-tRNA synthetase	1p32.3	AR	Alpers syndrome	hhh
	RARS2	611523	Arginyl-tRNA synthetase	6q16.1	AR	Pontocerebellar hypoplasia	iii
	SARS2	612804	Seryl-tRNA synthetase	19q13.2	AR	Hyperuricemia, pulmonary hypertension, renal failure	jjj
	TARS2	612805	Threonyl-tRNA synthetase	1q21.2	AR	Encephalomyopathy	kkk
	VAR2	612802	Valyl-tRNA synthetase	6p21.33	AR	Encephalomyopathy	lll

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
	<i>YARS2</i>	610957	Tyrosyl-tRNA synthetase	12p11.21	AR	Myopathy, lactic acidosis, and sideroblastic anemia-2	mmm
	<i>EFG1</i>	609060	Elongation factor G1 mitochondrial translation defect	3q25	AR	Severe hepatoen- cephalopathy and lactic acidosis	nnn
	<i>TSFM</i>	604723	Mitochondrial translation elongation	12q13-q14	AR	Encephalomyopathy, hypertrophic cardio- myopathy	ooo
	<i>TUFM</i>	602389	Mitochondrial translation Elongation	16p11.2	AR	Leukodystrophy with micropolygyria	ppp
	<i>GTPBP3</i>	608536	GTP-binding protein	19p13.11	AR	Cardiomyopathy, encephalopathy	qqq
	<i>MTFMT</i>	611766	Mitochondrial translation	15q22.31	AR	LS	rrr
	<i>MTO1</i>	614667	tRNA modifi- cation	6q13	AR	Cardiomyopathy	sss
	<i>TRMT5</i>	611023	mitochondrial tRNA meth- ylation	14q23.1	AR	Cardiomyopathy/ exercise intoler- ance	ttt
	<i>TRMT10C</i>	615423	TRNA methyl- transferase	3q12.3	AR	Hypotonia, feeding difficulties, deaf- ness	uuu
	<i>TRMU</i>	610230	mitochondrial translation	22q13.31	AR	Liver failure, deaf- ness	vvv
	<i>GFM1</i>	606639	Mitochondrial translation elongation	3q25.32	AR	Encephalopathy/ hepatic failure	www
	<i>GFM2</i>	606544	Mitochondrial translation elongation	5q13.3	AR	Neurodevelopmental disorder	xxx
	<i>C12orf65</i>	613541	Mitochondrial translation	12q24.31	AR	Dysmorphic features Encephalomyopa- thy, optic atrophy, axonal neuropathy, paraparesis	yyy
	<i>RMND1</i>	614917	Mitochondrial translation	6q25.1	AR	Encephalopathy	zzz
	<i>MRPL3</i>	607118	Mitochondrial translation	3q22.1	AR	Cardiomyopathy, mental retardation	aaaa

Continued

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Iron homeo- stasis	<i>MRPS7</i>	611974	Mitochondrial translation	17q25.1	AR	Deafness, hepatic and renal failure	bbbb
	<i>MRPL12</i>	602375	Mitochondrial translation	17q25.3	AR	Growth retardation, encephalopathy	cccc
	<i>MRPS16</i>	609204	Mitochondrial translation	10q22.1	AR	Neonatal lactic acidosis corpus callosum agenesis	dddd
	<i>MRPS22</i>	605810	Mitochondrial translation	3q23	AR	Cardiomyopathy, tubulopathy	eeee
	<i>MRPL44</i>	611849	Mitochondrial translation	2q36.1	AR	Cardiomyopathy	ffff
	<i>FRDA (FXN)</i>	606829	Frataxin Trinuc.* repeat,	9q13	AR	Friedreich ataxia, neuropathy, cardiomyopathy, diabetes	gggg
	<i>ABCB7</i>	301310	Iron transport	Xq13.1-q13.3	X-linked	X-linked sideroblastic anemia with ataxia	hhhh
	<i>GLRX5</i>	205950	Iron-sulfur cluster bio-synthesis	3p22.1	AR	Sideroblastic anemia	iiii
	<i>ISCU</i>	255125	Iron-sulfur cluster bio-synthesis	12q23.3	AR	Myopathy, lactic acidosis, exercise intolerance	jjjj
	<i>BOLA3</i>	613183	Iron-sulfur cluster bio-synthesis	2p13.1	AR	Encephalomyopathy, cardiomyopathy	kkkk
	<i>NFU1</i>	608100	Iron-sulfur cluster bio-synthesis	2p13.3	AR	Lactic acidosis multiple respiratory chain deficiency	llll
	<i>ISCA2</i>	615317	Iron-sulfur cluster bio-synthesis	14q24.3	AR	Leukodystrophy	mmmm
	<i>IBA57</i>	615316	Iron-sulfur cluster bio-synthesis	1q42.13	AR	Myopathy, encephalopathy	nnnn
	<i>LYRM4</i>	613311	Iron-sulfur cluster bio-synthesis	6p25.1	AR	Lactic acidosis, failure to thrive	oooo
	<i>LYRM7</i>	615831	Iron-sulfur cluster bio-synthesis	5q23.3-q31.1	AR	Encephalopathy, lactic acidosis	pppp
	<i>FDXL1</i>	614585	Iron-sulfur cluster bio-synthesis	19p13.2	AR	Myopathy, lactic acidosis	qqqq

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Coenzyme Q10 biogenesis	<i>COQ2</i>	609825	CoQ10 deficiency	4q21-q22	AR	Encephalomyopathy, nephropathy	rrrr
	<i>COQ4</i>	612898	CoQ10 deficiency	9q34.13	AR	Encephalomyopathy, mental retardation	ssss
	<i>COQ5</i>	616359	CoQ10 deficiency	12q24.31	AR	Encephalomyopathy, cerebellar ataxia	tttt
	<i>COQ6</i>	614647	CoQ10 deficiency	14q24.3	AR	Nephrotic syndrome, deafness	uuuu
	<i>COQ7</i>	601683	CoQ10 deficiency	16p12.3	AR	Hypotonia, cardiac hypertrophy	vvvv
	<i>COQ9</i>	612837	CoQ10 deficiency	16q13	AR	Neonatal lactic acidosis seizures, cardiomyopathy	wwww
	<i>APTX</i>	606350	CoQ10 deficiency	9p13.3	AR	Cerebellar ataxia	xxxx
	<i>PDSS1</i>	607429	CoQ10 deficiency	10p12.1	AR	Oculomotor apraxia	yyyy
						Deafness, valvulopathy, mental retardation	
	<i>PDSS2</i>	610564	CoQ10 deficiency	6q21	AR	LS, nephrotic syndrome	zzzz
Chaperone function	<i>CABC1</i>	606980	CoQ10 deficiency	1q42.2	AR	Cerebellar ataxia, lactic acidosis	aaaaa
	<i>SPG7</i>	607259	Paraplegin ATPase protease	16q24.3	AR	Spastic paraplegia	bbbbb
Mitochondrial integrity	<i>HSPD1</i>	118190	Mitochondrial chaperone	2q33.1	AR	Spastic paraplegia, leukodystrophy	ccccc
	<i>DLP1</i>	603850	Mitochondrial and peroxisomal fission	12p11.21	AD	Microcephaly, abnormal brain	ddddd
	<i>G4.5 (Tafazzin)</i>	302060	Cardiolipin defect	Xq28	X-linked	Development, optic atrophy, Lactic acidosis	eeeeee
	<i>RMRP</i>	250250	RNAse mitochondrial RNA processing	9p13-p12	AR	Barth syndrome, X-linked dilated cardiomyopathy	ffffff
						Metaphyseal chondrodysplasia or cartilage-hair hypoplasia	

Continued

TABLE 10.3 Nonstructural Nuclear Genes—cont'd

Complex	Name	OMIM	Function	Chromosome	Inheritance	Clinical Phenotype	References
Mitochondrial metabolism	<i>PDHA1</i>	308930	Pyruvate dehydrogenase	Xp22.2-p22.1	X-linked	LS	ggggg
	<i>ETHE1</i>	602473	E1- α subunit Ethylmalonic acid metabolism	19q13	AR	Encephalopathy, ethylmalonic aciduria	hhhhh
	<i>PUS1</i>	600462	Pseudouridine synthase	12q24.33	AR	Myopathy, lactic acidosis, and sideroblastic anemia	iiiiii
	<i>ATAD3</i>	617183	mitochondrial dynamics	1p36.33	AR/AD	Neurodevelopmental disorder, pontocerebellar hypoplasia, encephalopathy	jjjjj

See <https://mitomap.org/MITOMAP/NuclearGenesNonStructural> for additional reports and phenotypes. *AD*, Autosomal Dominant; *AR*, Autosomal Recessive; *CMT*, Charcot-Marie-Tooth; *CoQ10*, Coenzyme Q10; *CPEO*, Chronic progressive external ophthalmoplegia; *GRACILE syndrome*, Growth Retardation, Amino aciduria, Cholestasis, Iron overload, Lactic acidosis, and Early death; *LS*, Leigh Syndrome; *MDDS*, Mitochondrial DNA Depletion Syndrome; *MNGIE*, MyoNeuroGastroIntestinal Encephalopathy; *PEO*, Progressive external ophthalmoplegia; *SANDO*, Sensory Ataxic Neuropathy, Dysarthria, and Ophthalmoparesis; *SCAE*, Spinocerebellar Ataxia with Epilepsy.

^aDunning CJ, McKenzie M, Sugiana C, Lazarou M, Silke J, Connelly A, Fletcher JM, Kirby DM, Thorburn DR, Ryan MT. Human CIA30 is involved in the early assembly of mitochondrial complex I and mutations in its gene cause disease. *The EMBO J* 2000;26:3227–37.

^bOgilvie I, Kennaway NG, Shoubridge EA. A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy. *J Clin Invest* 2000;115:2784–92.

^cSaada A, Vogel RO, Hoefs SJ, van den Brand MA, Wessels HJ, Willems PH, Venselaar H, Shaag A, Barghuti F, Reish O, Shohat M, Huynen MA, Smeitink JA, van den Heuvel LP, Nijtmans LG. Mutations in NDUFAF3 (C3ORF60), encoding an NDUFAF4 (C6ORF66)-interacting complex I assembly protein, cause fatal neonatal mitochondrial disease. *Am J Hum Genet* 2000;84:718–27.

^dSaada A, Edvardson S, Rapoport M, Shaag A, Amry K, Miller C, Lorberboum-Galski H, Elpeleg O. C6ORF66 is an assembly factor of mitochondrial complex I. *Am J Hum Genet* 2000;82:32–8.

^eGerards M, Sluiter W, van den Bosch BJ, de Wit LE, Calis CM, Frentzen M, Akbari H, Schoonderwoerd K, Scholte HR, Jongbloed RJ, Hendrickx AT, de Coe IF, Smeets HJ. Defective complex I assembly due to C20orf7 mutations as a new cause of Leigh syndrome. *J Med Genet* 2000;47:507–12, Sugiana C, Pagliarini DJ, McKenzie M, Kirby DM, Salemi R, Abu-Amro KK, Dahl HH, Hutchison WM, Vascotto KA, Smith SM, Newbold RF, Christodoulou J, Calvo S, Mootha VK, Ryan MT, Thorburn DR. Mutation of C20orf7 disrupts complex I assembly and causes lethal neonatal mitochondrial disease. *Am J Hum Genet* 2000;83:468–78.

^fBianciardi L, Imperatore V, Fernandez-Vizarra E, Lopomo A, Falabella M, Furini S, Galluzzi P, Grosso S, Zeviani M, Renieri A, Mari F, Frullanti E. Exome sequencing coupled with mRNA analysis identifies NDUFAF6 as a Leigh gene. *Mol Genet Metab* 2016;119:214–22.

^gCalvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burrill NP, Rivas M, Guiducci C, Bruno DL, Goldberger OA, Redman MC, Wiltshire E, Wilson CJ, Altshuler D, Gabriel SB, Daly MJ, Thorburn DR, Mootha VK. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat Genet* 2010;42:851–8.

^hsee footnote g.

ⁱHaack TB, Danhauser K, Haberberger B, Hoser J, Strecker V, Boehm D, Uziel G, Lamantea E, Invernizzi F, Poulton J, Rolinski B, Iuso A, Biskup S, Schmidt T, Mewes HW, Wittig I, Meitinger T, Zeviani M, Prokisch H. Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nature Genet* 2010;42:1131–4.

^jGhezzi D, Goffrini P, Uziel G, Horvath R, Klopstock T, Lochmuller H, D'Adamo P, Gasparini P, Strom TM, Prokisch H, Invernizzi F, Ferrero I, Zeviani M. SDHAF1, encoding an LYR complex-II specific assembly factor, is mutated in SDH-defective infantile leukoencephalopathy. *Nature Genet* 2009;41:654–6.

- ^kHao HX, Khalimonchuk O, Schraders M, Dephoure N, Bayley JP, Kunst H, Devilee P, Cremers CW, Schiffman JD, Bentz BG, Gygi SP, Winge DR, Kremer H, Rutter J. SDH5, a gene required for flavination of succinate dehydrogenase, is mutated in paraganglioma. *Science* 2009;325:1139–42.
- ^lde Lonlay P, Valnot I, Barrientos A, Gorbatyuk M, Tzagoloff A, Taanman JW, Benayoun E, Chretien D, Kadhom N, Lombes A, de Baulny HO, Niaudet P, Munnich A, Rustin P, Rotig A. A mutant mitochondrial respiratory chain assembly protein causes complex III deficiency in patients with tubulopathy, encephalopathy and liver failure. *Nat Genet* 2001;29:57–60. Hinson JT, Fantin VR, Schonberger J, Breivik N, Siem G, McDonough B, Sharma P, Keogh I, Godinho R, Santos F, Esparza A, Nicolau Y, Selvaag E, Cohen BH, Hoppel CL, Tranebjaerg L, Eavey RD, Seidman JG, Seidman CE. Missense mutations in the BCS1L gene as a cause of the Bjornstad syndrome. *N Engl J Med* 2007;356:809–19. Visapaa I, Fellman V, Vesa J, Dasvarma A, Hutton JL, Kumar V, Payne GS, Makarow M, Van Coster R, Taylor RW, Turnbull DM, Suomalainen A, Peltonen L. GRACILE syndrome, a lethal metabolic disorder with iron overload, is caused by a point mutation in BCS1L. *Am J Hum Genet* 2002;71:863–76.
- ^mTucker EJ, Wanschers BF, Szklarczyk R, Mountford HS, Wijeyeratne XW, van den Brand MA, Leenders AM, Rodenburg RJ, Reljic B, Compton AG, Frazier AE, Bruno DL, Christodoulou J, Endo H, Ryan MT, Nijtmans LG, Huynen MA, Thorburn DR. Mutations in the UQCC1-interacting protein, UQCC2, cause human complex III deficiency associated with perturbed cytochrome b protein expression. *PLoS Genet* 2013;9:e1004034.
- ⁿWanschers BF, Szklarczyk R, van den Brand MA, Jonckheere A, Suijskens J, Smeets R, Rodenburg RJ, Stephan K, Helland IB, Elkamil A, Rootwelt T, Ott M, van den Heuvel L, Nijtmans LG, Huynen MA. A mutation in the human CBP4 ortholog UQCC3 impairs complex III assembly, activity and cytochrome b stability. *Hum Mol Genet* 2014;23:6356–65.
- ^oTiranti V, Hoertnagel K, Carrozzo R, Galimberti C, Munaro M, Granatiero M, Zelante L, Gasparini P, Marzella R, Rocchi M, Bayona-Bafaluy MP, Enriquez JA, Uziel G, Bertini E, Dionisi-Vici C, Franco B, Meitinger T, Zeviani M. Mutations of SURF-1 in Leigh Disease associated with cytochrome c oxidase deficiency. *Am J Hum Genet* 1998;63:1609–21. Zhu Z, Yao J, Johns T, Fu K, De Bie I, Macmillan C, Cuthbert AP, Newbold RF, Wang J, Chevrete M, Brown GK, Brown RM, Shoubridge EA. SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome. *Nat Genet* 1998;20:337–43.
- ^pValnot I, Osmond S, Gigarel N, Mehaye B, Amiel J, Cormier-Daire V, Munnich A, Bonnefont JP, Rustin P, Rotig A. Mutations of the SCO1 gene in mitochondrial cytochrome c oxidase deficiency with neonatal-onset hepatic failure and encephalopathy. *Am J Hum Genet* 2000;67:1104–9.
- ^qPapadopoulou LC, Sue CM, Davidson MM, Tanji K, Nishino I, Sadlock JE, Krishna S, Walker W, Selby J, Glerum DM, Coster RV, Lyon G, Scalais E, Lebel R, Kaplan P, Shanske S, De Vivo DC, Bonilla E, Hirano M, DiMauro S, Schon EA. Fatal infantile cardioencephalomyopathy with COX deficiency and mutations in SCO2, a COX assembly gene. *Nature Genet* 1999;23:333–7.
- ^rAntonicka H, Leary SC, Guercin GH, Agar JN, Horvath R, Kennaway NG, Harding CO, Jaksch M, Shoubridge EA. Mutations in COX10 result in a defect in mitochondrial heme A biosynthesis and account for multiple, early-onset clinical phenotypes associated with isolated COX deficiency. *Hum Mol Genet* 2003a;12:2693–702. Valnot I, Osmond S, Gigarel N, Mehaye B, Amiel J, Cormier-Daire V, Munnich A, Bonnefont JP, Rustin P, Rotig A. Mutations of the SCO1 gene in mitochondrial cytochrome c oxidase deficiency with neonatal-onset hepatic failure and encephalopathy. *Am J Hum Genet* 2000;67:1104–09.
- ^sWeraarpachai W, Sasarman F, Nishimura T, Antonicka H, Aure K, Rotig A, Lombes A, Shoubridge EA. Mutations in C12orf62, a factor that couples COX I synthesis with cytochrome c oxidase assembly, cause fatal neonatal lactic acidosis. *Am J Hum Genet* 2012;90:142–51.
- ^tAntonicka H, Mattman A, Carlson C.G, Glerum DM, Hoffbuhr KC, Leary SC, Kennaway NG, Shoubridge EA. Mutations in COX15 produce a defect in the mitochondrial heme biosynthetic pathway, causing early-onset fatal hypertrophic cardiomyopathy. *Am J Hum Genet* 2003b;72:101–14. Oquendo CE, Antonicka H, Shoubridge EA, Reardon W, Brown GK. Functional and genetic studies demonstrate that mutation in the COX15 gene can cause Leigh syndrome. *J Med Genet* 2004;41:540–4.
- ^uSzklarczyk R, Wanschers BF, Nijtmans LG, Rodenburg RJ, Zschocke J, Dikow N, van den Brand MA, Hendriks-Franssen MG, Gilissen C, Veltman JA, Nootboom M, Koopman WJ, Willems PH, Smeitink JA, Huynen MA, van den Heuvel LP. A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia. *Hum Mol Genet* 2004;22:656–67.
- ^vOstergaard E, Weraarpachai W, Ravn K, Born AP, Jonson L, Duno M, Wibrand F, Shoubridge EA, Vissing J. Mutations in COA3 cause isolated complex IV deficiency associated with neuropathy, exercise intolerance, obesity, and short stature. *J Med Genet* 2015;52:203–7.
- ^wHuigslout M, Nijtmans LG, Szklarczyk R, Baars MJ, van den Brand MA, Hendriksfranssen MG, van den Heuvel LP, Smeitink JA, Huynen MA, Rodenburg RJ. A mutation in C2orf64 causes impaired cytochrome c oxidase assembly and mitochondrial cardiomyopathy. *Am J Hum Genet* 2011;88:488–93.
- ^xBaertling F, M AMvdB, Hertecant JL, Al-Shamsi A, L, PvdH, Distelmaier F, Mayatepek E, Smeitink JA, Nijtmans LG, Rodenburg RJ. Mutations in COA6 cause cytochrome c oxidase deficiency and neonatal hypertrophic cardiomyopathy. *Hum Mutat* 2015;36:34–8.
- ^yMootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci USA* 2003;100:605–10.

- ^zGhezzi D, Saada A, D'Adamo P, Fernandez-Vizarra E, Gasparini P, Tiranti V, Elpeleg O, Zeviani M. FASTKD2 nonsense mutation in an infantile mitochondrial encephalomyopathy associated with cytochrome c oxidase deficiency. *Am J Hum Genet* 2008;83:415–23.
- ^{aa}Weraarpachai W, Antonicka H, Sasarman F, Seeger J, Schrank B, Kolesar JE, Lochmuller H, Chevrette M, Kaufman BA, Horvath R, Shoubridge EA. Mutation in TACO1, encoding a translational activator of COX I, results in cytochrome c oxidase deficiency and late-onset Leigh syndrome. *Nature Genet* 2009;41:833–7.
- ^{bb}De Meirleir L, Seneca S, Lissens W, De Clercq I, Eyskens F, Gerlo E, Smet J, Van Coster R. Respiratory chain complex V deficiency due to a mutation in the assembly gene ATP12. *J Med Genet* 2004;41:120–4.
- ^{cc}Cizkova A, Stranecky V, Mayr J.A, Tesarova M, Havlickova V, Paul J, Ivanek R, Kuss AW, Hansikova H, Kaplanova V, Vrbacky M, Hartmannova H, Noskova L, Honzik T, Drahota Z, Magner M, Hejzlarova K, Sperl W, Zeman J, Houstek J, Kmoch S. TMEM70 mutations cause isolated ATP synthase deficiency and neonatal mitochondrial encephalocardiomyopathy. *Nat Genet* 2008;40:1288–90.
- ^{dd}Naviaux RK, Nguyen KV. POLG mutations associated with Alpers' syndrome and mitochondrial DNA depletion. *Ann Neurol* 2004;55:706–12. Rovio AT, Marchington DR, Donat S, Schuppe HC, Abel J, Fritsche E, Elliott DJ, Laippala P, Ahola AL, McNay D, Harrison RF, Hughes B, Barrett T, Bailey DM, Mehmet D, Jequier AM, Hargreave TB, Kao SH, Cummins JM, Barton DE, Cooke HJ, Wei YH, Wichmann L, Poulton J, Jacobs HT. Mutations at the mitochondrial DNA polymerase (POLG) locus associated with male infertility. *Nat Genet* 2001;29:261–2. Van Goethem G, Deraut B, Lofgren A, Martin JJ, Van Broeckhoven C. Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nature Genet* 2001;28:211–2.
- ^{ee}Longley MJ, Clark S, Yu Wai Man C, Hudson G, Durham SE, Taylor RW, Nightingale S, Turnbull DM, Copeland WC, Chinnery PF. Mutant POLG2 disrupts DNA polymerase gamma subunits and causes progressive external ophthalmoplegia. *Am J Hum Genet* 2006;78:1026–34.
- ^{ff}Kaukonen J, Juselius JK, Tiranti V, Kyttala A, Zeviani M, Comi GP, Keranen S, Peltonen L, Suomalainen A. Role of adenine nucleotide translocator one in mtDNA maintenance. *Science* 2000;289:782–5. Strauss KA, Dubiner L, Simon M, Zaragoza M, Sengupta PP, Li P, Narula N, Dreike S, Platt J, Procaccio V, Ortiz-Gonzalez XR, Puffenberger EG, Kelley RI, Morton DH, Narula J, Wallace DC. Severity of cardiomyopathy associated with adenine nucleotide translocator-1 deficiency correlates with mtDNA haplogroup. *Proc Natl Acad Sci USA* 2013;110:3253–458.
- ^{gg}Spinazzola A, Viscomi C, Fernandez-Vizarra E, Carrara F, D'Adamo P, Calvo S, Marsano RM, Donnini C, Weiher H, Strisciuglio P, Parini R, Sarzi E, Chan A, DiMauro S, Rotig A, Gasparini P, Ferrero I, Mootha VK, Tiranti V, Zeviani M. MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nat Genet* 2006;38:570–5.
- ^{hh}Alexander C, Votruba M, Pesch UEA, Thiselton DL, Mayer S, Moore A, Rodriguez M, Kellner U, Leo-Kottler B, Auburger G, Bhat-tacharya SS, Wissinger B. OPA1, encoding a dynamin-related GTPase, is mutated in autosomal dominant optic atrophy linked to chromosome 3q28. *Nature Genet* 2000;26:211–5. Amati-Bonneau P, Valentino ML, Reynier P, Gallardo ME, Bornstein B, Boissiere A, Campos Y, Rivera H, de la Aleja JG, Carroccia R, Iommarini L, Labauge P, Figarella-Branger D, Marcorelles P, Furby A, Beauvais K, Letournel F, Liguori R, La Morgia C, Montagna P, Liguori M, Zanna C, Rugolo M, Cossarizza A, Wissinger B, Verry C, Schwarzenbacher R, Martin MA, Arenas J, Ayuso C, Garesse R, Lenaers G, Bonneau D, Carelli V. OPA1 mutations induce mitochondrial DNA instability and optic atrophy 'plus' phenotypes. *Brain* 2008;131:338–51. Delettre C, Lenaers G, Griffoin JM, Gigarel N, Lorenzo C, Belenguer P, Pelloquin L, Grosgeorge J, Turc-Carel C, Perret E, Astarie-Dequeker C, Lasquelles L, Arnaud B, Ducommun B, Kaplan J, Hamel CP. Nuclear gene OPA1, encoding a mitochondrial dynamin-related protein, is mutated in dominant optic atrophy. *Nat Genet* 2000;26:207–10.
- ⁱⁱKijima K, Numakura C, Izumino H, Umetsu K, Nezu A, Shiiki T, Ogawa M, Ishizaki Y, Kitamura T, Shozawa Y, Hayasaka K. Mitochondrial GTPase mitofusins two mutation in Charcot-Marie-Tooth neuropathy type 2A. *Hum Genet* 2005;116:23–7. Rouzier C, Bannwarth S, Chausseuot A, Chevrollier A, Verschueren A, Bonello-Palot N, Fragaki K, Cano A, Pouget J, Pellissier JF, Procaccio V, Chabrol B, Paquis-Flucklinger V. The MFN2 gene is responsible for mitochondrial DNA instability and optic atrophy 'plus' phenotype. *Brain* 2012;135:23–34. Zuchner S, Mersiyanova IV, Muglia M, Bissar-Tadmouri N, Rochelle J, Dadali EL, Zappia M, Nelis E, Patitucci A, Senderek J, Parman Y, Evgrafov O, Jonghe PD, Takahashi Y, Tsuji S, Pericak-Vance MA, Quattrone A, Battaloglu E, Polyakov AV, Timmerman V, Schroder JM, Vance JM. Mutations in the mitochondrial GTPase mitofusins two cause Charcot-Marie-Tooth neuropathy type 2A. *Nature Genet* 2004;36:449–51.
- ^{jj}Hudson G, Deschauer M, Busse K, Zierz S, Chinnery PF. Sensory ataxic neuropathy due to a novel C10orf2 mutation with probable germline mosaicism. *Neurology* 2005;64:371–3. Spelbrink JN, Li FY, Tiranti V, Nikali K, Yuan QP, Tariq M, Wanrooij S, Garrido N, Comi G, Morandi L, Santoro L, Toscano A, Fabrizi GM, Somer H, Croxson R, Beeson D, Poulton J, Suomalainen A, Jacobs HT, Zeviani M, Larsson C. Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nat Genet* 2001;28:223–31.
- ^{kk}Nishino I, Spinazzola A, Hirano, M. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. *Science* 1999;283:689–92.
- ^{ll}Mandel H, Szargel R, Labay V, Elpeleg O, Saada A, Shalata A, Anbinder Y, Berkowitz D, Hartman C, Barak M, Eriksson S, Cohen N. The deoxyguanosine kinase gene is mutated in individuals with depleted hepatocerebral mitochondrial DNA. *Nat Genet* 2001;29:337–41.
- ^{mm}Bourdon A, Minai L, Serre V, Jais JP, Sarzi E, Aubert S, Chretien D, de Lonlay P, Paquis-Flucklinger V, Arakawa H, Nakamura Y, Munnich A, Rotig A. Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. *Nature Genet* 2007;39:776–80. Shaibani A, Shchelochkov OA, Zhang S, Katsonis P, Lichtarge O, Wong LJ, Shinawi

- M. Mitochondrial neurogastrointestinal encephalopathy due to mutations in RRM2B. *Arch Neurol* 2009;66:1028–32. Tynismaa H, Ylikallio E, Patel M, Molnar MJ, Haller RG, Suomalainen A. A heterozygous truncating mutation in RRM2B causes autosomal-dominant progressive external ophthalmoplegia with multiple mtDNA deletions. *Am J Hum Genet* 2009;85:290–5.
- ⁿⁿElpeleg O, Miller C, Hershkovitz E, Bitner-Grindzicz M, Bondi-Rubinstein G, Rahman S, Pagnamenta A, Eshhar S, Saada A. Deficiency of the ADP-forming succinyl-CoA synthase activity is associated with encephalomyopathy and mitochondrial DNA depletion. *Am J Hum Genet* 2005;76:1081–6.
- ^{oo}Ostergaard E, Christensen E, Kristensen E, Mogensen B, Duno M, Shoubridge EA, Wibrand F. Deficiency of the alpha subunit of succinate-coenzyme A ligase causes fatal infantile lactic acidosis with mitochondrial DNA depletion. *Am J Hum Genet* 2007;81:383–7.
- ^{pp}Saada A, Shaag A, Mandel H, Nevo Y, Eriksson S, Elpeleg O. Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy. *Nat Genet* 2001;29:342–4.
- ^{qq}Stiles AR, Simon MT, Stover A, Eftekharian S, Khanlou N, Wang HL, Magaki S, Lee H, Partynski K, Dorrani N, Chang R, Martinez-Agosto JA, Abdenur JE. Mutations in TFAM, encoding mitochondrial transcription factor A, cause neonatal liver failure associated with mtDNA depletion. *Mol Genet Metab* 2016;119:91–9.
- ^{rr}Gai X, Ghezzi D, Johnson MA, Biagosch CA, Shamseldin HE, Haack TB, Reyes A, Tsukikawa M, Sheldon CA, Srinivasan S, Gorza M, Kremer LS, Wieland T, Strom TM, Polyak E, Place E, Consugar M, Ostrovsky J, Vidoni S, Robinson AJ, Wong LJ, Sondheim N, Salih MA, Al-Jishi E, Raab CP, Bean C, Furlan F, Parini R, Lamperti C, Mayr JA, Konstantopoulou V, Huemer M, Pierce EA, Meitinger T, Freisinger P, Sperl W, Prokisch H, Alkuraya FS, Falk MJ, Zeviani M. Mutations in FBXL4, encoding a mitochondrial protein, cause early-onset mitochondrial encephalomyopathy. *Am J Hum Genet* 2013;93:482–95.
- ^{ss}Kornblum C, Nicholls TJ, Haack TB, Scholer S, Peeva V, Danhauser K, Hallmann K, Zsurka G, Rorbach J, Iuso A, Wieland T, Sciacco M, Ronchi D, Comi GP, Moggio M, Quinzii CM, DiMauro S, Calvo SE, Mootha VK, Klopstock T, Strom TM, Meitinger T, Minczuk M, Kunz WS, Prokisch H. Loss-of-function mutations in MGME1 impair mtDNA replication and cause multisystemic mitochondrial disease. *Nat Genet* 2013;45:214–9.
- ^{tt}Jin H, May M, Tranebjaerg L, Kendall E, Fontan G, Jackson J, Subramony SH, Arena F, Lubs H, Smith S, Stevenson R, Schwartz C, Vetrie D. A novel X-linked gene, DDP, shows mutations in families with deafness (DFN-1), dystonia, mental deficiency and blindness. *Nature Genet* 1996;14:177–80.
- ^{uu}Davey KM, Parboosingh JS, McLeod DR, Chan A, Casey R, Ferreira P, Snyder FF, Bridge PJ, Bernier FP. Mutation of DNAJC19, a human homologue of yeast inner mitochondrial membrane co-chaperones, causes DCMA syndrome, a novel autosomal recessive Barth syndrome-like condition. *J Med Genet* 2006;43:385–93.
- ^{vv}Dallabona C, Diodato D, Kevelam SH, Haack TB, Wong LJ, Salomons GS, Baruffini E, Melchionda L, Mariotti C, Strom TM, Meitinger T, Prokisch H, Chapman K, Colley A, Rocha H, Ounap K, Schiffmann R, Salsano E, Savoiardo M, Hamilton EM, Abbink TE, Wolf NI, Ferrero I, Lamperti C, Zeviani M, Vanderver A, Ghezzi D, van der Knaap MS. Novel (ovario) leukodystrophy related to AARS2 mutations. *Neurology* 2014;82:2063–71. Gotz A, Tynismaa H, Euro L, Ellonen P, Hyotylainen T, Ojala T, Hamalainen RH, Tammela J, Raivio T, Oresic M, Karikoski R, Tammela O, Simola KO, Paetau A, Tyni T, Suomalainen A. Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy. *Am J Hum Genet* 2014;88:635–42.
- ^{www}Hallmann K, Zsurka G, Moskau-Hartmann S, Kirschner J, Korinthenberg R, Ruppert AK, Ozdemir O, Weber Y, Becker F, Lerche H, Elger CE, Thiele H, Nurnberg P, Sander T, Kunz WS. A homozygous splice-site mutation in CARS2 is associated with progressive myoclonic epilepsy. *Neurology* 2014;83:2183–7.
- ^{xx}Scheper GC, van der Klok T, van Andel RJ, van Berkel CG, Sissler M, Smet J, Muravina TI, Serkov SV, Uziel G, Bugiani M, Schiffmann R, Krageloh-Mann I, Smeitink JA, Florentz C, Van Coster R, Pronk JC, van der Knaap MS. Mitochondrial aspartyl-tRNA synthetase deficiency causes leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation. *Nat Genet* 2007;39:534–9.
- ^{yy}Steenweg ME, Ghezzi D, Haack T, Abbink TE, Martinelli D, van Berkel CG, Bley A, Diogo L, Grillo E, Te Water Naude J, Strom TM, Bertini E, Prokisch H, van der Knaap MS, Zeviani M. Leukoencephalopathy with thalamus and brainstem involvement and high lactate 'LTBL' caused by EARS2 mutations. *Brain* 2012;135:1387–94.
- ^{zz}Elo JM, Yadavalli SS, Euro L, Isohanni P, Gotz A, Carroll CJ, Valanne L, Alkuraya FS, Uusimaa J, Paetau A, Caruso EM, Pihko H, Ibba M, Tynismaa H, Suomalainen A. Mitochondrial phenylalanyl-tRNA synthetase mutations underlie fatal infantile Alpers encephalopathy. *Hum Mol Genet* 2012;21:4521–9. Yang Y, Liu W, Fang Z, Shi J, Che F, He C, Yao L, Wang E, Wu Y. A newly identified missense mutation in FARS2 causes autosomal-recessive spastic paraplegia. *Hum Mutat* 2016;37:165–9.
- ^{aaa}Antonellis A, Ellsworth RE, Sambuughin N, Puls I, Abel A, Lee-Lin SQ, Jordanova A, Kremensky I, Christodoulou K, Middleton LT, Sivakumar K, Ionasescu V, Funalot B, Vance JM, Goldfarb LG, Fischbeck KH, Green ED. Glycyl tRNA synthetase mutations in Charcot-Marie-Tooth disease type 2D and distal spinal muscular atrophy type V. *Am J Hum Genet* 2003;72:1293–9.
- ^{bbb}Pierce SB, Chisholm KM, Lynch ED, Lee MK, Walsh T, Opitz JM, Li W, Klevit RE, King MC. Mutations in mitochondrial histidyl tRNA synthetase HARS2 cause ovarian dysgenesis and sensorineural hearing loss of Perrault syndrome. *Proc Natl Acad Sci USA* 2011;108:6543–8.
- ^{ccc}Schwartzentruber J, Buhas D, Majewski J, Sasarman F, Papillon-Cavanagh S, Thiffault I, Sheldon KM, Massicotte C, Patry L, Simon M, Zare AS, McKernan KJ, Consortium FC, Michaud J, Boles RG, Deal CL, Desilets V, Shoubridge EA, Samuels ME. Mutation in the nuclear-encoded mitochondrial isoleucyl-tRNA synthetase IARS2 in patients with cataracts, growth hormone deficiency with

- short stature, partial sensorineural deafness, and peripheral neuropathy or with Leigh syndrome. *Hum Mutat* 2014;35:1285–89. Erratum *Hum Mutat* 2015:1236–1281.
- ^{ddd}McLaughlin HM, Sakaguchi R, Liu C, Igarashi T, Pehlivan D, Chu K, Iyer R, Cruz P, Cherukuri PF, Hansen NF, Mullikin JC, Program NCS, Biesecker LG, Wilson TE, Ionasescu V, Nicholson G, Searby C, Talbot K, Vance JM, Zuchner S, Szegedi K, Lupski JR, Hou YM, Green ED, Antonellis A. Compound heterozygosity for loss-of-function lysyl-tRNA synthetase mutations in a patient with peripheral neuropathy. *Am J Hum Genet* 2010;87:560–6, Santos-Cortez RL, Lee K, Azeem Z, Antonellis PJ, Pollock LM, Khan S, Irfanullah Andrade-Elizondo PB, Chiu I, Adams MD, Basit S, Smith JD, University of Washington Center for Mendelian G, Nickerson DA, McDermott BM, Jr, Ahmad W, Leal S. Mutations in KARS, encoding lysyl-tRNA synthetase, cause autosomal-recessive nonsyndromic hearing impairment DFNB89. *Am J Hum Genet* 2013;93:132–40.
- ^{eee}Casey JP, McGettigan P, Lynam-Lennon N, McDermott M, Regan R, Conroy J, Bourke B, O'Sullivan J, Crushell E, Lynch S, Ennis S. Identification of a mutation in LARS as a novel cause of infantile hepatopathy. *Mol Genet Metab* 2012;106:351–8.
- ^{fff}Pierce SB, Gersak K, Michaelson-Cohen R, Walsh T, Lee MK, Malach D, Kleit RE, King MC, Levy-Lahad E. Mutations in LARS2, encoding mitochondrial leucyl-tRNA synthetase, lead to premature ovarian failure and hearing loss in Perrault syndrome. *Am J Hum Genet* 2013;92:614–20.
- ^{ggg}Sofou K, Kollberg G, Holmstrom M, Davila M, Darin N, Gustafsson CM, Holme E, Oldfors A, Tulinius M, Asin-Cayuela J. Whole exome sequencing reveals mutations in NARS2 and PARS2, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. *Mol Genet Genom Med* 2015;3:59–68.
- ^{hhh}see footnote ggg.
- ⁱⁱⁱEdvardson S, Shaag A, Kolesnikova O, Gomori JM, Tarassov I, Einbinder T, Saada A, Elpeleg O. Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. *Am J Hum Genet* 2007;81:857–62.
- ⁱⁱⁱBelostotsky R, Ben-Shalom E, Rinat C, Becker-Cohen R, Feinstein S, Zeligson S, Segel R, Elpeleg O, Nassar S, Frishberg Y. Mutations in the mitochondrial seryl-tRNA synthetase cause hyperuricemia, pulmonary hypertension, renal failure in infancy and alkalosis, HUPRA syndrome. *Am J Hum Genet* 2011;88:193–200.
- ^{kkk}Diodato D, Melchionda L, Haack TB, Dallabona C, Baruffini E, Donnini C, Granata T, Ragona F, Balestri P, Margollicci M, Lamantea E, Nasca A, Powell CA, Minczuk M, Strom TM, Meitinger T, Prokisch H, Lamperti C, Zeviani M, Ghezzi D. VARS2 and TARS2 mutations in patients with mitochondrial encephalomyopathies. *Human Mutat* 2014;35:983–9.
- ^{lll}see footnote kkk.
- ^{mmm}Riley LG, Cooper S, Hickey P, Rudinger-Thirion J, McKenzie M, Compton A, Lim SC, Thorburn D, Ryan MT, Giege R, Bahlo M, Christodoulou J. Mutation of the mitochondrial tyrosyl-tRNA synthetase gene, YARS2, causes myopathy, lactic acidosis, and sideroblastic anemia–MLASA syndrome. *Am J Hum Genet* 2010;87:52–9.
- ⁿⁿⁿCoenen MJ, Antonicka H, Ugalde C, Sasarman F, Rossi R, Heister JG, Newbold RF, Trijbels FJ, van den Heuvel LP, Shoubridge EA, Smeitink JA. Mutant mitochondrial elongation factor G1 and combined oxidative phosphorylation deficiency. *N Engl J Med* 2004;351:2080–6.
- ^{ooo}Smeitink JA, Elpeleg O, Antonicka H, Diepstra H, Saada A, Smits P, Sasarman F, Vriend G, Jacob-Hirsch J, Shaag A, Rechavi G, Welling B, Horst J, Rodenburg RJ, van den Heuvel B, Shoubridge EA. Distinct clinical phenotypes associated with a mutation in the mitochondrial translation elongation factor EFTs. *Am J Hum Genet* 2006;79:869–77.
- ^{ppp}Valente L, Tiranti V, Marsano RM, Malfatti E, Fernandez-Vizarra E, Donnini C, Mereghetti P, De Gioia L, Burlina A, Castellani C, Comi GP, Savasta S, Ferrero I, Zeviani M. Infantile encephalopathy and defective mitochondrial DNA translation in patients with mutations of mitochondrial elongation factors EFG1 and EFTu. *Am J Hum Genet* 2007;80:44–58. Erratum *Am J Hum Genet* 2007;80:580.
- ^{qqq}Kopajtich R, Nicholls TJ, Rorbach J, Metodiev MD, Freisinger P, Mandel H, Vanlander A, Ghezzi D, Carrozzo R, Taylor RW, Marquard K, Murayama K, Wieland T, Schwarzmayer T, Mayr JA, Pearce SF, Powell CA, Saada A, Ohtake A, Invernizzi F, Lamantea E, Sommerville EW, Pyle A, Chinnery PF, Crushell E, Okazaki Y, Kohda M, Kishita Y, Tokuzawa Y, Assouline Z, Rio M, Feillet F, Mousson de Camaret B, Chretien D, Munnich A, Menten B, Sante T, Smet J, Regal L, Lorber A, Khoury A, Zeviani M, Strom TM, Meitinger T, Bertini ES, Van Coster R, Klopstock T, Rotig A, Haack TB, Minczuk M, Prokisch H. Mutations in GTPBP3 cause a mitochondrial translation defect associated with hypertrophic cardiomyopathy, lactic acidosis, and encephalopathy. *Am J Hum Genet* 2014;95:708–20.
- ^{rrr}Haack TB, Haberberger B, Frisch EM, Wieland T, Iuso A, Gorza M, Strecker V, Graf E, Mayr JA, Herberg U, Hennermann JB, Klopstock T, Kuhn KA, Ahting U, Sperl W, Wilichowski E, Hoffmann GF, Tesarova M, Hansikova H, Zeman J, Plecko B, Zeviani M, Wittig I, Strom TM, Schuelke M, Freisinger P, Meitinger T, Prokisch H. Molecular diagnosis in mitochondrial complex I deficiency using exome sequencing. *J Med Genet* 2012;49:277–83.
- ^{sss}Ghezzi D, Baruffini E, Haack TB, Invernizzi F, Melchionda L, Dallabona C, Strom TM, Parini R, Burlina AB, Meitinger T, Prokisch H, Ferrero I, Zeviani M. Mutations of the mitochondrial-tRNA modifier MTO1 cause hypertrophic cardiomyopathy and lactic acidosis. *Am J Hum Genet* 2012;90:1079–87.
- ^{ttt}Powell CA, Kopajtich R, D'Souza AR, Rorbach J, Kremer LS, Husain RA, Dallabona C, Donnini C, Alston CL, Griffin H, Pyle A, Chinnery PF, Strom TM, Meitinger T, Rodenburg RJ, Schottmann G, Schuelke M, Romain N, Haller RG, Ferrero I, Haack TB, Taylor RW, Prokisch H, Minczuk M. TRMT5 mutations cause a defect in post-transcriptional modification of mitochondrial tRNA associated with multiple respiratory-chain deficiencies. *Am J Hum Genet* 2015;97:319–28.

- ^{uuu}Metodiev MD, Thompson K, Alston CL, Morris AAM, He L, Assouline Z, Rio M, Bahi-Buisson N, Pyle A, Griffin H, Siira S, Filipovska A, Munnich A, Chinnery PF, McFarland R, Rotig A, Taylor RW. Recessive mutations in TRMT10C cause defects in mitochondrial RNA processing and multiple respiratory chain deficiencies. *Am J Hum Genet* 2016;98:993–1000. Erratum *Am J Hum Genet* 2016;1099–246.
- ^{vvv}Zeharia A, Shaag A, Pappo O, Mager-Heckel AM, Saada A, Beinat M, Karicheva O, Mandel H, Ofek N, Segel R, Marom D, Rotig A, Tarassov I, Elpeleg O. Acute infantile liver failure due to mutations in the TRMU gene. *Am J Hum Genet* 2009;85:401–7.
- ^{www}Valente L, Tiranti V, Marsano RM, Malfatti E, Fernandez-Vizarra E, Donnini C, Mereghetti P, De Gioia L, Burlina A, Castellan C, Comi GP, Savasta S, Ferrero I, Zeviani M. Infantile encephalopathy and defective mitochondrial DNA translation in patients with mutations of mitochondrial elongation factors EFG1 and EFTu. *Am J Hum Genet* 2007;80:44–58. Erratum: *Am. J. Hum. Genet* 2007;80:580.
- ^{xxx}Fukumura S, Ohba C, Watanabe T, Minagawa K, Shimura M, Murayama K, Ohtake A, Saitsu H, Matsumoto N, Tsutsumi H. Compound heterozygous GFM2 mutations with Leigh syndrome complicated by arthrogryposis multiplex congenita. *J Hum Genet* 2015;60:509–13.
- ^{yyy}Antonicka H, Ostergaard E, Sasarman F, Weraarpachai W, Wibrand F, Pedersen AM, Rodenburg RJ, van der Knaap MS, Smeitink JA, Chrzanowska-Lightowlers ZM, Shoubridge EA. Mutations in C12orf65 in patients with encephalomyopathy and a mitochondrial translation defect. *Am J Hum Genet* 2010;87:115–122. Shimazaki H, Takiyama Y, Ishiura H, Sakai C, Matsushima Y, Hatakeyama H, Honda J, Sakoe K, Naoi T, Namekawa M, Fukuda Y, Takahashi Y, Goto J, Tsuji S, Goto Y, Nakano I, Japan Spastic Paraplegia Research, C. A homozygous mutation of C12orf65 causes spastic paraplegia with optic atrophy and neuropathy (SPG55). *J Med Genet* 2012;49:777–84.
- ^{zzz}Janer A, Antonicka H, Lalonde E, Nishimura T, Sasarman F, Brown GK, Brown RM, Majewski J, Shoubridge EA. An RMND1 Mutation causes encephalopathy associated with multiple oxidative phosphorylation complex deficiencies and a mitochondrial translation defect. *Am J Hum Genet* 2012;91:737–43.
- ^{aaaa}Galmiche L, Serre V, Beinat M, Assouline Z, Lebre AS, Chretien D, Nietschke P, Benes V, Boddaert N, Sidi D, Brunelle F, Rio M, Munnich A, Rotig A. Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy. *Hum Mutat* 2011;32:1225–31.
- ^{bbbb}Menezes MJ, Guo Y, Zhang J, Riley LG, Cooper ST, Thorburn DR, Li J, Dong D, Li Z, Glessner J, Davis RL, Sue CM, Alexander SI, Arbuckle S, Kirwan P, Keating BJ, Xu X, Hakonarson H, Christodoulou J. Mutation in mitochondrial ribosomal protein S7 (MRPS7) causes congenital sensorineural deafness, progressive hepatic and renal failure and lactic acidemia. *Hum Mol Genet* 2015;24:2297–307.
- ^{cccc}Serre V, Rozanska A, Beinat M, Chretien D, Boddaert N, Munnich A, Rotig A, Chrzanowska-Lightowlers ZM. Mutations in mitochondrial ribosomal protein MRPL12 leads to growth retardation, neurological deterioration and mitochondrial translation deficiency. *Biochim Biophys Acta* 2013;1832:1304–12.
- ^{dddd}Miller C, Saada A, Shaul N, Shabtai N, Ben-Shalom E, Shaag A, HersHKovitz E, Elpeleg O. Defective mitochondrial translation caused by a ribosomal protein (MRPS16) mutation. *Ann Neurol* 2004;56:734–8.
- ^{eeee}Saada A, Shaag A, Arnon S, Dolfin T, Miller C, Fuchs-Telem D, Lombes A, Elpeleg O. Antenatal mitochondrial disease caused by mitochondrial ribosomal protein (MRPS22) mutation. *J Med Genet* 2007;44:784–6.
- ^{ffff}Carroll CJ, Isohanni P, Poyhonen R, Euro L, Richter U, Brilhante V, Gotz A, Lahtinen T, Paetau A, Pihko H, Battersby BJ, Tynismaa H, Suomalainen A. Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy. *J Med Genet* 2013;50:151–9.
- ^{gggg}Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Canizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, DeMichele G, Filla A, De Frutos R, Palau F, Patel P, DiDonato S, Mandel J, Coccoza S, Koenig M, Pandolfo M. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 1996;271:1423–7, Rotig A, de Lonlay P, Chretien D, Foury F, Koenig M, Sidi D, Munnich A, Rustin P. Aconitase and mitochondrial iron-sulphur protein deficiency in Friedreich ataxia. *Nat Genet* 1997;17:215–7.
- ^{hhhh}Allikmets R, Raskind WH, Hutchinson A, Schueck ND, Dean M, Koeller DM. Mutation of a putative mitochondrial iron transporter gene (ABC7) in X-linked sideroblastic anemia and ataxia (XLSA/A). *Hum Mol Genet* 1999;8:743–9.
- ⁱⁱⁱⁱGuernsey DL, Jiang H, Campagna DR, Evans SC, Ferguson M, Kellogg MD, Lachance M, Matsuoka M, Nightingale M, Rideout A, Saint-Amant L, Schmidt PJ, Orr A, Bottomley SS, Fleming MD, Ludman M, Dyack S, Fernandez CV, Samuels ME. Mutations in mitochondrial carrier family gene SLC25A38 cause nonsyndromic autosomal recessive congenital sideroblastic anemia. *Nat Genet* 2009;41:651–3.
- ^{jjjj}Mochel F, Knight MA, Tong WH, Hernandez D, Ayyad K, Taivassalo T, Andersen PM, Singleton A, Rouault TA, Fischbeck KH, Haller RG. Splice mutation in the iron-sulfur cluster scaffold protein ISCU causes myopathy with exercise intolerance. *Am J Hum Genet* 2008;82:652–60.
- ^{kkkk}Cameron JM, Janer A, Levandovskiy V, Mackay N, Rouault TA, Tong WH, Ogilvie I, Shoubridge EA, Robinson BH. Mutations in iron-sulfur cluster scaffold genes NFU1 and BOLA3 cause a fatal deficiency of multiple respiratory chain and 2-oxoacid dehydrogenase enzymes. *Am J Hum Genet* 2011;89:486–95.
- ^{llll}see footnote kkkk.
- ^{mmmm}Al-Hassnan ZN, Al-Dosary M, Alfadhel M, Fageih EA, Alsagob M, Kenana R, Almass R, Al-Harazi OS, Al-Hindi H, Malibari OI, Almutari FB, Tulbah S, Alhadeq F, Al-Sheddi T, Alamro R, AlAsmari A, Almuntashri M, Alshaalan H, Al-Mohanna FA, Colak D, Kaya N. ISCA2 mutation causes infantile neurodegenerative mitochondrial disorder. *J Med Genet* 2015;52:186–94.

- nnnnAjit Bolar N, Vanlander AV, Wilbrecht C, Van der Aa N, Smet J, De Paepe B, Vandeweyer G, Kooy F, Eyskens F, De Latter E, Delanghe G, Govaert P, Leroy JG, Loeys B, Lill R, Van Laer L, Van Coster R. Mutation of the iron-sulfur cluster assembly gene IBA57 causes severe myopathy and encephalopathy. *Hum Mol Genet* 2013;22:2590–602. Lossos A, Stumpfig C, Stevanin G, Gausson M, Zimmerman BE, Mundwiler E, Asulin M, Chamma L, Sheffer R, Misk A, Dotan S, Gomori JM, Ponger P, Brice A, Lerer I, Meiner V, Lill R. Fe/S protein assembly gene IBA57 mutation causes hereditary spastic paraplegia. *Neurology* 2015;84:659–67.
- ooooLim SC, Friemel M, Marum JE, Tucker EJ, Bruno DL, Riley LG, Christodoulou J, Kirk EP, Boneh A, DeGennaro CM, Springer M, Mootha VK, Rouault TA, Leimkuhler S, Thorburn DR, Compton AG. Mutations in LYRM4, encoding iron-sulfur cluster biogenesis factor ISD11, cause deficiency of multiple respiratory chain complexes. *Hum Mol Genet* 2013;22:4460–73.
- ppppInvernizzi F, Tigano M, Dallabona C, Donnini C, Ferrero I, Cremonte M, Ghezzi D, Lamperti C, Zeviani MA. homozygous mutation in LYRM7/MZM1L associated with early onset encephalopathy, lactic acidosis, and severe reduction of mitochondrial complex III activity. *Hum Mutat* 2013;34:1619–22.
- qqqqSpiegel R, Saada A, Halvardson J, Soiferman D, Shaag A, Edvardson S, Horovitz Y, Khayat M, Shalev SA, Feuk L, Elpeleg O. Deleterious mutation in FDX1L gene is associated with a novel mitochondrial muscle myopathy. *Eur J Hum Genet* 2014;22:902–6.
- rrrrQuinzii C, Naini A, Salviati L, Trevisson E, Navas P, Dimauro S, Hirano M. A mutation in *para*-hydroxybenzoate-polyprenyl transferase (COQ2) causes primary coenzyme Q10 deficiency. *Am J Hum Genet* 2006;78:345–9.
- ssssSalviati L, Trevisson E, Rodriguez Hernandez MA, Casarin A, Pertegato V, Doimo M, Cassina M, Agosto C, Desbats MA, Sartori G, Sacconi S, Memo L, Zuffardi O, Artuch R, Quinzii C, Dimauro S, Hirano M, Santos-Ocana C, Navas P. Haploinsufficiency of COQ4 causes coenzyme Q10 deficiency. *J Med Genet* 2012;49:187–91.
- ttttMalicdan MCV, Vilboux T, Ben-Zeev B, Guo J, Eliyahu A, Pode-Shakked B, Dori A, Kakani S, Chandrasekharappa SC, Ferreira CR, Shelestovich N, Marek-Yagel D, Pri-Chen H, Blatt I, Niederhuber JE, He L, Toro C, Taylor RW, Deeken J, Yardeni T, Wallace DC, Gahl WA, Anikster Y. A novel inborn error of the coenzyme Q10 biosynthesis pathway: cerebellar ataxia and static encephalomyopathy due to COQ5 C-methyltransferase deficiency. *Hum Mutat* 2018;39:69–79.
- uuuuHeeringa SF, Chernin G, Chaki M, Zhou W, Sloan AJ, Ji Z, Xie LX, Salviati L, Hurd TW, Vega-WarnerV, Killen PD, Raphael Y, Ashraf S, Ovunc B, Schoeb DS, McLaughlin HM, Airik R, Vlangos CN, Gbadegesin R, Hinkes B, Saisawat P, Trevisson E, Doimo M, Casarin A, Pertegato V, Giorgi G, Prokisch H, Rotig A, Nurnberg G, Becker C, Wang S, Ozaltin F, Topaloglu R, Bakaloglu A, Bakaloglu SA, Muller D, Beissert A, Mir S, Berdeli A, Varpizen S, Zenker M, MatejasV, Santos-Ocana C, Navas P, Kusakabe T, Kispert A, Akman S, Soliman NA, Krick S, Mundel P, Reiser J, Nurnberg P, Clarke CF, Wiggins RC, Faul C, Hildebrandt F. COQ6 mutations in human patients produce nephrotic syndrome with sensorineural deafness. *J Clin Invest* 2011;121:2013–24.
- vvvFreyer C, Stranneheim H, Naess K, Mourier A, Felser A, Maffezzini C, Lesko N, Bruhn H, Engvall M, Wiborn R, Barbaro M, Hinze Y, Magnusson M, Andeer R, Zetterstrom RH, von Döbeln U, Wredenberg A, Wedell A. Rescue of primary ubiquinone deficiency due to a novel COQ7 defect using 2,4-dihydroxybenzoic acid. *J Med Genet* 2015;52:779–83.
- wwwwDuncan AJ, Bitner-Glindzic M, Meunier B, Costello H, Hargreaves IP, Lopez LC, Hirano M, Quinzii CM, Sadowski MI, Hardy J, Singleton A, Clayton PT, Rahman S. A nonsense mutation in COQ9 causes autosomal-recessive neonatal-onset primary coenzyme Q10 deficiency: a potentially treatable form of mitochondrial disease. *Am J Hum Genet* 2009;84:558–66.
- xxxxQuinzii CM, Kattah AG, Naini A, Akman HO, Mootha VK, DiMauro S, Hirano M. Coenzyme Q deficiency and cerebellar ataxia associated with an aprataxin mutation. *Neurology* 2005;64:539–41.
- yyyyMollet J, Giurgea I, Schlemmer D, Dallner G, Chretien D, Delahodde, A. Bacq D, de Lonlay P, Munnich A, Rotig A. Prenyl-diphosphate synthase, subunit 1 (PDSS1) and OH-benzoate polyprenyltransferase (COQ2) mutations in ubiquinone deficiency and oxidative phosphorylation disorders. *J Clin Invest* 2007;117:765–72.
- zzzzLopez LC, Schuelke M, Quinzii CM, Kanki T, Rodenburg RJ, Naini A, Dimauro S, Hirano, M. Leigh syndrome with nephropathy and CoQ10 deficiency due to decaprenyl diphosphate synthase subunit 2 (PDSS2) mutations. *Am J Hum Genet* 2006;79:1125–9.
- aaaaMollet J, Delahodde A, Serre V, Chretien D, Schlemmer D, Lombes A, Boddaert N, Desguerre I, de Lonlay P, de Baulny HO, Munnich A, Rotig A. CABP1 gene mutations cause ubiquinone deficiency with cerebellar ataxia and seizures. *Am J Hum Genet* 2008;82:623–30.
- bbbbbCasari G, De Fusco M, Ciarmatori S, Zeviani M, Mora M, Fernandez P, De Michele G, Filla A, Coccoza S, Marconi R, Durr A, Fontaine B, Ballabio A. Spastic paraplegia and OXPHOS impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease. *Cell* 1998;93:973–83.
- ccccMagen D, Georgopoulos C, Bross P, Ang D, Segev Y, Goldsher D, Nemirovski A, Shahar E, Ravid S, Luder A, Heno B, Gershoni-Baruch R, Skorecki K, Mandel H. Mitochondrial hsp60 chaperonopathy causes an autosomal-recessive neurodegenerative disorder linked to brain hypomyelination and leukodystrophy. *Am J Hum Genet* 2008;83:30–42.
- dddddWaterham HR, Koster J, van Roermund CW, Mooyer PA, Wanders RJ, Leonard JV. A lethal defect of mitochondrial and peroxisomal fission. *N Engl J Med* 2007;356:1736–41.
- eeeeBione S, D'Adamo P, Maestrini E, Gedeon AK, Bolhuis PA, Toniolo DA novel X-linked gene, G4.5. is responsible for Barth syndrome. *Nat Genet* 1996;12:385–9, D'Adamo P, Fassone L, Gedeon A, Janssen EA, Bione S, Bolhuis PA, Barth PG, Wilson M, Haan E, Orstavik KH, Patton MA, Green AJ, Zammarchi E, Donati MA, Toniolo D. The X-linked gene G4.5 is responsible for different infantile dilated cardiomyopathies. *Am J Hum Genet* 1997;61:862–7.
- ffffRidanpaa M, Sistonen P, Rockas S, Rimoin DL, Makitie O, Kaitila I. Worldwide mutation spectrum in cartilage-hair hypoplasia: ancient founder origin of the major70A→G mutation of the untranslated RMRP. *Eur J Hum Genet* 2002;10:439–47, Ridanpaa M, van Eenennaam H, Pelin K, Chadwick R, Johnson, C, Yuan B, vanVenrooij W, Pruijn G, Salmela R, Rockas S, Makitie O, Kaitila I, de

- la Chapelle A. Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, Cartilage-Hair Hypoplasia. *Cell* 2001;104:195–203.
- Mattheews PM, Marchington DR, Squier M, Land J, Brown RM, Brown GK. Molecular genetic characterization of an X-linked form of Leigh's syndrome. *Ann Neurol* 1993;33:652–5.
- Tiranti V, Briem E, Lamantea E, Mineri R, Papaleo E, Degioia L, Forlani F, Rinaldo P, Dickson P, Abu-Libdeh B, Cindro-Heberle L, Owaidha M, Jack RM, Christensen E, Burlina A, Zeviani M. ETHE1 mutations are specific to ethylmalonic encephalopathy. *J Med Genet* 2006;43:340–6.
- Tiranti V, D'Adamo P, Briem E, Ferrari G, Mineri R, Lamantea E, Mandel H, Balestri P, Garcia-Silva MT, Vollmer B, Rinaldo P, Hahn SH, Leonard J, Rahman S, Dionisi-Vici C, Garavaglia B, Gasparini P, Zeviani M. Ethylmalonic encephalopathy is caused by mutations in ETHE1, a gene encoding a mitochondrial matrix protein. *Am J Hum Genet* 2004;74:239–52.
- Bykhovskaya Y, Casas K, Mengesha E, Inbal A, Fischel-Ghodsian N. Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). *Am J Hum Genet* 2004;74:1303–08.
- Harel T, Yoon WH, Garone C, Gu S, Coban-Akdemir Z, Eldomery MK, Posey JE, Jhangiani SN, Rosenfeld JA, Cho MT, Fox S, Withers M, Brooks SM, Chiang T, Duraine L, Erdin S, Yuan B, Shao Y, Moussallem E, Lamperti C, Donati MA, Smith JD, McLaughlin HM, Eng CM, Walkiewicz M, Xia F, Pippucci T, Magini P, Seri M, Zeviani M, Hirano M, Hunter JV, Srour M, Zagnoni S, Lewis RA, Muzny DM, Lotze TE, Boerwinkle E, Baylor-Hopkins Center for Mendelian G, University of Washington Center for Mendelian G, Gibbs RA, Hickey SE, Graham BH, Yang Y, Buhas D, Martin DM, Potocki L, Graziano C, Bellen HJ, Lupski JR Recurrent de novo and biallelic variation of ATAD3A, encoding a mitochondrial membrane protein, results in distinct neurological syndromes. *Am J Hum Genet* 2016;99:831–45.

OXPPOS is acutely sensitive to toxins linking mitochondrial genetics to environmental challenges. Hence, a mitochondrial etiology of the common diseases explains the organ-specificity, the complex genetics, the age-related onset and progression, and the environmental involvement that contribute to common diseases [9].

A large number of studies have correlated mtDNA haplogroups with predisposition to common “complex diseases” [81]. A meta-analysis of the literature [129] affirmed that European haplogroup H is at increased risk for Alzheimer disease (AD), haplogroup K and haplogroups J-T are protective for Parkinson disease (PD), Asian haplogroups A and F as well as the control region nt 16189C affected increased risk of type 2 diabetes mellitus (T2DM), European haplogroups J and W are associated with increased longevity, and nt 10398G is associated with breast cancer risk. Overall, European haplogroups H, K, and J have significant effects of common disease risk [129].

Considerable evidence has accumulated implicating mitochondrial dysfunction in diabetes and metabolic syndrome [150,151]. Among Europeans, there is the suggestion that haplogroup U [152,153] and haplogroups J/T [154–156] may be at increased risk of T2DM. However, several other studies have not observed associations between mtDNA haplogroups and T2DM in Europeans. This is likely the product of the interaction between mtDNA and nDNA factors, as a carefully

controlled study of Israeli Jewish patients with diabetes revealed that subhaplogroup J1 is 2.4-fold underrepresented in T2DM patients whose parents are not diabetic versus patients whose parents are diabetic, suggesting that subhaplogroup J1 increases the risk of T2DM in association with additional nDNA genetic factors [155]. Among Japanese and Koreans, haplogroups D5 and F have been associated with increased risk and N9a with decreased risk of T2DM [157]. These results have been corroborated in gene expression studies in cybrids harboring D5, F, and N9a mtDNAs, which revealed that D5 and F had similar gene expression profiles while N9a cybrids showed an increased expression of OXPPOS genes relative to the D5 and F cybrids [158]. In the Chinese Uyghur, haplogroups H and D4 were at increased risk for T2DM [159], while in the Han Chinese, N9a and M8a [160] and M9 [161] were found to be at increased the risk of diabetes and N9a at increased risk of diabetic nephropathy. In a separate Japanese study, M8a was also associated with diabetes and B4c was associated with obesity [162]. In cybrid studies, haplogroup F1 was associated with significantly lower complex I activity and respiration than B4c and M9 [163]. In relation to single mtDNA nucleotide variants, the 10398G>A variant is associated with increased risk of T2DM in India [164] and Japan [161] while the mtDNA control region variant at nt 16189C is associated with increased T2DM risk when present in multiple mtDNA lineages [161,165].

Haplogroups have also been reported to affect the risk of developing associated diabetes sequelae [153,166]. For example, in Israeli Jews, H is associated with retinopathy, H3 with neuropathy, U3 with nephropathy, and V with renal failure, [167]. In Han Chinese, N9a was associated with increased risk of diabetic nephropathy [160], haplogroup T with increased the risk of coronary artery disease [168], T [169] and B4c [162] with obesity, and K with ischemic stroke [170], while N9a is protective of myocardial infarction in Japanese [171].

Haplogroups J and U are associated with predisposition to age-related macular degeneration [172] while the European haplogroup H mtDNAs correlate with reduced macular degeneration risk [173]. Haplogroups J and T are protective of osteoarthritis in the Spanish [174], and T and U have been associated with reduced sperm motility [175,176]. In athletes, haplogroups J and Uk are enriched in sprinters, haplogroup I is enriched in long distance runners in the Finnish [177], while haplogroup L0 is enriched in Kenyan elite athletes [178]. Haplogroup J has been correlated with increased longevity in Europeans [179–182] and D with increased longevity in Asians [183]. The longevity-associated Asian haplogroup D4b2 harbors a homoplasmic polymorphism in the 12S rRNA-embedded MOTS-c peptide sequence (K14Q) [184], and the normal MOTS-c peptide has been shown to act systemically to suppress inflammation [185].

Differences in mtDNA content have also been correlated with metabolic disease. These include diabetes [186,187], cardiometabolic phenotypes (lipids, glycemic and inflammatory traits, blood pressure) [188], and ischemic stroke [189].

Evidence for a mitochondrial etiology of AD and PD continues to accumulate. Numerous publications have reported systemic mitochondrial complex IV defects in AD [190–193] and complex I defects in PD [194–196]. Direct evidence that mtDNA variants can contribute to AD and PD came with the discovery that an mtDNA tRNA^{Gln} nt 4336A>G mutation is enriched in AD (3.2%), PD (5.3%), and AD/PD (6.8%) patients versus controls (0.4%). This variant arose in Europe about 8500 to 17,000 years ago and is associated with mtDNA haplogroup H5a [197], and its AD/PD association has been confirmed multiple times [151,198,199]. Since the tRNA^{Gln} nt 4336A>G mutation would partially impair mitochondrial protein synthesis, it would preferentially impair the synthesis of the mtDNA complex IV

(COI-III) and complex I (*ND1-6*) genes, thus predisposing to AD and PD. The mtDNA haplogroup K (Uk) also imparts increased risk of AD [200], and haplogroup Uk containing 143B(TK⁻) cybrids have reduced COX activity [127]. Finally, mitochondrial dysfunction has been verified in cultured cells from AD and Down syndrome with dementia patients [201].

Predisposing mtDNA variants can then be exacerbated by the age-related accumulation of developmental and somatic mutations. Somatic mtDNA mutations have been shown to be increased in AD brains and Down syndrome brains with dementia and to be associated with altered mtDNA transcript levels and mtDNA copy number [72,202–207].

A strong case can also be made for a mitochondrial etiology of PD. A subset of PD patients shows autosomal inheritance, and the cloning of these associated nDNA genes has revealed that most, if not all, are involved in protection of mitochondrial function through mitophagy, mitochondrial inner membrane integrity, mitochondrial redox control, and mitigation of oxidative damage [205,208]. Haplogroup H has been associated with increased risk, and haplogroups J and Uk with decreased risk, for developing PD [209–211], and somatic mtDNA mutations have been found increased in PD brains [205] and in neurons of PD patients [74,75].

Like PD, amyotrophic lateral sclerosis (ALS) is genetically complex. ALS also shows strong indications of underlying mitochondrial dysfunction [212] with ALS fibroblasts showing an overall dysregulated bioenergetics [213]. Several ALS nuclear gene mutations have been found to result in mitochondrial alterations. ALS Cu/Zn superoxide dismutase (*SOD1*) mutations result in multiple alterations in mitochondria function [214,215], and the ALS R15L dominant mutation in the *CHCHD10* gene causes ALS, frontotemporal dementia, and PD and results in defects in complex I assembly, impaired respiration, and impaired mitochondrial bioenergetics-driven proliferation [216]. The protein of the ALS-associated gene, *TDP-43*, is partially localized in the mitochondrion and binds to mitochondrial tRNAs and precursor RNAs encoded by the L-strand, thus regulating the processing of mitochondrial transcripts [217].

Mitochondrial dysfunction has repeatedly been implicated in autism spectrum disorder (ASD). A meta-analysis of publications reporting OXPHOS dysfunction and alterations in mitochondria-related metabolites has concluded that ASD children have

a spectrum of mitochondrial alterations of differing severities [218]. Extensive nDNA genomic studies of ASD patients have identified a variety of heterozygous copy number variants (CNVs) [219,220] and loss-of-function (LOF) mutations [221–223]. However, each variant accounts for a only few cases [221,223–226], and the cumulative risk currently accounted for by nDNA mutations is only about 20% [226]. LOF mutations in genes from ASD patients overlap with metabolic and cardiac abnormalities [221] as well as genes implicated in intellectual disability, attention-deficit/hyperactivity disorders, and schizophrenia [226–228]. Hence, ASD is a polygenic disorder [228] with the associated genes being involved in metabolic, neurological, and cardiovascular disorders.

An in-depth study of a well-characterized autism cohort revealed that many of the CNVs delete bioenergetics genes [229] and a number of ASD LOF gene mutations have been found to affect mitochondria-related functions [221,223,230,231]. More recently, both mtDNA haplogroups and heteroplasmic mtDNA mutations have been associated with ASD. Relative to the most common European haplogroup (H-HV), European haplogroups I, J, K, X, T, and U, and Asian-Native American haplogroups A and M are at significantly increased risk of ASD with odds ratios (ORs) ranging from 2.06 to 3.55. Since haplogroups I, J, K, X, T, and U represent about 55% of the European mtDNA lineages, mtDNA haplogroups may make a major contribution to the differential risk of ASD [232]. Moreover, ASD probands are more likely to harbor deleterious heteroplasmic mtDNA mutations than are their unaffected siblings. Nonsynonymous mutations were enriched about 1.5-fold and potentially pathogenic mutations were enriched about 2.2-fold in ASD probands, giving an OR of 2.55 [233].

There is growing evidence that psychiatric disorders also have a bioenergetics etiology [234]. Several meta-analyses and reviews have now reported that mitochondrial dysfunction is associated with neuropsychiatric disease [235–238]. A recent compilation of proteomic data from autopsy brains of neuropsychiatric patients revealed marked changes in mitochondrial and bioenergetic proteins, with 92 differentially expressed proteins in schizophrenia, 95 in bipolar disorder, and 41 in major depressive disorder, five being common to all three [239].

A role for mitochondrial dysfunction is also accumulating for acute illnesses such as infection, inflammation,

and trauma. Haplogroup H has been associated with protection against sepsis [240], haplogroup U with increased serum IgE levels [241], and haplogroup Uk with reduced AIDS progression and preservation of CD4 T lymphocyte numbers [242]. Mitochondrial dysfunction is also being linked to inflammation. Mitochondrial dysfunction prompts the release of mitochondrial damage-associated molecular patterns (DAMPs) including mtDNA, *N*-formyl peptides, and cardiolipin [243], and these activate the mitochondrially associated “inflammasome.” The inflammasome is composed of the NLRP3, ASC, and pro-caspase-1 polypeptides, and on binding of DAMPs, caspase-1 is activated; it cleaves pro-interleukin (IL)-1 β , and the activated IL-1 β activates the NF- κ B/AP-1 pathway leading to inflammation [244,245]. Mitochondrial dysfunction also alters T-cell biology. T regulatory cells are oxidative while T effector cells are glycolytic, so partial OXPHOS defects preferentially impair T regulatory cells, thus enhancing T effector activity and inflammation [246]. Mitochondrial function is also compromised by trauma as indicated by altered cellular mtDNA copy number [247].

Thus, mitochondrial biology is being implicated in an increasingly broad array of previously intractable clinical problems.

10.8 DIAGNOSIS OF MITOCHONDRIAL DISEASES

One of the major challenges of mitochondrial medicine is the accurate diagnosis of mitochondrial disease due to the extreme clinical and genetic heterogeneities. Traditional approaches for diagnosing mitochondrial diseases involved identifying patients with neuromuscular disease, muscle biopsy with histological and ultrastructure analyses, muscle OXPHOS enzyme assays, and ^{31}P MRI spectroscopy with exercise stress tests [8]. Several schemes have been proposed for diagnosing primary mitochondrial diseases. A consensus statement was reported [248–251].

With the discovery of mtDNA molecular defects [5–7], molecular diagnostics has become increasingly important. The use of massive parallel sequencing technologies is now permitting the analysis of both mtDNA and nDNA gene sequence variation in a fast and efficient way. This has significantly improved the molecular diagnosis of primary mitochondrial diseases and led to the discovery of many new mitochondrial genes and

associated mitochondrial functions [252–254]. By combining whole exome sequencing with RNA-sequencing, additional nonexonic variants can also be detected [255].

Still, more general screening approaches are required. Recently, testing for elevated Fibroblast Growth Factor 21 (FGF21) [256–258] and Growth Differentiation Factor 15 (GDF15) [259,260] have been proposed for detecting some aspects of mitochondrial disease. Moreover, approaches combining multiple OMIC data sets are proving informative. Relevant data sets include gene expression profiles, metabolomics, and methylomics, which can reveal key pathways or biomarkers for mitochondrial disorders [120,261–264].

However, we are still testing patients for phenotypes that have been seen in primary mitochondrial disease patients. To determine the true extent of mitochondrial dysfunction in both rare and common diseases will require the development of noninvasive, simple, versatile, sensitive, and accurate tests for detecting mitochondrial dysfunction in patients.

10.9 METABOLIC THERAPIES OF MITOCHONDRIAL DISEASES

As interest in the mitochondrion as an important factor in the etiology of not only primary mitochondrial diseases but also common metabolic and generative diseases has grown, there has been increasing interest in developing therapies for mitochondrial dysfunction. For many years, treatments focused on mitochondrial cofactors since many of the essential vitamins are mitochondrial coenzymes. This has resulted in the use of “mitochondrial cocktails.” A summary of common and potential therapeutic compounds was generated by Manji et al. [237] with the following classification: *scavenging ROS*: vitamin C, vitamin E, selenium, taurine, hypotaurine, α -lipoic acid, *N*-acetylcysteine, mitoquinone, Szeto–Schiller peptides, CoQ, idebenone, minocycline, diaminophenothiazines, polyphenols, despramipexole; *stimulating cellular antioxidant pathways* (e.g., *NRF2*): triterpenoids, fumarates, polyphenols, *N*-acetylcysteine, α -lipoic acid; *modulating calcium flux*: xestospongins C (IP3R antagonist), dihydropyridines (L-type calcium channel blockers); *targeting electron transport chain*: CoQ, idebenone, diaminophenothiazines, vitamin E quinones; *targeting antiapoptotic mechanisms*: rasagiline, thiazolidinediones, BI-11A7; *targeting mPTP*: olesoxime, minocycline, latrepirdine,

rasagiline, cyclosporine, NIM811; *other mechanisms*: uridine. However, there have been few definitive clinical trials proving the efficacy of these various approaches for primary mitochondrial diseases [265].

The therapeutic efficacy of mitochondrial metabolites has been greatly enhanced by combining molecules such as CoQ and vitamin E with the heterocyclic, cationic, triphenylphosphonium moiety (TPP⁺). TPP⁺ is permeable to the mitochondrial inner membrane, and its positive charge pulls the cationic compound into the negatively charged mitochondrial matrix [266,267]. These compounds have shown positive results in a variety of animal models of common diseases such as AD and ischemia–reperfusion injury [268,269].

Administration of CoQ and its derivatives has received the most attention. For LHON, the CoQ analogue idebenone (2,3-dimethoxy-5-methyl-6-(10-hydroxydecyl)-1,4-benzoquinone) has been reported to be beneficial in both open-label studies [270,271] and a double-blind trial [272]. In vitro studies have reported that idebenone can increase complex I activity [273] and act on the mtPTP [274]. The CoQ analogue EPI-743 has also been used to treat both a range of mitochondrial disease patients [275] and LHON patients [276]. In open-label studies, most mitochondrial diseases patients report beneficial effects of EPI-743 [275,276].

The phytoestrogens genistein + daidzein + equol (G + D + E), may be beneficial, particularly for LHON. These herbaceuticals preferentially bind to estrogen receptor β (ER β) but not ER α . About 20% of ER β is located in the mitochondrial matrix, and treatment of cells with estradiol doubles the mitochondrial Mn superoxide dismutase (MnSOD) activity within 60 min [277]. In cybrids harboring LHON mtDNAs, G + D + E increased the oxygen consumption rate and reduced ROS production in association with the upregulation of *SIRT1*, *PGC-1 α* , *NRF1*, and *TFAM* [278].

Supplementation with uridine or triacetyluridine may be therapeutic since mitochondrial ETC inhibition is required for de novo uridine synthesis. One of the steps in uridine synthesis is the reduction of dihydro-orotic acid to orotic acid via the mitochondrial enzyme (dihydro-orotic acid dehydrogenase), which uses mitochondrial CoQ as the electron acceptor. Similarly, the Szeto–Schiller peptides, specifically SS-31, selectively target the mitochondrial inner membrane by binding to cardiolipin and preventing cytochrome c from becoming a peroxidase [279].

One potential mitochondrial physiological target could be an increased mitochondrial copy number. Analysis of the mtDNA copy number in white blood cells of individuals harboring mild mtDNA LHON mutations but who have not yet lost their vision, designated as “carriers,” revealed that the carriers have a higher mtDNA copy number and mitochondrial quantity than affected individuals [280]. LHON patients who smoke have decreased mtDNA copy number and increased penetrance of blindness, indicating that heavy smoking may offset the beneficial effects of higher mtDNA copy in potential carriers, resulting in late-onset blindness [281,282]. Hence, increasing mtDNA copy number may be a general approach to treating mitochondrial disease.

One way of modulating mtDNA copy number could be through drugs that stimulate mitochondrial biogenesis. Multiple transcription factors provide promising drug targets including nuclear receptors such as the peroxisome proliferator-activated receptors (PPARs), the PPAR γ -coactivator-1 (PGC-1 α) biosynthetic pathways, the estrogen receptor-related (ERR) receptors, AMP kinase agonists, and SIRTUIN agonists [283–285]. An example of this class of drugs is bezafibrate [286–288]. Since it has been shown that cellular and tissue phenotypes are mediated by phase changes in the expression of classes of nuclear coded genes [120], drugs that modulate the epigenome including histone acetyl transferases (HATs) and histone deacetylases (HDACs) may also be promising [284].

Finally, modulating mitochondrial quality control could prove beneficial. Mitochondria undergo continuous cycles of fission and fusion in association with mitophagy. Stimulating mitophagy could increase the clearance of defective mitochondria [289–291].

10.10 GENETIC THERAPIES OF MITOCHONDRIAL DISEASES

Both somatic and germline therapies are now being developed for mitochondrial disease. Gene therapy treatment of somatic tissues involves both the direct complementation of nDNA mitochondrial gene mutations and various efforts to complement pathogenic mtDNA mutations. In a mouse model of *Ant1* deficiency, the nDNA null mutation was successfully complemented in muscle by injection of an adeno-associated virus (AAV)-borne *Ant1* cDNA [292]. This approach

will likely be used to treat various mitochondrial myopathies and cardiomyopathies in the years ahead.

Efforts to introduce DNA into the mitochondrion have been reported using both DNA–protein aggregates [293] and by targeting AAV to mitochondrial membranes [294]. The most actively pursued approach for genetically treating mitochondrial disease has been the allotopic expression of mtDNA genes. In this procedure, mtDNA genes are cloned and the mtDNA genetic code adjusted to be optimal for the nucleus–cytosol compartment. The mtDNA-derived cDNA is then linked to an N-terminal mitochondrial targeting peptide, nuclear gene expression elements added, and the construct introduced into the nucleus via AAV. Expression of the transgene results in the cytosolic synthesis of the mitochondrial polypeptide which is then imported into the mitochondrion by the N-terminal targeting peptide. When the polypeptide is processed, the hope is that it will be incorporated into the appropriate mitochondrial enzyme complex. This approach is being actively pursued for treating the common LHON *ND4* nt 11778A>G (R340H) mutation by injecting the virus into the eye vitreous to transduce the affected retinal ganglion cells [295–304]. A newer approach could be to have the nucleus express allotopic mRNA that is imported into the mitochondrion. Inside the mitochondrion, the imported mRNA could be translated on mitochondrial ribosomes and directly introduced into the appropriate complex [305–307]. This approach could have the advantage that the mRNA with the mitochondrial genetic code could not be translated in the cytosol, thus avoiding the accumulation of toxic cytosolic protein aggregates.

Most recently, mtDNA mutations have been eliminated through germline mitochondrial replacement therapy. Two procedures are being developed: zygote pronuclear transfer (PNT) and oocyte spindle transfer (ST). In PNT procedure the oocyte of the affected woman is fertilized and the two pronuclei are transferred via a micropipette into an enucleated recipient zygote from a woman with normal mtDNAs [308,309]. In the ST procedure, the spindle from the metaphase II oocyte is transferred into an enucleated oocyte with normal mtDNAs followed by fertilization and implantation [310–314]. The birth of one spindle transfer “three-parent baby” has been reported in the literature. This child was born to a mother who had lost two children to Leigh syndrome due to high levels of the mtDNA *ATP6*

8993T>G (L156R) mutation. One implanted ST blastocyst gave rise to an apparently normal boy carrying between 2% and 9% 8993G mutant mtDNA in his various tissues [315,316].

10.11 CONCLUSION

It is now 30 years since the first molecular causes of mitochondrial disease were reported. In the interim, an entirely new field of mitochondrial medicine has arisen, which now explains the molecular and biochemical bases of a broad new class of bioenergetic diseases. More importantly, the knowledge gained from the identification and characterization of primary mitochondrial diseases has provided new perspectives on the etiology and genetics of the common metabolic and degenerative diseases and aging. The development of therapeutic modalities for the treatment of primary mitochondrial diseases thus has the potential for application to the common diseases. Hence, the mitochondrial medicine paradigm may revolutionize the conceptual framework of Western medicine in the twenty-first century with broad implications for the health and well-being of patients.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants MH108592, NS021328, CA182384, and OD010944 and U.S. Department of Defense grant W81XWH-16-1-0400 & 0401 awarded to D.C.W. and supported by grants from: “Fondation pour la Recherche Médicale” DPM20121125554, and AFM Mitoscreen 17122 awarded to V.P.

REFERENCES

- [1] Picard M, Wallace DC, Burelle Y. The rise of mitochondria in medicine. *Mitochondrion* 2016;30:105–16.
- [2] Luft R, Ikkos D, Palmieri G, Ernster L, Afzelius BA. A case of severe hypermetabolism of nonthyroid origin with a defect in the maintenance of mitochondrial respiratory control: a correlated clinical, biochemical, and morphological study. *J Clin Invest* 1962;41:1776–804.
- [3] DiMauro S. Mitochondrial encephalomyopathies. In: Rosenberg, Prusiner SB, DiMauro S, Barchi RL, Kunkel LM, editors. *The Molecular and Genetic Basis of Neurological Disease*. Stoneham (MA): Butterworth-Heinemann; 1993. p. 665–94.
- [4] Morgan-Hughes JA, Hanna MG. Mitochondrial encephalomyopathies: the enigma of genotype versus phenotype. *Biochim Biophys Acta* 1999;1410:125–45.
- [5] Holt IJ, Harding AE, Morgan-Hughes JA. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* 1988;331:717–9.
- [6] Shoffner JM, Lott MT, Lezza AM, Seibel P, Ballinger SW, Wallace DC. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA^{Lys} mutation. *Cell* 1990;61:931–7.
- [7] Wallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AM, Elsas LJ, Nikoskelainen EK. Mitochondrial DNA mutation associated with Leber’s hereditary optic neuropathy. *Science* 1988a;242:1427–30.
- [8] Wallace DC, Zheng X, Lott MT, Shoffner JM, Hodge JA, Kelley RI, Epstein CM, Hopkins LC. Familial mitochondrial encephalomyopathy (MERRF): Genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease. *Cell* 1988b;55:601–10.
- [9] Wallace DC. Mitochondrial bioenergetic etiology of disease. *J Clin Invest* 2013;123:1405–12.
- [10] Chinnery PF, Johnson MA, Wardell TM, Singh-Kler R, Hayes C, Brown DT, Taylor RW, Bindoff LA, Turnbull DM. The epidemiology of pathogenic mitochondrial DNA mutations. *Ann Neurol* 2000;48:188–93.
- [11] Cree LM, Samuels DC, Chinnery PF. The inheritance of pathogenic mitochondrial DNA mutations. *Biochim Biophys Acta* 2009;1792:1097–102.
- [12] Gorman GS, Schaefer AM, Ng Y, Gomez N, Blakely EL, Alston CL, Feeney C, Horvath R, Yu-Wai-Man P, Chinnery PF, Taylor RW, Turnbull DM, McFarland R. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann Neurol* 2015;77:753–9.
- [13] Schaefer AM, McFarland R, Blakely EL, He L, Whitaker RG, Taylor RW, Chinnery PF, Turnbull DM. Prevalence of mitochondrial DNA disease in adults. *Ann Neurol* 2008;63:35–9.
- [14] Elliott HR, Samuels DC, Eden JA, Relton CL, Chinnery PF. Pathogenic mitochondrial DNA mutations are common in the general population. *Am J Hum Genet* 2008;83:254–60.
- [15] Wallace DC, Lott MT, Procaccio V. Mitochondrial Medicine: The Mitochondrial Biology and Genetics of Metabolic and Degenerative Diseases, Cancer, and Aging (Chapter 11). In: Rimoin DL, Pyeritz RE, Korf BR, editors. *Emery and Rimoin’s Principles and Practice of Medical Genetics*. 6th ed. Philadelphia: Churchill Livingstone Elsevier; 2013. p. 1–153.
- [16] MITOMAP. A Human Mitochondrial Genome Database. 2018. <http://www.mitomap.org>.

- [17] Pham TD, Pham PQ, Li J, Letai AG, Wallace DC, Burke PJ. Cristae remodeling causes acidification detected by integrated graphene sensor during mitochondrial outer membrane permeabilization. *Sci Rep* 2016;6:35907.
- [18] Mayr JA. Lipid metabolism in mitochondrial membranes. *J Inherit Metab Dis* 2015;38:137–44.
- [19] Bricker DK, Taylor EB, Schell JC, Orsak T, Boutron A, Chen YC, Cox JE, Cardon CM, Van Vranken JG, Dephourse N, Redin C, Boudina S, Gygi SP, Brivet M, Thummel CS, Rutter J. A mitochondrial pyruvate carrier required for pyruvate uptake in yeast, *Drosophila*, and humans. *Science* 2012;337:96–100.
- [20] Wellen KE, Hatzivassiliou G, Sachdeva UM, Bui TV, Cross JR, Thompson CB. ATP-citrate lyase links cellular metabolism to histone acetylation. *Science* 2009;324:1076–80.
- [21] Gururaja Rao S. Mitochondrial changes in cancer. *Handb Exp Pharmacol* 2017;240:211–27.
- [22] Wallace DC. Mitochondria and cancer. *Nat Rev Cancer* 2012;12:685–98.
- [23] Denton RM. Regulation of mitochondrial dehydrogenases by calcium ions. *Biochim Biophys Acta* 2009;1787:1309–16.
- [24] Palmieri F. The mitochondrial transporter family SLC25: identification, properties and physiopathology. *Mol Aspects Med* 2013;34:465–84.
- [25] Palmieri F. Mitochondrial transporters of the SLC25 family and associated diseases: a review. *J Inherit Metab Dis* 2014;37:565–75.
- [26] Palmieri F, Monne M. Discoveries, metabolic roles and diseases of mitochondrial carriers: A review. *Biochim Biophys Acta* 2016;1863:2362–78.
- [27] Valsecchi F, Ramos-Espíritu LS, Buck J, Levin LR, Manfredi G. cAMP and mitochondria. *Physiology (Bethesda)* 2013;28:199–209.
- [28] Ghafourifar P, Cadenas E. Mitochondrial nitric oxide synthase. *Trends Pharm Sci* 2005;26:190–5.
- [29] Wirth C, Brandt U, Hunte C, Zickermann V. Structure and function of mitochondrial complex I. *Biochim Biophys Acta* 2016;1857:902–14.
- [30] Zhu J, Vinothkumar KR, Hirst J. Structure of mammalian respiratory complex I. *Nature* 2016;536:354–8.
- [31] Iwata S, Lee JW, Okada K, Lee JK, Iwata M, Rasmussen B, Link TA, Ramaswamy S, Jap BK. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc₁ complex. *Science* 1998;281:64–71.
- [32] Solmaz SR, Hunte C. Structure of complex III with bound cytochrome c in reduced state and definition of a minimal core interface for electron transfer. *J Biol Chem* 2008;283:17542–9.
- [33] Saraste M. Oxidative phosphorylation at the fin de siècle. *Science* 1999;283:1488–93.
- [34] Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science* 1996;272:1136–44.
- [35] Yoshikawa S, Shimada A. Reaction mechanism of cytochrome c oxidase. *Chem Rev* 2015;115:1936–89.
- [36] Letts JA, Sazanov LA. Clarifying the supercomplex: the higher-order organization of the mitochondrial electron transport chain. *Nat Struct Mol Biol* 2017;24:800–8.
- [37] Schagger H. Blue-Native Gels to Isolate Protein Complexes from Mitochondria. In: Pon LA, Schon EA, editors. *Mitochondria*. San Diego (CA): Academic Press; 2001. p. 231–44.
- [38] Schagger H, Pfeiffer K. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *EMBO J* 2000;19:1777–83.
- [39] Wu M, Gu J, Guo R, Huang Y, Yang M. Structure of mammalian respiratory supercomplex I₁III₂IV₁. *Cell* 2016;167:1598–609. e1510.
- [40] Walker JE. The ATP synthase: the understood, the uncertain and the unknown. *Biochem Soc Trans* 2013;41:1–16.
- [41] He J, Ford HC, Carroll J, Ding S, Fearnley IM, Walker JE. Persistence of the mitochondrial permeability transition in the absence of subunit c of human ATP synthase. *Proc Natl Acad Sci USA* 2017;114:3409–14.
- [42] Martin JL, Ishmukhametov R, Hornung T, Ahmad Z, Frasch WD. Anatomy of F₁-ATPase powered rotation. *Proc Natl Acad Sci USA* 2014;111:3715–20.
- [43] Kaukonen J, Juselius JK, Tiranti V, Kyttälä A, Zeviani M, Comi GP, Keranen S, Peltonen L, Suomalainen A. Role of adenine nucleotide translocator 1 in mtDNA maintenance. *Science* 2000;289:782–5.
- [44] Palmieri L, Alberio S, Pisano I, Lodi T, Meznaric-Petrusa M, Zidar J, Santoro A, Scarcia P, Fontanesi F, Lamantea E, Ferrero I, Zeviani M. Complete loss-of-function of the heart/muscle-specific adenine nucleotide translocator is associated with mitochondrial myopathy and cardiomyopathy. *Hum Mol Genet* 2005;14:3079–88.
- [45] Strauss KA, Dubiner L, Simon M, Zaragoza M, Sengupta PP, Li P, Narula N, Dreike S, Platt J, Procaccio V, Ortiz-Gonzalez XR, Puffenberger EG, Kelley RI, Morton DH, Narula J, Wallace DC. Severity of cardiomyopathy associated with adenine nucleotide translocator-1 deficiency correlates with mtDNA haplogroup. *Proc Natl Acad Sci USA* 2013;110:3253–458.

- [46] Baughman JM, Perocchi F, Girgis HS, Plovanich M, Belcher-Timme CA, Sancak Y, Bao XR, Strittmatter L, Goldberger O, Bogorad RL, Koteliensky V, Mootha VK. Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* 2011;476:341–5.
- [47] De Stefani D, Raffaello A, Teardo E, Szabo I, Rizzuto R. A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* 2011;476:336–40.
- [48] Bhola PD, Letai A. Mitochondria - judges and executioners of cell death sentences. *Mol Cell* 2016;61:695–704.
- [49] Sarosiek KA, Ni Chonghaile T, Letai A. Mitochondria: gatekeepers of response to chemotherapy. *Trends Cell Biol* 2013;23:612–9.
- [50] Alavian KN, Beutner G, Lazrove E, Sacchetti S, Park HA, Licznernski P, Li H, Nabili P, Hockensmith K, Graham M, Porter Jr GA, Jonas EA. An uncoupling channel within the c-subunit ring of the F1FO ATP synthase is the mitochondrial permeability transition pore. *Proc Natl Acad Sci USA* 2014;111:10580–5.
- [51] Bernardi P, Rasola A, Forte M, Lippe G. The mitochondrial permeability transition pore: channel formation by F-ATP synthase, integration in signal transduction, and role in pathophysiology. *Physiol Rev* 2015;95:1111–55.
- [52] Giorgio V, von Stockum S, Antoniel M, Fabbro A, Fogolari F, Forte M, Glick GD, Petronilli V, Zoratti M, Szabo I, Lippe G, Bernardi P. Dimers of mitochondrial ATP synthase form the permeability transition pore. *Proc Natl Acad Sci USA* 2013;110:5887–92.
- [53] Jonas EA, Porter Jr GA, Beutner G, Mnatsakanyan N, Alavian KN. Cell death disguised: the mitochondrial permeability transition pore as the c-subunit of the F(1)F(O) ATP synthase. *Pharmacol Res* 2015;99:382–92.
- [54] Teixeira FK, Sanchez CG, Hurd TR, Seifert JR, Czech B, Preall JB, Hannon GJ, Lehmann R. ATP synthase promotes germ cell differentiation independent of oxidative phosphorylation. *Nat Cell Biol* 2015;17:689–96.
- [55] Shanmughapriya S, Rajan S, Hoffman NE, Higgins AM, Tomar D, Nemani N, Hines KJ, Smith DJ, Eguchi A, Vallem S, Shaikh F, Cheung M, Leonard NJ, Stalakis RS, Wolfers MP, Ibetti J, Chuprun JK, Jog NR, Houser SR, Koch WJ, Elrod JW, Madesh M. SPG7 Is an essential and conserved component of the mitochondrial permeability transition pore. *Mol Cell* 2015;60:47–62.
- [56] van der Bliek AM, Shen Q, Kawajiri S. Mechanisms of mitochondrial fission and fusion. *Cold Spring Harbor. Perspect Biol* 2013;5.
- [57] Narendra D, Walker JE, Youle R. Mitochondrial quality control mediated by PINK1 and Parkin: links to parkinsonism. *Cold Spring Harbor Perspect Biol* 2012;4:a011338.
- [58] Wallace DC, Bunn CL, Eisenstadt JM. Cytoplasmic transfer of chloramphenicol resistance in human tissue culture cells. *Journal of Cell Biology* 1975;67:174–88.
- [59] Shuster RC, Rubenstein AJ, Wallace DC. Mitochondrial DNA in anucleate human blood cells. *Biochem Biophys Res Commun* 1988;155:1360–5.
- [60] Wallace DC. Cytoplasmic inheritance of chloramphenicol resistance in mammalian cells. Chapter 12. In: Shay JW, editor. *Techniques in Somatic Cell Genetics*. New York: Plenum Press; 1982. p. 159–87.
- [61] Wallace DC, Bunn CL, Eisenstadt JM. Mitotic segregation of cytoplasmic inherited genes for chloramphenicol resistance in mammalian cells. II: Fusions with human cell lines. *Somatic Cell Genet* 1977;3:93–119.
- [62] Wallace DC, Eisenstadt JM. The expression of cytoplasmically inherited genes for chloramphenicol resistance in interspecific somatic cell hybrids and cybrids. *Somatic Cell Genet* 1979;5. 573–396.
- [63] Wallace DC, Pollack Y, Bunn CL, Eisenstadt JM. Cytoplasmic inheritance in mammalian tissue culture cells. *In Vitro* 1976;12:758–76.
- [64] Case JT, Wallace DC. Maternal inheritance of mitochondrial DNA polymorphisms in cultured human fibroblasts. *Somatic Cell Genet* 1981;7:103–8.
- [65] Giles RE, Blanc H, Cann HM, Wallace DC. Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci USA* 1980;77:6715–9.
- [66] Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, Wan J, Kim SJ, Mehta H, Hevener AL, de Cabo R, Cohen P. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab* 2015;21:443–54.
- [67] Yen K, Lee C, Mehta H, Cohen P. The emerging role of the mitochondrial-derived peptide humanin in stress resistance. *J Mol Endocrinol* 2013;50:R11–9.
- [68] Brown WM, Prager EM, Wan A, Wilson AC. Mitochondrial DNA sequences in primates: tempo and mode of evolution. *J Mol Evol* 1982;18:225–39.
- [69] Neckelmann N, Li K, Wade RP, Shuster R, Wallace DC. cDNA sequence of a human skeletal muscle ADP/ATP translocator: lack of a leader peptide, divergence from a fibroblast translocator cDNA, and coevolution with mitochondrial DNA genes. *Proc Natl Acad Sci USA* 1987;84:7580–4.
- [70] Wallace DC, Ye JH, Neckelmann SN, Singh G, Webster KA, Greenberg BD. Sequence analysis of cDNAs for the human and bovine ATP synthase b-subunit: mitochondrial DNA genes sustain seventeen times more mutations. *Curr Genet* 1987;12:81–90.

- [71] Corral-Debrinski M, Horton T, Lott MT, Shoffner JM, Beal MF, Wallace DC. Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age. *Nat Genet* 1992a;2:324–9.
- [72] Coskun PE, Wyrembak J, Derbereva O, Melkonian G, Doran E, Lott IT, Head E, Cotman CW, Wallace DC. Systemic mitochondrial dysfunction and the etiology of Alzheimer's disease and down syndrome dementia. *J Alzheimers Dis* 2010;20(Suppl. 2):S293–310.
- [73] Muller-Hocker J, Seibel P, Schneiderbanger K, Kadenbach B. Different in situ hybridization patterns of mitochondrial DNA in cytochrome c oxidase-deficient extraocular muscle fibres in the elderly. *Virchows Arch A Pathol Anat Histopathol* 1993;422:7–15.
- [74] Bender A, Krishnan KJ, Morris CM, Taylor GA, Reeve AK, Perry RH, Jaros E, Hersheson JS, Betts J, Klopstock T, Taylor RW, Turnbull DM. High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. *Nat Genet* 2006;38:515–7.
- [75] Kraytsberg Y, Kudryavtseva E, McKee AC, Geula C, Kowall NW, Khrapko K. Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nature Genet* 2006;38:518–20.
- [76] Khrapko K, Bodyak N, Thilly WG, van Orsouw NJ, Zhang X, Collier HA, Perls TT, Upton M, Vijg J, Wei JY. Cell-by-cell scanning of whole mitochondrial genomes in aged human heart reveals a significant fraction of myocytes with clonally expanded deletions. *Nucleic Acids Res* 1999;27:2434–41.
- [77] Wallace DC, Chalkia D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor. Perspect Biol* 2013;5:a021220.
- [78] Fan W, Waymire K, Narula N, Li P, Rocher C, Coskun PE, Vannan MA, Narula J, MacGregor GR, Wallace DC. A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science* 2008;319:958–62.
- [79] Sharples MS, Marciniak C, Eckel-Mahan K, McManus MJ, Crimi M, Waymire K, Lin CS, Masubuchi S, Friend N, Koike M, Chalkia D, MacGregor GR, Sassone-Corsi P, Wallace DC. Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* 2012;151:333–43.
- [80] Stewart JB, Freyer C, Elson JL, Wredenberg A, Cansu Z, Trifunovic A, Larsson NG. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol* 2008;6:e10.
- [81] Wallace DC. Mitochondrial DNA variation in human radiation and disease. *Cell* 2015;163:33–8.
- [82] Calvo SE, Mootha VK. The mitochondrial proteome and human disease. *Annu Rev Genom Hum Genet* 2010;11:25–44.
- [83] Stojanovski D, Bohnert M, Pfanner N, van der Laan M. Mechanisms of protein sorting in mitochondria. *Cold Spring Harbor. Perspect Biol* 2012;4:a011320.
- [84] Li K, Warner CK, Hodge JA, Minoshima S, Kudoh J, Fukuyama R, Maekawa M, Shimizu Y, Shimizu N, Wallace DC. A human muscle adenine nucleotide translocator gene has four exons, is located on chromosome 4, and is differentially expressed. *J Biol Chem* 1989;264:13998–4004.
- [85] Neckelmann N, Warner CK, Chung A, Kudoh J, Minoshima S, Fukuyama R, Maekawa M, Shimizu Y, Shimizu N, Liu JD, Wallace DC. The human ATP synthase beta subunit gene: sequence analysis, chromosome assignment, and differential expression. *Genomics* 1989;5:829–43.
- [86] Procaccio V, Mousson B, Beugnot R, Duborjal H, Feillet F, Putet G, Pignot-Paintrand I, Lombes A, De Coe R, Smeets H, Lunardi J, Issartel JP. Nuclear DNA origin of mitochondrial complex I deficiency in fatal infantile lactic acidosis evidenced by transnuclear complementation of cultured fibroblasts. *J Clin Investig* 1999;104:83–92.
- [87] Koopman WJ, Willems PH, Smeitink JA. Monogenic mitochondrial disorders. *N Engl J Med* 2012;366:1132–41.
- [88] Potluri P, Davila A, Ruiz-Pesini E, Mishmar D, O'Hearn S, Hancock S, Simon MC, Scheffler I, Wallace DC, Procaccio V. A novel NDUFA1 mutation leads to a progressive mitochondrial complex I-specific neurodegenerative disease. *Mol Genet Metab* 2009;96:189–95.
- [89] Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli Sforza LL, Wallace DC. Ethnic variation in Hpa I endonuclease cleavage patterns of human mitochondrial DNA. *Proc Natl Acad Sci USA* 1981;78:5768–72.
- [90] Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. *Gene* 1999;238:211–30.
- [91] Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL. Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *J Mol Evol* 1983;19:255–71.
- [92] Wallace DC. Mitochondrial DNA sequence variation in human evolution and disease. *Proc Natl Acad Sci USA* 1994;91:8739–46.
- [93] Wallace DC. 1994 William Allan Award Address. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *Am J Hum Genet* 1995;57:201–23.

- [94] Mishmar D, Ruiz-Pesini EE, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC. Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 2003;100:171–6.
- [95] Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 2004;303:223–6.
- [96] Ruiz-Pesini E, Wallace DC. Evidence for adaptive selection acting on the tRNA and rRNA genes of the human mitochondrial DNA. *Human Mutat* 2006;27:1072–81.
- [97] Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. *Nature* 1987;325:31–6.
- [98] Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, Sherry ST, Wallace DC. The structure of human mitochondrial DNA variation. *J Mol Evol* 1991;33:543–55.
- [99] Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet* 1995;57:133–49.
- [100] Schurr TG, Donham BP, Morreale SC, Panter-Brick C, Donham DL, Armelagos GJ, Wallace DC. Genetic diversity in modern African populations and its use for reconstructing ancient and modern population movements. In: Reed DM, editor. *Biomolecular Archeology, Genetic Approaches to the Past. Occasional Paper #31*. Carbondale: Center for Archaeological Investigations, Southern Illinois University; 2005. p. 169–207.
- [101] Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC. Classification of European mtDNAs from an analysis of three European populations. *Genetics* 1996;144:1835–50.
- [102] Schurr TG, Ballinger SW, Gan YY, Hodge JA, Merriwether DA, Lawrence DN, Knowler WC, Weiss KM, Wallace DC. Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. *Am J Hum Genet* 1990;46:613–23.
- [103] Brown MD, Hosseini SH, Torroni A, Bandelt HJ, Allen JC, Schurr TG, Scozzari R, Cruciani F, Wallace DC. mtDNA Haplogroup X: an ancient link between Europe/Western Asia and North America? *Am J Hum Genet* 1998;63:1852–61.
- [104] Kazuno AA, Munakata K, Nagai T, Shimozone S, Tanaka M, Yoneda M, Kato N, Miyawaki A, Kato T. Identification of mitochondrial DNA polymorphisms that alter mitochondrial matrix pH and intracellular calcium dynamics. *PLoS Genet* 2006;2:e128.
- [105] Wallace DC, Lott MT. Leber Hereditary Optic Neuropathy: exemplar of an mtDNA disease. *Handb Exp Pharmacol* 2017;240:339–76.
- [106] Holt IJ, Harding AE, Petty RK, Morgan-Hughes JA. A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *Am J Hum Genet* 1990;46:428–33.
- [107] Sadun AA, La Morgia C, Carelli V. Leber's Hereditary Optic Neuropathy. *Curr Treat Options Neurol* 2011;13:109–17.
- [108] Huoponen K, Vilkkij J, Aula P, Nikoskelainen EK, Savontaus ML. A new mtDNA mutation associated with Leber hereditary optic neuroretinopathy. *Am J Hum Genet* 1991;48:1147–53.
- [109] Johns DR, Neufeld MJ, Park RD. An ND-6 mitochondrial DNA mutation associated with Leber hereditary optic neuropathy. *Biochem Biophys Res Commun* 1992;187:1551–7.
- [110] Jun AS, Brown MD, Wallace DC. A mitochondrial DNA mutation at np 14459 of the ND6 gene associated with maternally inherited Leber's hereditary optic neuropathy and dystonia. *Proc Natl Acad Sci USA* 1994;91:6206–10.
- [111] Malfatti E, Bugiani M, Invernizzi F, de Souza CF, Farina L, Carrara F, Lamantea E, Antozzi C, Confalonieri P, Sanseverino MT, Giugliani R, Uziel G, Zeviani M. Novel mutations of ND genes in complex I deficiency associated with mitochondrial encephalopathy. *Brain* 2007;130:1894–904.
- [112] De Vries DD, Van Engelen BG, Gabreels FJ, Ruitenbeek W, Van Oost BA. A second missense mutation in the mitochondrial ATPase 6 gene in Leigh's syndrome. *Ann Neurol* 1993;34:410–2.
- [113] Santorelli FM, Shanske S, Jain KD, Tick D, Schon EA, DiMauro S. A T-C mutation at nt 8993 of mitochondrial DNA in a child with Leigh syndrome. *Neurology* 1994;44:972–4.
- [114] Trounce I, Neill S, Wallace DC. Cytoplasmic transfer of the mtDNA nt 8993 TG (ATP6) point mutation associated with Leigh syndrome into mtDNA-less cells demonstrates cosegregation with a decrease in state III respiration and ADP/O ratio. *Proc Natl Acad Sci USA* 1994;91:8334–8.
- [115] Ortiz RG, Newman NJ, Shoffner JM, Kaufman AE, Koontz DA, Wallace DC. Variable retinal and neurologic manifestations in patients harboring the mitochondrial DNA 8993 mutation. *Arch Ophthalmol* 1993;111:1525–30.
- [116] Santorelli FM, Shanske S, Macaya A, DeVivo DC, DiMauro S. The mutation at nt 8993 of mitochondrial DNA is a common cause of Leigh's syndrome. *Ann Neurol* 1993;34:827–34.

- [117] Tatuch Y, Christodoulou J, Feigenbaum A, Clarke JTR, Wherret J, Smith C, Rudd N, Petrova-Benedict R, Robinson BH. Heteroplasmic mtDNA mutation (T-G) at 8993 can cause Leigh disease when the percentage of abnormal mtDNA is high. *Am J Hum Genet* 1992;50:852–8.
- [118] Goto Y, Nonaka I, Horai S. A mutation in the tRNA^{Leu(UUR)} gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature* 1990;348:651–3.
- [119] Desquiret-Dumas V, Gueguen N, Barth M, Chevrollier A, Hancock S, Wallace DC, Amati-Bonneau P, Henrion D, Bonneau D, Reynier P, Procaccio V. Metabolically induced heteroplasmy shifting and l-arginine treatment reduce the energetic defect in a neuronal-like model of MELAS. *Biochim Biophys Acta* 2012;1822:1019–29.
- [120] Picard M, Zhang J, Hancock S, Derbeneva O, Golhar R, Golik P, O'Hearn S, Levy S, Potluri P, Lvova M, Davila A, Lin CS, Perin JC, Rappaport EF, Hakonarson H, Trounce IA, Procaccio V, Wallace DC. Progressive increase in mtDNA 3243A>G heteroplasmy causes abrupt transcriptional reprogramming. *Proc Natl Acad Sci USA* 2014;111:E4033–42.
- [121] Wang Q, Ito M, Adams K, Li BU, Klopstock T, Maslim A, Higashimoto T, Herzog J, Boles RG. Mitochondrial DNA control region sequence variation in migraine headache and cyclic vomiting syndrome. *Am J Med Genet* 2004;131A:50–8.
- [122] Brown MD, Starikovskaya E, Derbeneva O, Hosseini S, Allen JC, Mikhailovskaya IE, Sukernik RI, Wallace DC. The role of mtDNA background in disease expression: A new primary LHON mutation associated with Western Eurasian haplogroup J. *Hum Genet* 2002;110:130–8.
- [123] Brown MD, Sun F, Wallace DC. Clustering of Caucasian Leber hereditary optic neuropathy patients containing the 11778 or 14484 mutations on an mtDNA lineage. *Am J Hum Genet* 1997;60:381–7.
- [124] Brown MD, Torroni A, Reckord CL, Wallace DC. Phylogenetic analysis of Leber's hereditary optic neuropathy mitochondrial DNAs indicates multiple independent occurrences of the common mutations. *Hum Mutat* 1995;6:311–25.
- [125] Torroni A, Petrozzi M, D'Urbano L, Sellitto D, Zeviani M, Carrara F, Carducci C, Leuzzi V, Carelli V, Barboni P, De Negri A, Scozzari R. Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. *Am J Hum Genet* 1997;60:1107–21.
- [126] Carelli V, Achilli A, Valentino ML, Rengo C, Semino O, Pala M, Olivieri A, Mattiazzi M, Pallotti F, Carrara F, Zeviani M, Leuzzi V, Carducci C, Valle G, Simionati B, Mendieta L, Salomao S, Belfort R, Sadun AA, Torroni A. Haplogroup effects and recombination of mitochondrial DNA: novel clues from the analysis of Leber Hereditary Optic Neuropathy pedigrees. *Am J Hum Genet* 2006;78:564–74.
- [127] Gomez-Duran A, Pacheu-Grau D, Lopez-Gallardo E, Diez-Sanchez C, Montoya J, Lopez-Perez MJ, Ruiz-Pesini E. Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. *Hum Mol Genet* 2010;19:3343–53.
- [128] Vergani L, Martinuzzi A, Carelli V, Cortelli P, Montagna P, Schievano G, Carrozzo R, Angelini C, Lugaesi E. MtDNA mutations associated with Leber's hereditary optic neuropathy: studies on cytoplasmic hybrid (cybrid) cells. *Biochem Biophys Res Commun* 1995;210:880–8.
- [129] Marom S, Friger M, Mishmar D. MtDNA meta-analysis reveals both phenotype specificity and allele heterogeneity: a model for differential association. *Sci Rep* 2017;7:43449.
- [130] Zeviani M, Moraes CT, DiMauro S, Nakase H, Bonilla E, Nakase H, Bonilla E, Schon EA, Rowland LP. Deletions of mitochondrial DNA in Kearns-Sayre syndrome. *Neurology* 1988;38:1339–46.
- [131] Shoffner JM, Lott MT, Voljavec AS, Soueidan SA, Costigan DA, Wallace DC. Spontaneous Kearns-Sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial DNA deletion: a slip-replication model and metabolic therapy. *Proc Natl Acad Sci USA* 1989;86:7952–6.
- [132] Rotig A, Colonna M, Blanche S, Fischer A, LeDeist F, Frezal J, Saudubray JM, Munnich A. Deletion of blood mitochondrial DNA in pancytopenia. *Lancet* 1988;2:567–8.
- [133] Rotig A, Colonna M, Bonnefont JP, Blanche S, Fischer A, Saudubray JM, Munnich A. Mitochondrial DNA deletion in Pearson's marrow-pancreas syndrome. *Lancet* 1989;1:902–3.
- [134] Ballinger SW, Shoffner JM, Gebhart S, Koontz DA, Wallace DC. Mitochondrial diabetes revisited. *Nat Genet* 1994;7:458–9.
- [135] Kujoth GC, Hiona A, Pugh TD, Someya S, Panzer K, Wohlgemuth SE, Hofer T, Seo AY, Sullivan R, Jobling WA, Morrow JD, Van Remmen H, Sedivy JM, Yamasoba T, Tanokura M, Weindrich R, Leeuwenburgh C, Prolla TA. Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* 2005;309:481–4.

- [136] Schriener SE, Linford NJ, Martin GM, Treuting P, Ogburn CE, Emond M, Coskun PE, Ladiges W, Wolf N, Van Remmen H, Wallace DC, Rabinovitch PS. Extension of murine life span by overexpression of catalase targeted to mitochondria. *Science* 2005;308:1909–11.
- [137] Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, Bruder CE, Bohlooly YM, Gidlof S, Oldfors A, Wibom R, Tornell J, Jacobs HT, Larsson NG. Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* 2004;429:417–23.
- [138] Gerber S, Ding MG, Gerard X, Zwicker K, Zanolighi X, Rio M, Serre V, Hanein S, Munnich A, Rotig A, Bianchi L, Amati-Bonneau P, Elpeleg O, Kaplan J, Brandt U, Rozet JM. Compound heterozygosity for severe and hypomorphic NDUF52 mutations cause non-syndromic LHON-like optic neuropathy. *J Med Genet* 2017;54:346–56.
- [139] Antonicka H, Choquet K, Lin ZY, Gingras AC, Kleinman CL, Shoubridge EA. A pseudouridine synthase module is essential for mitochondrial protein synthesis and cell viability. *EMBO Rep* 2017;18:28–38.
- [140] Simon M, Richard EM, Wang X, Shahzad M, Huang VH, Qaiser TA, Potluri P, Mahl SE, Davila A, Nazli S, Hancock S, Yu M, Gargus J, Chang R, Al-Sheqaih N, Newman WG, Abdenur J, Starr A, Hegde R, Dorn T, Busch A, Park E, Wu J, Schwenzer H, Flierl A, Florentz C, Sissler M, Khan SN, Li R, Guan MX, Friedman TB, Wu DK, Procaccio V, Riazuddin S, Wallace DC, Ahmed ZM, Huang T, Riazuddin S. Mutations of human NARS2, encoding the mitochondrial asparaginyl-tRNA synthetase, cause nonsyndromic Deafness and Leigh Syndrome. *PLoS Genet* 2015;11:e1005097.
- [141] Meng F, Cang X, Peng Y, Li R, Zhang Z, Li F, Fan Q, Guan AS, Fischel-Ghosian N, Zhao X, Guan MX. Biochemical evidence for a nuclear modifier allele (A10S) in TRMU (Methylaminomethyl-2-thiouridylate-methyltransferase) related to mitochondrial tRNA modification in the phenotypic manifestation of deafness-associated 12S rRNA mutation. *J Biol Chem* 2017;292:2881–92.
- [142] Copeland WC. The mitochondrial DNA polymerase in health and disease. *Sub Cell Biochem* 2010;50:211–22.
- [143] Naviaux RK, Nguyen KV. POLG mutations associated with Alpers' syndrome and mitochondrial DNA depletion. *Ann Neurol* 2004;55:706–12.
- [144] Van Goethem G, Dermaut B, Lofgren A, Martin JJ, Van Broeckhoven C. Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nature Genet* 2001;28:211–2.
- [145] Spelbrink JN, Li FY, Tiranti V, Nikali K, Yuan QP, Tariq M, Wanrooij S, Garrido N, Comi G, Morandi L, Santoro L, Toscano A, Fabrizi GM, Somer H, Croxen R, Beeson D, Poulton J, Suomalainen A, Jacobs HT, Zeviani M, Larsson C. Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nature Genet* 2001;28:223–31.
- [146] Mandel H, Szargel R, Labay V, Elpeleg O, Saada A, Shalata A, Anbinder Y, Berkowitz D, Hartman C, Barak M, Eriksson S, Cohen N. The deoxyguanosine kinase gene is mutated in individuals with depleted hepatocerebral mitochondrial DNA. *Nature Genet* 2001;29:337–41.
- [147] Saada A, Shaag A, Mandel H, Nevo Y, Eriksson S, Elpeleg O. Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy. *Nature Genet* 2001;29:342–4.
- [148] Nishino I, Spinazzola A, Hirano M. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. *Science* 1999;283:689–92.
- [149] El-Hattab AW, Craigen WJ, Scaglia F. Mitochondrial DNA maintenance defects. *Biochim Biophys Acta* 2017;1863:1539–55.
- [150] Kwak SH, Park KS, Lee KU, Lee HK. Mitochondrial metabolism and diabetes. *J Diabetes Investig* 2010;1:161–9.
- [151] Wallace DC. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu Rev Genet* 2005;39:359–407.
- [152] Martikainen MH, Ronnema T, Majamaa K. Prevalence of mitochondrial diabetes in southwestern Finland: a molecular epidemiological study. *Acta Diabetologica* 2013;50:737–41.
- [153] Martikainen MH, Ronnema T, Majamaa K. Association of mitochondrial DNA haplogroups and vascular complications of diabetes mellitus: A population-based study. *Diabetes Vasc Dis Res* 2015;12:302–4.
- [154] Crispim D, Canani LH, Gross JL, Tschiedel B, Souto KE, Roisenberg I. The European-specific mitochondrial cluster J/T could confer an increased risk of insulin-resistance and type 2 diabetes: an analysis of the m.4216T>C and m.4917A>G variants. *Ann Hum Genet* 2006;70:488–95.
- [155] Feder J, Ovadia O, Blech I, Cohen J, Wainstein J, Harman-Boehm I, Glaser B, Mishmar D. Parental diabetes status reveals association of mitochondrial DNA haplogroup J1 with type 2 diabetes. *BMC Med Genet* 2009;10:60.
- [156] Mohlke KL, Jackson AU, Scott LJ, Peck EC, Suh YD, Chines PS, Watanabe RM, Buchanan TA, Conneely KN, Erdos MR, Narisu N, Enloe S, Valle TT, Tuomilehto J, Bergman RN, Boehnke M, Collins FS. Mitochondrial polymorphisms and susceptibility to type 2 diabetes-related traits in Finns. *Hum Genet* 2005;118:245–54.

- [157] Fuku N, Park KS, Yamada Y, Nishigaki Y, Cho YM, Matsuo H, Segawa T, Watanabe S, Kato K, Yokoi K, Nozawa Y, Lee HK, Tanaka M. Mitochondrial haplogroup N9a confers resistance against type 2 diabetes in Asians. *Am J Hum Genet* 2007;80:407–15.
- [158] Hwang S, Kwak SH, Bhak J, Kang HS, Lee YR, Koo BK, Park KS, Lee HK, Cho YM. Gene expression pattern in transmitochondrial cytoplasmic hybrid cells harboring type 2 diabetes-associated mitochondrial DNA haplogroups. *PLoS One* 2011;6:e22116.
- [159] Jiang W, Li R, Zhang Y, Wang P, Wu T, Lin J, Yu J, Gu M. Mitochondrial DNA mutations associated with type 2 diabetes mellitus in Chinese Uyghur population. *Sci Rep* 2017;7:16989.
- [160] Niu Q, Zhang W, Wang H, Guan X, Lu J, Li W. Effects of mitochondrial haplogroup N9a on type 2 diabetes mellitus and its associated complications. *Exp Ther Med* 2015;10:1918–24.
- [161] Liao WQ, Pang Y, Yu CA, Wen JY, Zhang YG, Li XH. Novel mutations of mitochondrial DNA associated with type 2 diabetes in Chinese Han population. *Tohoku J Exp Med* 2008;215:377–84.
- [162] Guo LJ, Oshida Y, Fuku N, Takeyasu T, Fujita Y, Kurata M, Sato Y, Ito M, Tanaka M. Mitochondrial genome polymorphisms associated with type-2 diabetes or obesity. *Mitochondrion* 2005;5:15–33.
- [163] Ji F, Sharpley MS, Derbeneva O, Alves LS, Qian P, Wang Y, Chalkia D, Lvova M, Xu J, Yao W, Simon M, Platt J, Xu S, Angelin A, Davila A, Huang T, Wang PH, Chuang LM, Moore LG, Qian G, Wallace DC. Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. *Proc Natl Acad Sci USA* 2012;109:7391–6.
- [164] Sharma V, Sharma I, Singh VP, Verma S, Pandita A, Singh V, Rai E, Sharma S. mtDNA G10398A variation provides risk to type 2 diabetes in population group from the Jammu region of India. *Meta Gene* 2014;2:269–73.
- [165] Poulton J, Luan J, Macaulay V, Hennings S, Mitchell J, Wareham NJ. Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case-control study. *Hum Mol Genet* 2002;11:1581–3.
- [166] Estopinal CB, Chocron IM, Parks MB, Wade EA, Roberson RM, Burgess LG, Brantley Jr MA, Samuels DC. Mitochondrial haplogroups are associated with severity of diabetic retinopathy. *Investig Ophthalmol Vis Sci* 2014;55:5589–95.
- [167] Achilli A, Olivieri A, Pala M, Hooshar Kashani B, Carossa V, Perego UA, Gandini F, Santoro A, Battaglia V, Grugni V, Lancioni H, Sirolla C, Bonfigli AR, Cormio A, Boemi M, Testa I, Semino O, Ceriello A, Spazfumo L, Gadaleta MN, Marra M, Testa R, Franceschi C, Torroni A. Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. *PLoS One* 2011;6:e21029.
- [168] Kofler B, Mueller EE, Eder W, Stanger O, Maier R, Weger M, Haas A, Winker R, Schmut O, Paulweber B, Iglseider B, Renner W, Wiesbauer M, Aigner I, Santic D, Zimmermann FA, Mayr JA, Sperl W. Mitochondrial DNA haplogroup T is associated with coronary artery disease and diabetic retinopathy: a case control study. *BMC Med Genet* 2009;10:35.
- [169] Nardelli C, Labruna G, Liguori R, Mazzaccara C, Ferrigno M, Capobianco V, Pezzuti M, Castaldo G, Farinaro E, Contaldo F, Buono P, Sacchetti L, Pisanisi F. Haplogroup T is an obesity risk factor: mitochondrial DNA haplotyping in a morbid obese population from southern Italy. *BioMed Res Int* 2013;2013:631082.
- [170] Chinnery PF, Elliott HR, Syed A, Rothwell PM. Mitochondrial DNA haplogroups and risk of transient ischaemic attack and ischaemic stroke: a genetic association study. *Lancet Neurol* 2010;9:498–503.
- [171] Nishigaki Y, Yamada Y, Fuku N, Matsuo H, Segawa T, Watanabe S, Kato K, Yokoi K, Yamaguchi S, Nozawa Y, Tanaka M. Mitochondrial haplogroup N9b is protective against myocardial infarction in Japanese males. *Hum Genet* 2007;120:827–36.
- [172] Udar N, Atilan SR, Memarzadeh M, Boyer DS, Chwa M, Lu S, Maguen B, Langberg J, Coskun P, Wallace DC, Nesburn AB, Khatibi N, Hertzog D, Le K, Hwang D, Kenney MC. Mitochondrial DNA haplogroups associated with age-related macular degeneration. *Investig Ophthalmol Vis Sci* 2009;50:2966–74.
- [173] Jones MM, Manwaring N, Wang JJ, Rohtchina E, Mitchell P, Sue CM. Mitochondrial DNA haplogroups and age-related maculopathy. *Arch Ophthalmol* 2007;125:1235–40.
- [174] Soto-Hermida A, Fernandez-Moreno M, Oreiro N, Fernandez-Lopez C, Rego-Perez I, Blanco FJ. mtDNA haplogroups and osteoarthritis in different geographic populations. *Mitochondrion* 2014;15:18–23.
- [175] Montiel-Sosa F, Ruiz-Pesini E, Enriquez JA, Marcuello A, Diez-Sanchez C, Montoya J, Wallace DC, Lopez-Perez MJ. Differences of sperm motility in mitochondrial DNA haplogroup U sublineages. *Gene* 2006;368:21–7.
- [176] Ruiz-Pesini E, Lapena AC, Diez-Sanchez C, Perez-Martos A, Montoya J, Alvarez E, Diaz M, Urries A, Montoro L, Lopez-Perez MJ, Enriquez JA. Human mtDNA haplogroups associated with high or reduced spermatozoa motility. *Am J Hum Genet* 2000;67:682–96.
- [177] Niemi AK, Majamaa K. Mitochondrial DNA and ACTN3 genotypes in Finnish elite endurance and sprint athletes. *Eur J Hum Genet* 2005;13:965–9.

- [178] Scott RA, Fuku N, Onywera VO, Boit M, Wilson RH, Tanaka M, Goodwin WH, Pitsiladis YP. Mitochondrial haplogroups associated with elite Kenyan athlete status. *Med Sci Sports Exerc* 2009;41:123–8.
- [179] De Benedictis G, Rose G, Carrieri G, De Luca M, Falcone E, Passarino G, Bonafe M, Monti D, Baggio G, Bertolini S, Mari D, Mattace R, Franceschi C. Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. *FASEB J* 1999;13:1532–6.
- [180] Ivanova R, Lepage V, Charron D, Schachter F. Mitochondrial genotype associated with French Caucasian centenarians. *Gerontology* 1998;44:349.
- [181] Niemi AK, Hervonen A, Hurme M, Karhunen PJ, Jylha M, Majamaa K. Mitochondrial DNA polymorphisms associated with longevity in a Finnish population. *Hum Genet* 2003;112:29–33.
- [182] Rose G, Passarino G, Carrieri G, Altomare K, Greco V, Bertolini S, Bonafe M, Franceschi C, De Benedictis G. Paradoxes in longevity: sequence analysis of mtDNA haplogroup J in centenarians. *Eur J Hum Genet* 2001;9:701–7.
- [183] Tanaka M, Gong JS, Zhang J, Yoneda M, Yagi K. Mitochondrial genotype associated with longevity. *Lancet* 1998;351:185–6.
- [184] Fuku N, Pareja-Galeano H, Zempo H, Alis R, Arai Y, Lucia A, Hirose N. The mitochondrial-derived peptide MOTS-c: a player in exceptional longevity? *Aging Cell* 2015;14:921–3.
- [185] Zhai D, Ye Z, Jiang Y, Xu C, Ruan B, Yang Y, Lei X, Xiang A, Lu H, Zhu Z, Yan Z, Wei D, Li Q, Wang L, Lu Z. MOTS-c peptide increases survival and decreases bacterial load in mice infected with MRSA. *Mol Immunol* 2017;92:151–60.
- [186] Cho SB, Koh I, Nam HY, Jeon JP, Lee HK, Han BG. Mitochondrial DNA copy number augments performance of A1C and oral glucose tolerance testing in the prediction of type 2 diabetes. *Sci Rep* 2017;7:43203.
- [187] Moore AZ, Ding J, Tuke MA, Wood AR, Bandinelli S, Frayling TM, Ferrucci L. Influence of cell distribution and diabetes status on the association between mitochondrial DNA copy number and aging phenotypes in the InCHIANTI study. *Aging Cell* 2017. [ePub ahead of print] <https://doi.org/10.1111/accel.12683>.
- [188] Guyatt AL, Burrows K, Guthrie PAI, Ring S, McArdle W, Day INM, Ascione R, Lawlor DA, Gaunt TR, Rodriguez S. Cardiometabolic phenotypes and mitochondrial DNA copy number in two cohorts of UK women. *Mitochondrion* 2017. [ePub ahead of print] <https://doi.org/10.1016/j.mito.2017.08.007>.
- [189] Lien LM, Chiou HY, Yeh HL, Chiu SY, Jeng JS, Lin HJ, Hu CJ, Hsieh FI, Wei YH. Significant association between low mitochondrial DNA content in peripheral blood leukocytes and ischemic stroke. *J Am Heart Assoc* 2017;6.
- [190] Bonda DJ, Lee HP, Lee HG, Friedlich AL, Perry G, Zhu X, Smith MA. Novel therapeutics for Alzheimer's disease: an update. *Curr Opin Drug Discov Dev* 2010;13:235–46.
- [191] Diana FF, Silva Esteves AR, Oliveira CR, Cardoso SM. Mitochondria: the common upstream driver of a-beta and tau pathology in Alzheimer's Disease. *Curr Alzheimer Res* 2011;8:563–72.
- [192] Reddy PH, Beal MF. Are mitochondria critical in the pathogenesis of Alzheimer's disease? *Brain Research. Brain Res Rev* 2005;49:618–32.
- [193] Sullivan PG, Brown MR. Mitochondrial aging and dysfunction in Alzheimer's disease. *Prog Neuro Psychopharmacol Biol Psychiatry* 2005;29:407–10.
- [194] Benecke R, Strumper P, Weiss H. Electron transfer complexes I and IV of platelets are abnormal in Parkinson's disease but normal in Parkinson-plus syndromes. *Brain* 1993;116:1451–63.
- [195] Schapira AH, Cooper JM, Dexter D, Jenner P, Clark JB, Marsden CD. Mitochondrial complex I deficiency in Parkinson's disease. *Lancet* 1989;1:1269.
- [196] Wallace DC, Shoffner JM, Watts RL, Juncos JL, Torroni A. Mitochondrial oxidative phosphorylation defects in Parkinson's disease. *Ann Neurol* 1992;32:113–4.
- [197] Shoffner JM, Brown MD, Torroni A, Lott MT, Cabell MR, Mirra SS, Beal MF, Yang C, Gearing M, Salvo R, Watts RL, Juncos JL, Hansen LA, Crain BJ, Fayad M, Reckord CL, Wallace DC. Mitochondrial DNA variants observed in Alzheimer disease and Parkinson disease patients. *Genomics* 1993;17:171–84.
- [198] Hutchin T, Cortopassi G. A mitochondrial DNA clone is associated with increased risk for Alzheimer disease. *Proc Natl Acad Sci USA* 1995;92:6892–5.
- [199] Santoro A, Balbi V, Balducci E, Pirazzini C, Rosini F, Tavano F, Achilli A, Siviero P, Minicuci N, Bellavista E, Mishto M, Salvioli S, Marchegiani F, Cardelli M, Olivieri F, Nacmias B, Chiamenti AM, Benussi L, Ghidoni R, Rose G, Gabelli C, Binetti G, Sorbi S, Crepaldi G, Passarino G, Torroni A, Franceschi C. Evidence for sub-haplogroup h5 of mitochondrial DNA as a risk factor for late onset Alzheimer's disease. *PLoS One* 2010;5:e12037.
- [200] Lakatos A, Derbeneva O, Younes D, Keator D, Bakken T, Lvova M, Brandon M, Guffanti G, Reglodi D, Saykin A, Weiner M, Macciardi F, Schork N, Wallace DC, Potkin SG. Association between mitochondrial DNA variations and Alzheimer's disease in the ADNI cohort. *Neurobiol Aging* 2010;31:1355–63.

- [201] Coskun P, Helguera P, Nemati Z, Bohannon RC, Thomas J, Samuel SE, Argueta J, Doran E, Wallace DC, Lott IT, Busciglio J. Metabolic and growth rate alterations in lymphoblastic cell lines discriminate between Down Syndrome and Alzheimer's Disease. *J Alzheimers Dis* 2017;55:737–48.
- [202] Corral-Debrinski M, Horton T, Lott MT, Shoffner JM, McKee AC, Beal MF, Graham BH, Wallace DC. Marked changes in mitochondrial DNA deletion levels in Alzheimer brains. *Genomics* 1994;23:471–6.
- [203] Corral-Debrinski M, Shoffner JM, Lott MT, Wallace DC. Association of mitochondrial DNA damage with aging and coronary atherosclerotic heart disease. *Mutat Res* 1992b;275:169–80.
- [204] Corral-Debrinski M, Stepien G, Shoffner JM, Lott MT, Kanter K, Wallace DC. Hypoxemia is associated with mitochondrial DNA damage and gene induction. Implications for cardiac disease. *JAMA* 1991;266:1812–6.
- [205] Coskun P, Wyrembak J, Schriener SE, Chen HW, Marciniack C, LaFerla F, Wallace DC. A mitochondrial etiology of Alzheimer and Parkinson disease. *Biochim Biophys Acta* 2012;1820:553–64.
- [206] Coskun PE, Beal MF, Wallace DC. Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. *Proc Natl Acad Sci USA* 2004;101:10726–31.
- [207] Horton TM, Graham BH, Corral-Debrinski M, Shoffner JM, Kaufman AE, Beal BF, Wallace DC. Marked increase in mitochondrial DNA deletion levels in the cerebral cortex of Huntington's Disease patients. *Neurology* 1995;45:1879–83.
- [208] Helley MP, Pinnell J, Sportelli C, Tieu K. Mitochondria: a common target for genetic mutations and environmental toxicants in Parkinson's Disease. *Front Genet* 2017;8:177.
- [209] Ghezzi D, Marelli C, Achilli A, Goldwurm S, Pezzoli G, Barone P, Pellicchia MT, Stanzione P, Brusa L, Bentivoglio AR, Bonuccelli U, Petrozzi L, Abbruzzese G, Marchese R, Cortelli P, Grimaldi D, Martinelli P, Ferrarese C, Garavaglia B, Sangiorgi S, Carelli V, Torroni A, Albanese A, Zeviani M. Mitochondrial DNA haplogroup K is associated with a lower risk of Parkinson's disease in Italians. *Eur J Hum Genet* 2005;13:748–52.
- [210] Khusnutdinova E, Gilyazova I, Ruiz-Pesini E, Derbeneva O, Khusainova R, Khidiyatova I, Magzhanov R, Wallace DC. A mitochondrial etiology of neurodegenerative diseases: evidence from Parkinson's disease. *Ann NY Acad Sci* 2008;1147:1–20.
- [211] van der Walt JM, Nicodemus KK, Martin ER, Scott WK, Nance MA, Watts RL, Hubble JP, Haines JL, Koller WC, Lyons K, Pahwa R, Stern MB, Colcher A, Hiner BC, Jankovic J, Ondo WG, Allen Jr FH, Goetz CG, Small GW, Mastaglia F, Stajich JM, McLaurin AC, Middleton LT, Scott BL, Schmechel DE, Pericak-Vance MA, Vance JM. Mitochondrial polymorphisms significantly reduce the risk of Parkinson disease. *Am J Hum Genet* 2003;72:804–11.
- [212] Khalil B, Lievens JC. Mitochondrial quality control in amyotrophic lateral sclerosis: towards a common pathway? *Neural Regen Res* 2017;12:1052–61.
- [213] Konrad C, Kawamata H, Bredvik KG, Arreguin AJ, Cajamarca SA, Hupf JC, Ravits JM, Miller TM, Margakis NJ, Hales CM, Glass JD, Gross S, Mitumoto H, Manfredi G. Fibroblast bioenergetics to classify amyotrophic lateral sclerosis patients. *Mol Neurodegener* 2017;12:76.
- [214] Moller A, Bauer CS, Cohen RN, Webster CP, De Vos KJ. Amyotrophic lateral sclerosis-associated mutant SOD1 inhibits anterograde axonal transport of mitochondria by reducing Miro1 levels. *Hum Mol Genet* 2017;26:4668–79.
- [215] Wang H, Yi J, Li X, Xiao Y, Dhakal K, Zhou J. ALS-associated mutation SOD1(G93A) leads to abnormal mitochondrial dynamics in osteocytes. *Bone* 2018;106:126–38.
- [216] Straub IR, Janer A, Weraarpachai W, Zinman L, Robertson J, Rogaeva E, Shoubridge EA. Loss of CH-CHD10-CHCHD2 complexes required for respiration underlies the pathogenicity of a CHCHD10 mutation in ALS. *Hum Mol Genet* 2018;27:178–89.
- [217] Izumikawa K, Nobe Y, Yoshikawa H, Ishikawa H, Miura Y, Nakayama H, Nonaka T, Hasegawa M, Egawa N, Inoue H, Nishikawa K, Yamano K, Simpson RJ, Taoka M, Yamauchi Y, Isobe T, Takahashi N. TDP-43 stabilises the processing intermediates of mitochondrial transcripts. *Sci Rep* 2017;7:7709.
- [218] Rossignol DA, Frye RE. Mitochondrial dysfunction in autism spectrum disorders: a systematic review and meta-analysis. *Mol Psychiatry* 2012;17:290–314.
- [219] Leppa VM, Kravitz SN, Martin CL, Andrieux J, Le Caignec C, Martin-Coignard D, DyBuncio C, Sanders SJ, Lowe JK, Cantor RM, Geschwind DH. Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. *Am J Hum Genet* 2016;99:540–54.
- [220] Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K, Law P, Qiu S, Lord C, Sebat J, Ye K, Wigler M. A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci USA* 2007;104:12831–6.
- [221] De Rubeis S, He X, Goldberg AP, Poultnery CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, Kosmicki J, Shih-Chen F, Aleksic B, Biscaldi M, Bolton PF, Brownfeld JM, Cai J, Campbell

- NG, Carracedo A, Chahrouh MH, Chiocchetti AG, Coon H, Crawford EL, Curran SR, Dawson G, Duketis E, Fernandez BA, Gallagher L, Geller E, Guter SJ, Hill RS, Ionita-Laza J, Jimenez Gonzalez P, Kilpinen H, Klauck SM, Kolevzon A, Lee I, Lei J, Lehtimäki T, Lin CF, Ma'ayan A, Marshall CR, McInnes AL, Neale B, Owen MJ, Ozaki N, Parellada M, Parr JR, Purcell S, Puura K, Rajagopalan D, Rehnstrom K, Reichenberg A, Sabo A, Sachse M, Sanders SJ, Schafer C, Schulte-Ruther M, Skuse D, Stevens C, Szatmari P, Tammimies K, Valladares O, Voran A, Li-San W, Weiss LA, Willsey AJ, Yu TW, Yuen RK, Study DDD, Homozygosity Mapping Collaborative for, A., Consortium, U.K., Cook EH, Freitag CM, Gill M, Hultman CM, Lehner T, Palotie A, Schellenberg GD, Sklar P, State MW, Sutcliffe JS, Walsh CA, Scherer SW, Zwick ME, Barrett JC, Cutler DJ, Roeder K, Devlin B, Daly MJ, Buxbaum JD. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 2014;515:209–15.
- [222] Kosmicki JA, Samocha KE, Howrigan DP, Sanders SJ, Slowikowski K, Lek M, Karczewski KJ, Cutler DJ, Devlin B, Roeder K, Buxbaum JD, Neale BM, MacArthur DG, Wall DP, Robinson EB, Daly MJ. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nature Genet* 2017;49:504–10.
- [223] Krumm N, O'Roak BJ, Shendure J, Eichler EE. A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci* 2014;37:95–105.
- [224] Bishop SL, Farmer C, Bal V, Robinson EB, Willsey AJ, Werling DM, Havdahl KA, Sanders SJ, Thurm A. Identification of developmental and behavioral markers associated with genetic abnormalities in autism spectrum disorder. *Am J Psychiatry* 2017;174:576–85.
- [225] Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, Wong LC, Estabillio JA, Gadowski TE, Hong O, Fajardo KV, Bhandari A, Owen R, Baughn M, Yuan J, Solomon T, Moyzis AG, Maile MS, Sanders SJ, Reiner GE, Vaux KK, Strom CM, Zhang K, Muotri AR, Akshoomoff N, Leal SM, Pierce K, Courchesne E, Iakoucheva LM, Corsello C, Sebat J. Frequency and complexity of de novo structural mutation in autism. *Am J Hum Genet* 2016;98:667–79.
- [226] Robinson EB, St Pourcain B, Anttila V, Kosmicki JA, Bulik-Sullivan B, Grove J, Maller J, Samocha KE, Sanders SJ, Ripke S, Martin J, Hollegaard MV, Werge T, Hougaard DM, iPsych-SSI Broad Autism Group, Neale BM, Evans DM, Skuse D, Mortensen PB, Borglum AD, Ronald A, Smith GD, Daly MJ. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature Genet* 2016;48:552–5.
- [227] Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* 2017;8:21.
- [228] Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, Samocha KE, Goldstein JI, Okbay A, Bybjerg-Grauholm J, Werge T, Hougaard DM, Taylor J, i, P.-B.A.G., Psychiatric Genomics Consortium Autism, G., Skuse D, Devlin B, Anney R, Sanders SJ, Bishop S, Mortensen PB, Borglum AD, Smith GD, Daly MJ, Robinson EB. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet* 2017;49:978–85.
- [229] Smith M, Flodman PL, Gargus JJ, Simon MT, Verrell K, Haas R, Reiner GE, Naviaux R, Osann K, Spence MA, Wallace DC. Mitochondrial and ion channel gene alterations in autism. *Biochim Biophys Acta* 2012;1817:1796–802.
- [230] Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, Keaney 3rd JF, Klei L, Mandell JD, Moreno-De-Luca D, Poultney CS, Robinson EB, Smith L, Solli-Nowlan T, Su MY, Teran NA, Walker MF, Werling DM, Beaudet AL, Cantor RM, Fombonne E, Geschwind DH, Grice DE, Lord C, Lowe JK, Mane SM, Martin DM, Morrow EM, Talkowski ME, Sutcliffe JS, Walsh CA, Yu TW, Autism Sequencing Consortium, Ledbetter DH, Martin CL, Cook EH, Buxbaum JD, Daly MJ, Devlin B, Roeder K, State MW. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* 2015;87:1215–33.
- [231] Yoon JC, Ng A, Kim BH, Bianco A, Xavier RJ, Elledge SJ. Wnt signaling regulates mitochondrial physiology and insulin sensitivity. *Genes Dev* 2010;24:1507–18.
- [232] Chalkia D, Singh LN, Leipzig J, Lvova M, Derbeneva O, Lakatos A, Hadley D, Hakonarson H, Wallace DC. Association between mitochondrial DNA haplogroup variation and autism spectrum disorders. *JAMA Psychiatry* 2017;74:1161–8.
- [233] Wang Y, Picard M, Gu Z. Genetic evidence for elevated pathogenicity of mitochondrial DNA heteroplasmy in Autism Spectrum Disorder. *PLoS Genet* 2016;12:e1006391.
- [234] Wallace DC. A mitochondrial etiology of neuropsychiatric disorders. *JAMA Psychiatry* 2017;74:863–4.
- [235] Anglin RE, Mazurek MF, Tarnopolsky MA, Rosebush PI. The mitochondrial genome and psychiatric illness. *American Journal of Medical Genetics. Part B, Neuropsychiatr Genet* 2012a;159B:749–59.

- [236] Anglin RE, Tarnopolsky MA, Mazurek MF, Rosebush PI. The psychiatric presentation of mitochondrial disorders in adults. *J Neuropsychiatry Clin Neurosci* 2012b;24:394–409.
- [237] Manji H, Kato T, Di Prospero NA, Ness S, Beal MF, Krams M, Chen G. Impaired mitochondrial function in psychiatric disorders. *Nature Reviews. Neuroscience* 2012;13:293–307.
- [238] Rosebush PI, Anglin RE, Rasmussen S, Mazurek MF. Mental illness in patients with inherited mitochondrial disorders. *Schizophr Res* 2017;187:33–7.
- [239] Zuccoli GS, Saia-Cereda VM, Nascimento JM, Martins-de-Souza D. The energy metabolism dysfunction in psychiatric disorders postmortem brains: focus on proteomic evidence. *Front Neurosci* 2017;11:493.
- [240] Baudouin SV, Saunders D, Tiangyou W, Elson JL, Poynter J, Pyle A, Keers S, Turnbull DM, Howell N, Chinnery PF. Mitochondrial DNA and survival after sepsis: a prospective study. *Lancet* 2005;366:2118–21.
- [241] Raby BA, Klanderman B, Murphy A, Mazza S, Camargo Jr CA, Silverman EK, Weiss ST. A common mitochondrial haplogroup is associated with elevated total serum IgE levels. *J Allergy Clin Immunol* 2007;120:351–8.
- [242] Hendrickson SL, Hutcheson HB, Ruiz-Pesini E, Poole JC, Lautenberger J, Sezgin E, Kingsley L, Goedert JJ, Vlahov D, Donfield S, Wallace DC, O'Brien SJ. Mitochondrial DNA haplogroups influence AIDS progression. *AIDS* 2008;22:2429–39.
- [243] Dela Cruz CS, Kang MJ. Mitochondrial dysfunction and damage associated molecular patterns (DAMPs) in chronic inflammatory diseases. *Mitochondrion* 2017. [ePub ahead of print] <http://www.sciencedirect.com/science/article/pii/S1567724917301897>.
- [244] Dan Dunn J, Alvarez LA, Zhang X, Soldati T. Reactive oxygen species and mitochondria: A nexus of cellular homeostasis. *Redox Biol* 2015;6:472–85.
- [245] Nadeau-Vallee M, Obari D, Quiniou C, Lubell WD, Olson DM, Girard S, Chemtob S. A critical role of interleukin-1 in preterm labor. *Cytokine Growth Factor Rev* 2016;28:37–51.
- [246] Angelin A, Gil-de-Gomez L, Dahiya S, Jiao J, Guo L, Levine MH, Wang Z, Quinn 3rd WJ, Kopinski PK, Wang L, Akimova T, Liu Y, Bhatti TR, Han R, Laskin BL, Baur JA, Blair IA, Wallace DC, Hancock WW, Beier UH. Foxp3 reprograms T cell metabolism to function in low-glucose, high-lactate environments. *Cell Metab* 2017;25:1282–93.
- [247] Kilbaugh TJ, Lvova M, Karlsson M, Zhang Z, Leipzig J, Wallace DC, Margulies SS. Peripheral blood mitochondrial DNA as a biomarker of cerebral mitochondrial dysfunction following traumatic brain injury in a porcine model. *PLoS One* 2015;10:e0130927.
- [248] Newman NJ, Yu-Wai-Man P, Sadun AA, Karanjia R, Carelli V. Management of ophthalmologic manifestations of mitochondrial diseases. *Genet Med* 2017;19.
- [249] Parikh S, Goldstein A, Karaa A, Koenig MK, Anselm I, Brunel-Guitton C, Christodoulou J, Cohen BH, Dimmock D, Enns GM, Falk MJ, Feigenbaum A, Frye RE, Ganesh J, Griesemer D, Haas R, Horvath R, Korson M, Kruer MC, Mancuso M, McCormack S, Josee Raboisson M, Reimschisel T, Salvarinova R, Saneto RP, Scaglia F, Shoffner J, Stacpoole PW, Sue CM, Tarnopolsky M, Van Karnebeek C, Wolfe LA, Zolkipli Cunningham Z, Rahman S, Chinnery PF. Response to Newman et al. *Genet Med* 2017a;19.
- [250] Parikh S, Goldstein A, Karaa A, Koenig MK, Anselm I, Brunel-Guitton C, Christodoulou J, Cohen BH, Dimmock D, Enns GM, Falk MJ, Feigenbaum A, Frye RE, Ganesh J, Griesemer D, Haas R, Horvath R, Korson M, Kruer MC, Mancuso M, McCormack S, Raboisson MJ, Reimschisel T, Salvarinova R, Saneto RP, Scaglia F, Shoffner J, Stacpoole PW, Sue CM, Tarnopolsky M, Van Karnebeek C, Wolfe LA, Cunningham ZZ, Rahman S, Chinnery PF. Patient care standards for primary mitochondrial disease: a consensus statement from the Mitochondrial Medicine Society. *Genet Med* 2017b;19:689.
- [251] Parikh S, Goldstein A, Koenig MK, Scaglia F, Enns GM, Saneto R, Anselm I, Cohen BH, Falk MJ, Greene C, Gropman AL, Haas R, Hirano M, Morgan P, Sims K, Tarnopolsky M, Van Hove JLK, Wolfe L, DiMauro S. Diagnosis and management of mitochondrial disease: a consensus statement from the Mitochondrial Medicine Society. *Genet Med* 2015;17:689–701.
- [252] Dinwiddie DL, Smith LD, Miller NA, Atherton AM, Farrow EG, Strenk ME, Soden SE, Saunders CJ, King-smore SF. Diagnosis of mitochondrial disorders by concomitant next-generation sequencing of the exome and mitochondrial genome. *Genomics* 2013;102:148–56.
- [253] Tang S, Wang J, Zhang VW, Li FY, Landsverk M, Cui H, Truong CK, Wang G, Chen LC, Graham B, Scaglia F, Schmitt ES, Craigen WJ, Wong LJ. Transition to next generation analysis of the whole mitochondrial genome: a summary of molecular defects. *Human Mutat* 2013;34:882–93.
- [254] Wong LJ. Next generation molecular diagnosis of mitochondrial disorders. *Mitochondrion* 2013;13:379–87.
- [255] Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, Haack TB, Graf E, Schwarzmayer T, Terrile C, Konarikova E, Repp B, Kastenmuller G, Adamski J, Lichtner P, Leonhardt C, Funalot B, Donati A, Tiranti V, Lombes A, Jardel C, Glaser D, Taylor RW, Ghezzi D, Mayr JA, Rotig A, Freisinger P, Distelmaier F, Strom

- TM, Meitinger T, Gagneur J, Prokisch H. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun* 2017;8:15824.
- [256] Davis R, Liang C, Sue C. A new diagnostic paradigm for mitochondrial disease (Abstract 20). *J Clin Neurosci* 2014;21:2039.
- [257] Morovat A, Weerasinghe G, Nesbitt V, Hofer M, Agnew T, Quaghebeur G, Sergeant K, Fratter C, Guha N, Mirzazadeh M, Poulton J. Use of FGF-21 as a biomarker of mitochondrial disease in clinical practice. *J Clin Med* 2017;6:80.
- [258] Suomalainen A, Elo JM, Pietilainen KH, Hakonen AH, Sevastianova K, Korpela M, Isohanni P, Marjavaara SK, Tyni T, Kiuru-Enari S, Pihko H, Darin N, Ounap K, Kluijtmans LA, Paetau A, Buzkova J, Bindoff LA, Annunen-Rasila J, Uusimaa J, Rissanen A, Yki-Jarvinen H, Hirano M, Tulinius M, Smeitink J, Tynjismaa H. FGF-21 as a biomarker for muscle-manifesting mitochondrial respiratory chain deficiencies: a diagnostic study. *Lancet Neurol* 2011;10:806–18.
- [259] Montero R, Yubero D, Villarroja J, Henares D, Jou C, Rodriguez MA, Ramos F, Nascimento A, Ortez CI, Campistol J, Perez-Duenas B, O'Callaghan M, Pineda M, Garcia-Cazorla A, Oferil JC, Montoya J, Ruiz-Pesini E, Emperador S, Meznaric M, Campderros L, Kalko SG, Villarroja F, Artuch R, Jimenez-Mallebrera C. GDF-15 is elevated in children with mitochondrial diseases and is induced by mitochondrial dysfunction. *PLoS One* 2016;11:e0148709.
- [260] Yatsuga S, Fujita Y, Ishii A, Fukumoto Y, Arahata H, Kakuma T, Kojima T, Ito M, Tanaka M, Saiki R, Koga Y. Growth differentiation factor 15 as a useful biomarker for mitochondrial disorders. *Ann Neurol* 2015;78:814–23.
- [261] Chao de la Barca JM, Simard G, Amati-Bonneau P, Safiedeen Z, Prunier-Mirebeau D, Chupin S, Gadras C, Tessier L, Gueguen N, Chevrollier A, Desquirit-Dumas V, Ferre M, Bris C, Kouassi Nzoughe J, Bocca C, Lereux S, Verny C, Milea D, Bonneau D, Lenaers G, Martinez MC, Procaccio V, Reynier P. The metabolomic signature of Leber's hereditary optic neuropathy reveals endoplasmic reticulum stress. *Brain* 2016;139:2864–76.
- [262] Esterhuizen K, van der Westhuizen FH, Louw R. Metabolomics of mitochondrial disease. *Mitochondrion* 2017;35:97–110.
- [263] Quiros PM, Prado MA, Zamboni N, D'Amico D, Williams RW, Finley D, Gygi SP, Auwerx J. Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J Cell Biol* 2017;216:2027–45.
- [264] Vivian CJ, Brinker AE, Graw S, Koestler DC, Legendre C, Gooden GC, Salhia B, Welch DR. Mitochondrial genomic backgrounds affect nuclear DNA methylation and gene expression. *Cancer Res* 2017;77:6202–14.
- [265] Pfeffer G, Majamaa K, Turnbull DM, Thorburn D, Chinnery PF. Treatment for mitochondrial disorders. *Cochrane Database System Rev* 2012:CD004426.
- [266] Murphy MP. Selective targeting of bioactive compounds to mitochondria. *Trends Biotechnol* 1997;15:326–30.
- [267] Smith RA, Kelso GF, James AM, Murphy MP. Targeting coenzyme Q derivatives to mitochondria. *Methods Enzymol* 2004;382:45–67.
- [268] McManus MJ, Murphy MP, Franklin JL. The mitochondria-targeted antioxidant MitoQ prevents loss of spatial memory retention and early neuropathology in a transgenic mouse model of Alzheimer's disease. *J Neurosci* 2011;31:15703–15.
- [269] Pell VR, Chouchani ET, Murphy MP, Brookes PS, Krieg T. Moving forwards by blocking back-flow: the yin and yang of MI therapy. *Circ Res* 2016;118:898–906.
- [270] Carelli V, La Morgia C, Valentino ML, Rizzo G, Carbonelli M, De Negri AM, Sadun F, Carta A, Guerriero S, Simonelli F, Sadun AA, Aggarwal D, Liguori R, Avoni P, Baruzzi A, Zeviani M, Montagna P, Barboni P. Idebenone treatment In Leber's Hereditary Optic Neuropathy. *Brain* 2011;134:e188.
- [271] Mashima Y, Kigasawa K, Wakakura M, Oguchi Y. Do idebenone and vitamin therapy shorten the time to achieve visual recovery in Leber hereditary optic neuropathy? *J Neuro Ophthalmol* 2000;20:166–70.
- [272] Klopstock T, Yu-Wai-Man P, Dimitriadis K, Rouleau J, Heck S, Bailie M, Atawan A, Chattopadhyay S, Schubert M, Garip A, Kernt M, Petraki D, Rummey C, Leinonen M, Metz G, Griffiths PG, Meier T, Chinnery PF. A randomized placebo-controlled trial of idebenone in Leber's hereditary optic neuropathy. *Brain* 2011;134:2677–86.
- [273] Angebault C, Gueguen N, Desquirit-Dumas V, Chevrollier A, Guillet V, Verny C, Cassereau J, Ferre M, Milea D, Amati-Bonneau P, Bonneau D, Procaccio V, Reynier P, Loiseau D. Idebenone increases mitochondrial complex I activity in fibroblasts from LHON patients while producing contradictory effects on respiration. *BMC Res Notes* 2011;4:557.
- [274] Giorgio V, Petronilli V, Ghelli A, Carelli V, Rugolo M, Lenaz G, Bernardi P. The effects of idebenone on mitochondrial bioenergetics. *Biochim Biophys Acta* 2012;1817:363–9.
- [275] Enns GM, Kinsman SL, Perlman SL, Spicer KM, Abdenur JE, Cohen BH, Amagata A, Barnes A, Kheifets V, Shrader WD, Thoolen M, Blankenberg F, Miller G. Initial experience in the treatment of inherited mitochondrial disease with EPI-743. *Mol Genet Metab* 2012;105:91–102.

- [276] Sadun AA, Chicani CF, Ross-Cisneros FN, Barboni P, Thoolen M, Shrader WD, Kubis K, Carelli V, Miller G. Effect of EPI-743 on the clinical course of the mitochondrial disease Leber Hereditary Optic Neuropathy. *Arch Neurol* 2012;69:331–8.
- [277] Pedram A, Razandi M, Wallace DC, Levin ER. Functional estrogen receptors in the mitochondria of breast cancer cells. *Mol Biol Cell* 2006;17:2125–37.
- [278] Pisano A, Preziuso C, Iommarini L, Perli E, Grazioli P, Campese AF, Maresca A, Montopoli M, Masuelli L, Sadun AA, d'Amati G, Carelli V, Ghelli A, Giordano C. Targeting estrogen receptor beta as preventive therapeutic strategy for Leber's hereditary optic neuropathy. *Hum Mol Genet* 2015;24:6921–31.
- [279] Szeto HH. First-in-class cardiolipin-protective compound as a therapeutic agent to restore mitochondrial bioenergetics. *Br J Pharmacol* 2014;171:2029–50.
- [280] Giordano C, Iommarini L, Giordano L, Maresca A, Pisano A, Valentino ML, Caporali L, Liguori R, Deceglie S, Roberti M, Fanelli F, Fracasso F, Ross-Cisneros FN, D'Adamo P, Hudson G, Pyle A, Yu-Wai-Man P, Chinnery PF, Zeviani M, Salomao SR, Berezovsky A, Belfort Jr R, Ventura DF, Moraes M, Moraes Filho M, Barboni P, Sadun F, De Negri A, Sadun AA, Tancredi A, Mancini M, d'Amati G, Loguercio Polosa P, Cantatore P, Carelli V. Efficient mitochondrial biogenesis drives incomplete penetrance in Leber's hereditary optic neuropathy. *Brain* 2014;137:335–53.
- [281] Carelli V, d'Adamo P, Valentino ML, La Morgia C, Ross-Cisneros FN, Caporali L, Maresca A, Loguercio Polosa P, Barboni P, De Negri A, Sadun F, Karanjia R, Salomao SR, Berezovsky A, Chicani F, Moraes M, Moraes Filho M, Belfort Jr R, Sadun AA. Parsing the differences in affected with LHON: genetic versus environmental triggers of disease conversion. *Brain* 2016;139:e17.
- [282] Giordano L, Deceglie S, d'Adamo P, Valentino ML, La Morgia C, Fracasso F, Roberti M, Cappellari M, Petrosillo G, Ciaravolo S, Parente D, Giordano C, Maresca A, Iommarini L, Del Dotto V, Ghelli AM, Salomao SR, Berezovsky A, Belfort Jr R, Sadun AA, Carelli V, Loguercio Polosa P, Cantatore P. Cigarette toxicity triggers Leber's hereditary optic neuropathy by affecting mtDNA copy number, oxidative phosphorylation and ROS detoxification pathways. *Cell Death Dis* 2015;6:e2021.
- [283] Pei L, Wallace DC. Mitochondrial etiology of neuropsychiatric disorders. *Biol Psychiatry* 2017. [ePub ahead of print] <https://doi.org/10.1016/j.biopsych.2017.11.018>.
- [284] Wallace DC, Fan W. Energetics, epigenetics, mitochondrial genetics. *Mitochondrion* 2010;10:12–31.
- [285] Wallace DC, Fan W, Procaccio V. Mitochondrial energetics and therapeutics. *Annu Rev Pathol* 2010;5:297–348.
- [286] Johri A, Calingasan NY, Hennessey TM, Sharma A, Yang L, Wille E, Chandra A, Beal MF. Pharmacologic activation of mitochondrial biogenesis exerts widespread beneficial effects in a transgenic mouse model of Huntington's disease. *Hum Mol Genet* 2012;21:1124–37.
- [287] Wenz T, Diaz F, Spiegelman BM, Moraes CT. Activation of the PPAR/PGC-1alpha pathway prevents a bioenergetic deficit and effectively improves a mitochondrial myopathy phenotype. *Cell Metab* 2008;8:249–56.
- [288] Wenz T, Diaz F, Spiegelman BM, Moraes CT. Retraction notice to: Activation of the PPAR/PGC-1alpha pathway prevents a bioenergetic deficit and effectively improves a mitochondrial myopathy phenotype. *Cell Metab* 2016;24:889.
- [289] Burte F, Carelli V, Chinnery PF, Yu-Wai-Man P. Disturbed mitochondrial dynamics and neurodegenerative disorders. *Nat Rev Neurol* 2015;11:11–24.
- [290] Georgakopoulos ND, Wells G, Campanella M. The pharmacological regulation of cellular mitophagy. *Nat Chem Biol* 2017;13:136–46.
- [291] Sorrentino V, Menzies KJ, Auwerx J. Repairing mitochondrial dysfunction in disease. *Annu Rev Pharmacol Toxicol* 2017;58. [ePub ahead of print] <https://doi.org/10.1146/annurev-pharmtox-010716-104908>.
- [292] Flierl A, Chen Y, Coskun PE, Samulski RJ, Wallace DC. Adeno-associated virus-mediated gene transfer of the heart/muscle adenine nucleotide translocator (ANT) in mouse. *Gene Ther* 2005;12:570–8.
- [293] Flierl A, Jackson C, Cottrell B, Murdock D, Seibel P, Wallace DC. Targeted delivery of DNA to the mitochondrial compartment via import sequence-conjugated peptide nucleic acid. *Mol Ther* 2003;7:550–7.
- [294] Yu H, Koilkonda RD, Chou TH, Porciatti V, Ozdemir SS, Chiodo V, Boye SL, Boye SE, Hauswirth WW, Lewin AS, Guy J. Gene delivery to mitochondria by targeting modified adenoassociated virus suppresses Leber's hereditary optic neuropathy in a mouse model. *Proc Natl Acad Sci USA* 2012;109:E1238–47.
- [295] Bonnet C, Kaltimbacher V, Ellouze S, Augustin S, Benit P, Forster V, Rustin P, Sahel JA, Corral-Debrinski M. Allotopic mRNA localization to the mitochondrial surface rescues respiratory chain defects in fibroblasts harboring mitochondrial DNA mutations affecting complex I or v subunits. *Rejuvenation Res* 2007;10:127–44.
- [296] Cwerman-Thibault H, Sahel JA, Corral-Debrinski M. Mitochondrial medicine: to a new era of gene therapy for mitochondrial DNA mutations. *J Inherit Metab Dis* 2011;34:327–44.

- [297] Ellouze S, Augustin S, Bouaita A, Bonnet C, Simonutti M, Forster V, Picaud S, Sahel JA, Corral-Debrinski M. Optimized allotopic expression of the human mitochondrial ND4 prevents blindness in a rat model of mitochondrial dysfunction. *Am J Hum Genet* 2008;83:373–87.
- [298] Guy J, Qi X, Koilkonda RD, Arguello T, Chou TH, Ruggeri M, Porciatti V, Lewin AS, Hauswirth WW. Efficiency and safety of AAV-mediated gene delivery of the human ND4 complex I subunit in the mouse visual system. *Investig Ophthalmol Vis Sci* 2009;50:4205–14.
- [299] Guy J, Qi X, Pallotti F, Schon EA, Manfredi G, Carelli V, Martinuzzi A, Hauswirth WW, Lewin AS. Rescue of a mitochondrial deficiency causing Leber Hereditary Optic Neuropathy. *Ann Neurol* 2002;52:534–42.
- [300] Koilkonda RD, Chou TH, Porciatti V, Hauswirth WW, Guy J. Induction of rapid and highly efficient expression of the human ND4 complex I subunit in the mouse visual system by self-complementary adeno-associated virus. *Arch Ophthalmol* 2010;128:876–83.
- [301] Koilkonda RD, Hauswirth WW, Guy J. Efficient expression of self-complementary AAV in ganglion cells of the ex vivo primate retina. *Mol Vis* 2009;15:2796–802.
- [302] Qi X, Sun L, Lewin AS, Hauswirth WW, Guy J. The mutant human ND4 subunit of complex I induces optic neuropathy in the mouse. *Investig Ophthalmol Vis Sci* 2007;48:1–10.
- [303] Sylvestre J, Margeot A, Jacq C, Dujardin G, Corral-Debrinski M. The role of the 3' untranslated region in mRNA sorting to the vicinity of mitochondria is conserved from yeast to human cells. *Mol Biol Cell* 2003a;14:3848–56.
- [304] Sylvestre J, Vialette S, Corral Debrinski M, Jacq C. Long mRNAs coding for yeast mitochondrial proteins of prokaryotic origin preferentially localize to the vicinity of mitochondria. *Genome Biol* 2003b;4:R44.
- [305] Towheed A, Markantone DM, Crain AT, Celotto AM, Palladino MJ. Small mitochondrial-targeted RNAs modulate endogenous mitochondrial protein expression *in vivo*. *Neurobiol Dis* 2014;69:15–22.
- [306] Wang G, Chen HW, Oktay Y, Zhang J, Allen EL, Smith GM, Fan KC, Hong JS, French SW, McCaffery JM, Lightowlers RN, Morse 3rd HC, Koehler CM, Teitell MA. PNPASE regulates RNA import into mitochondria. *Cell* 2010;142:456–67.
- [307] Wang G, Shimada E, Koehler CM, Teitell MA. PNPASE and RNA trafficking into mitochondria. *Biochim Biophys Acta* 2012;1819:998–1007.
- [308] Craven L, Tuppen HA, Greggains GD, Harbottle SJ, Murphy JL, Cree LM, Murdoch AP, Chinnery PF, Taylor RW, Lightowlers RN, Herbert M, Turnbull DM. Pronuclear transfer in human embryos to prevent transmission of mitochondrial DNA disease. *Nature* 2010;465:82–5.
- [309] Hyslop LA, Blakeley P, Craven L, Richardson J, Fogarty NM, Fragouli E, Lamb M, Wamaitha SE, Prathalingam N, Zhang Q, O'Keefe H, Takeda Y, Arizzi L, Alfarawati S, Tuppen HA, Irving L, Kalleas D, Choudhary M, Wells D, Murdoch AP, Turnbull DM, Niakan KK, Herbert M. Towards clinical application of pronuclear transfer to prevent mitochondrial DNA disease. *Nature* 2016;534:383–6.
- [310] Kang E, Wu J, Gutierrez NM, Koski A, Tippner-Hedges R, Agaronyan K, Platero-Luengo A, Martinez-Redondo P, Ma H, Lee Y, Hayama T, Van Dyken C, Wang X, Luo S, Ahmed R, Li Y, Ji D, Kayali R, Cinnioglu C, Olson S, Jensen J, Battaglia D, Lee D, Wu D, Huang T, Wolf DP, Temiakov D, Belmonte JC, Amato P, Mitalipov S. Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature* 2016;540:270–5.
- [311] Paull D, Emmanuele V, Weiss KA, Treff N, Stewart L, Hua H, Zimmer M, Kahler DJ, Goland RS, Noggle SA, Prosser R, Hirano M, Sauer MV, Egli D. Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants. *Nature* 2013;493:632–7.
- [312] Tachibana M, Amato P, Sparman M, Woodward J, Sanchis DM, Ma H, Gutierrez NM, Tippner-Hedges R, Kang E, Lee HS, Ramsey C, Masterson K, Battaglia D, Lee D, Wu D, Jensen J, Patton P, Gokhale S, Stouffer R, Mitalipov S. Towards germline gene therapy of inherited mitochondrial diseases. *Nature* 2013;493:627–31.
- [313] Tachibana M, Sparman M, Sritanandomchai H, Ma H, Clepper L, Woodward J, Li Y, Ramsey C, Kolotushkina O, Mitalipov S. Mitochondrial gene replacement in primate offspring and embryonic stem cells. *Nature* 2009;461:367–72.
- [314] Wolf DP, Hayama T, Mitalipov S. Mitochondrial genome inheritance and replacement in the human germline. *EMBO J* 2017;36:2177–81. Corrigendum: *EMBO J*. 2017 (Sep) 2136(2117):2659.
- [315] Slone J, Zhang J, Huang T. Experience from the first live-birth derived from oocyte nuclear transfer as a treatment strategy for mitochondrial diseases. *J Mol Genet Med* 2017;11:1000258.
- [316] Zhang J, Liu H, Luo S, Lu Z, Chavez-Badiola A, Liu Z, Yang M, Merhi Z, Silber SJ, Munne S, Konstantinidis M, Wells D, Tang JJ, Huang T. Live birth derived from oocyte spindle transfer to prevent mitochondrial disease. *Reprod Biomed Online* 2017;34:361–8. Corrigendum: *Reprod Biomed Online* 2017 (Jul) 2035(2011):2049.

Multifactorial Inheritance and Complex Diseases

Allison Fialkowski, T. Mark Beasley,
Hemant K. Tiwari

Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL,
United States

11.1 INTRODUCTION

If a disease or condition is caused by a single locus of large effect it is called a single-gene or *monogenic* disease, disorder, or, more generically, condition. There are over 10,000 such examples, which include cystic fibrosis, Huntington disease, Duchenne muscular dystrophy, and Marfan syndrome. A single-gene disease may have *locus heterogeneity* if that disease is caused by specific mutations on different genes, but this is more properly considered a special case of an oligogenic disorder. For example, osteogenesis imperfecta is caused by a single mutation in a type I collagen gene on either chromosome 7 or chromosome 17. *Oligogenic* disorders are explained by a few loci with large effects (for examples, see [1]). In contrast to oligogenic traits, *polygenic* inheritance is due to many loci with small effects at each locus. Thus, the term polygenic is generally used to describe multiple factors that are exclusively genetic. Any of these genetic effects, with or without the combination of an environmental effect, can give rise to a *multifactorial* disorder. Therefore, multifactorial diseases are caused by the simultaneous action of multiple genetic and/or environmental factors.

In contrast to *dichotomous* traits (i.e., affected vs. unaffected), *quantitative* traits are measured on a continuous scale, and most of them are thought to be multifactorial (e.g., blood pressure, body mass index). Some quantitative traits may be due to major gene

effects with a multifactorial background. Multifactorial inheritance is responsible for the majority of modern deleterious health conditions, such as heart disease and diabetes. Atopic syndrome, diabetes, cancer, spina bifida/anencephaly, pyloric stenosis, cleft palate, congenital hip dysplasia, club foot, and many other diseases and complex phenotypes also result from multifactorial inheritance.

Definitions and terminology: The *polygenic model* has its origins in Fisher's seminal work [2], which showed that "many small, equal and additive loci" would result in a Gaussian (or normal) distribution for a phenotype. Similarly, the combined additive effects of many genetic and environmental factors will also produce an approximately Gaussian phenotypic distribution. To illustrate, suppose (naïvely) that a quantitative trait such as percentage body fat is determined by a single gene with two codominant¹ alleles, *A* and *a*, which have equal frequency ($P = .50$). Assume individuals with an *A* allele tend to have a higher value of the trait, while individuals with an *a* allele tend to have a lower value of the trait. If *A* has an additive effect, then there are three distinct phenotypic groups, namely, high (2), intermediate (1), and low (0). If the allele frequencies of *A* and *a* are both

¹ "An allele 'a' is said to be *codominant* with respect to the wild-type allele 'A' if the *A/a* heterozygote fully expresses both of the phenotypes associated with the *a/a* and *A/A* homozygotes" (<http://www.informatics.jax.org/glossary/codominant>).

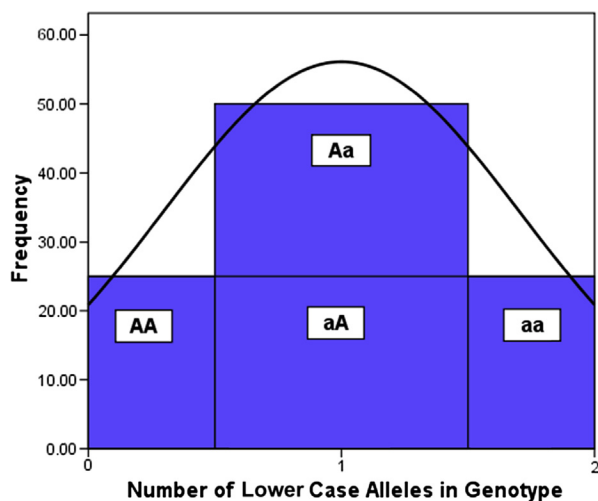


Figure 11.1 Expected phenotype distribution for a trait with a single causal locus with allele frequency of 50% and in Hardy-Weinberg equilibrium.

TABLE 11.1 Frequency Distribution of Genotypic Values for Two Loci With No Linkage Disequilibrium

	AA	Aa	aa
BB	0.0625	0.1250	0.1250
Bb	0.1250	0.2500	0.1250
bb	0.0625	0.1250	0.0625

0.50, then 25% of individuals would be expected to be *aa* and of low percentage fat, 50% would be expected to be *Aa* and of moderate percentage fat, and 25% would be expected to be *AA* and of high percentage fat. Fig. 11.1 gives the distribution of the trait in a population.

Now, suppose that the trait is determined by two loci. The second locus also has two codominant alleles, *B* for high and *b* for low expression of the trait, with *B* having an allele frequency of 0.50 and the same effect magnitude as the *A* allele. There are now nine possible genotypes (see Table 11.1).

An individual can possess 0, 1, 2, 3, or 4 “high” trait alleles. Assuming that the combined effects of the two loci are also additive², there are five distinct

TABLE 11.2 Genotypic Values of Two-Locus Genotypes

	AA	Aa	aa
BB	4	3	2
Bb	3	2	1
bb	2	1	0

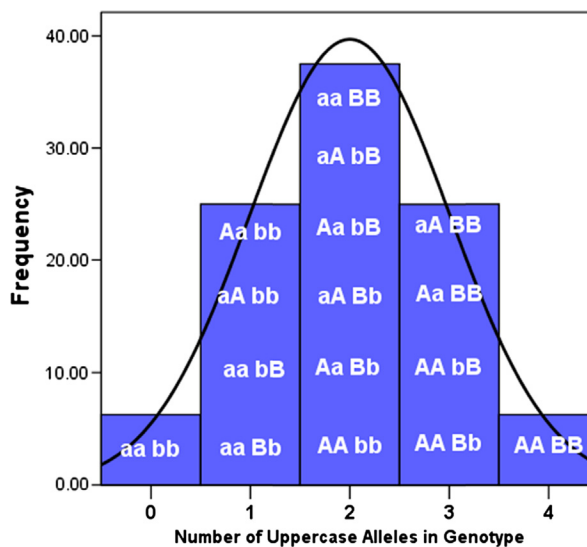


Figure 11.2 Expected phenotype distribution for a trait with two independently segregating causal loci of equal effect and allele frequency.

phenotypes with respect to the number of high trait alleles (see Table 11.2).

The trait distribution with respect to genotypic value distribution is shown in Fig. 11.2. As can be seen in Fig. 11.2, even with two loci, the distribution of the phenotype starts to look Gaussian. An example of a three-locus system with equal allele frequencies, no linkage disequilibrium (LD)³, and equal additive effects, is shown in Fig. 11.3. It can be seen that six diallelic loci are enough to produce population frequencies virtually indistinguishable from a normal curve.

Many traits (or diseases) are treated as dichotomous variables because they appear to be either present or absent (e.g., cancer). By definition, dichotomous variables do not

²That is, not *epistatic*, where epistatic refers to an interaction (in the statistical, not necessarily biochemical, sense) between two different loci, such that the effect of genotype at one locus depends on the genotype at another locus.

³*Linkage disequilibrium* is defined as the nonrandom association between alleles at *linked* (or adjacent) loci. Two loci are said to be linked if they are sufficiently close on the same chromosome such that they do not segregate independently.

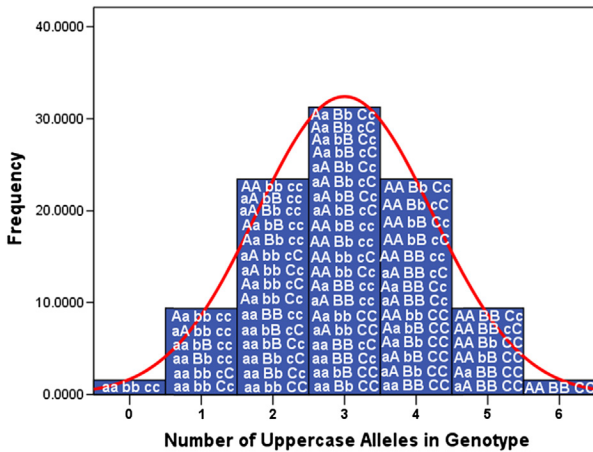


Figure 11.3 Expected phenotype distribution for a trait with three independently segregating causal loci of equal effect and allele frequency.

approximate a Gaussian distribution. These diseases may still be polygenic or multifactorial, because they do not follow the patterns expected of Mendelian (single-gene) diseases. A common explanation is that an underlying *liability* distribution exists for multifactorial diseases [3]. Individuals on the low end of the distribution have little chance of developing the disease because they possess few of the alleles or environmental factors that jointly cause the disease. By contrast, individuals on the high end of the liability distribution have a greater chance of developing the disease because they possess many of the alleles and/or environmental factors that jointly cause the disease. The liability distribution is assumed to be continuous (representing the sum of a large number of independent genetic and environmental factors) and normally distributed within the population. It is also commonplace to assume that all correlations between relatives are due to shared genes but not to shared environment. For multifactorial diseases that are either present or absent, there is a hypothesized *threshold of liability* that must be crossed before the disease is manifest [3].

For example, consider the development of the cleft palate. Early in embryonic development, the palatal arches are in a vertical position. Through embryonic and fetal development, the head grows larger, moving the arches farther apart, and the tongue increases in size, making it more difficult to move. Also, the arches themselves are growing and turning horizontally. There is a critical stage in development by which the two arches must meet and fuse. Head growth, tongue growth, and palatal arch

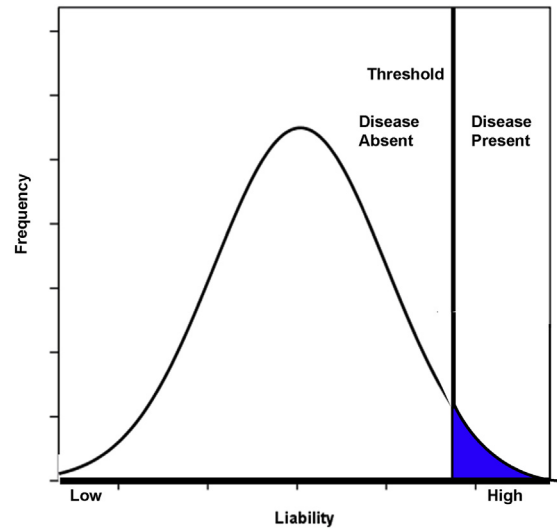


Figure 11.4 Liability distribution for a multifactorial disease. To be affected with the disease, an individual must exceed the threshold.

growth are all subject to many genetic and environmental factors. If the two arches start to grow in time, grow at the proper rate, and begin to move soon enough to the horizontal, they will meet and fuse despite head size and tongue growth. The result is no cleft palate. They may fuse well ahead of the critical developmental stage or just barely make it in time; it is impossible to know; however, if they do not meet by the critical stage, a cleft palate results. If they are close together at the critical stage, a small cleft will result, perhaps only a bifurcated uvula. If they are far apart, a more severe cleft will result. That critical difference in liability is called the *threshold*. Beyond the threshold, disease results; below the threshold, normal development is observed. Thus, the underlying liability is distributed as the normal curve shown in Fig. 11.4.

Some diseases may have more than one threshold, and commonly two liability thresholds are present, as defined by factors such as gender, race, age of onset, etc., causing different levels of severity [4]. Examples include pyloric stenosis (sex dimorphism for liability) [5] and orofacial cleft syndrome/cleft lip and palate (two thresholds for fetal mortality and disease) [6,7]. The latter model proposes a lower threshold level of liability resulting in a cleft formation and a higher level causing fetal death (predominantly in males).

It should be emphasized that, like any other loci, the individual loci underlying a polygenic or multifactorial trait are generally assumed to follow the Mendelian

principles of *random segregation* and *independent assortment*⁴. The difference is that they act together to influence the trait. Thus, the multifactorial model assumes:

1. several, but not an unlimited number, of loci are involved in the expression of the trait;
2. the loci act in concert in an additive fashion, each adding or subtracting a small amount from the phenotype; and
3. the environment interacts with the genotype to produce the final phenotype.

11.2 DETERMINING THE GENETIC COMPONENT OF A TRAIT

Historically, the genetic study of any trait can be divided into four broad categories: familial aggregation, segregation analysis, linkage analysis, and association studies. This paradigm was useful in discovering genes for many monogenic disorders.

11.2.1 Familial Aggregation

The first step of any genetic analysis is to establish a genetic component to the disease. Also, one must establish the relative size of the genetic effect in comparison with other sources of variation, such as common household effect and random environmental effect. *Familial aggregation* can be established using family-based or twin/adoption studies. Since family members share genes and environment, familial aggregation of the trait could be due to genetics and environment together. In general, very few traits are influenced only by genes or only by the environment. The detection and estimation of familial aggregation is a first step in the genetic analysis of any multifactorial trait. Twin and adoption studies are traditionally used to determine the genetic component of the trait [8–11]. Because monozygotic (MZ) twins share all their genes, any difference between them for a particular trait should be due solely to environmental effects. If the trait is completely influenced by genes, then MZ twins should have essentially identical expression of the trait; however, this is not true for dizygotic (DZ) twins because, on average, they share only 50% of their genes.

⁴Good descriptions of these principles can be found at <https://www.thoughtco.com/mendels-law-373515> and at <https://www.thoughtco.com/independent-assortment-373514>.

Twin studies to determine the genetic component of the threshold character are based on comparing *concordance rates* of MZ and DZ twins. If both members of a twin pair have the same status of a dichotomous trait (i.e., either both have the disease or both do not have the disease), they are *concordant*. If they do not share the trait status, they are *discordant*. The concordance rate is the proportion of concordant twin pairs among all those with the trait. Significantly higher concordance rates in MZ twins compared with DZ twins is considered evidence for a significant genetic component of the disease. The significance of the difference can be easily tested by a 2×2 contingency table using a χ^2 test (see Table 11.3).

Concordance rates are not appropriate for continuous traits, so correlation coefficients can be used instead of concordance rates [12].

For continuous traits, the familial aggregation is usually measured by *heritability*, the proportion of variability of the trait explained by genetic variation. Heritability can be defined either using total genetic effects (sum of additive, dominant, and epistatic effects) or using only additive effects. The former quantity is called *heritability in the broad sense* and is given by:

$$h^2 = \text{Var}(G) / \text{Var}(T)$$

where $\text{Var}(G)$ and $\text{Var}(T)$ are genetic and total variance, respectively. The latter quantity is called *heritability in the narrow sense* and is given by:

$$h^2 = \text{Var}(A) / \text{Var}(T)$$

where $\text{Var}(A)$ is the additive genetic variance.

TABLE 11.3 Using Twin Concordance and Discordance Rates to Test for a Genetic Component of a Disease

Twins	Concordant Pair	Discordant Pair	Total Pairs
Monozygotic (MZ)	n_{11}	n_{12}	n_{MZ}
Dizygotic (DZ)	n_{21}	n_{22}	n_{DZ}
	n_C	n_D	n
$\chi^2_1 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_C n_D n_{MZ} n_{DZ}}$			

One cannot conclude the number of genes or which genes are involved in the etiology of the trait from a heritability estimate. Although the absence of familial aggregation is generally thought to rule out a genetic contribution to the trait, there are some unlikely, yet plausible, scenarios in which this is not true. These include phenotypic competition within families [13] that counters genetic effects and an extreme form of epistasis referred to by some as emergence [14]. It is also important to emphasize that because heritability is a population-specific estimate, it can vary from population to population.

The method to determine the degree of genetic component of a continuous trait is based on a comparison of the variance of the differences between MZ twins and the differences between DZ twins. Since MZ twins share all their genes, the variance of the trait between MZ twins (V_{MZ}) must be due to environmental variance (V_E), so in this case we have $V_{MZ} = V_E$. The variance of the trait between the DZ twins (V_{DZ}) could be due to both environment (V_E) and shared genes (V_G). So, genetic variance is $V_G = V_{DZ} - V_{MZ}$, and therefore the heritability, h^2 , is defined as:

$$h^2 = \frac{V_{DZ} - V_{MZ}}{V_{DZ}}$$

Heritability ranges between 0 and 1, with 0 meaning a solely environmentally determined trait and 1 meaning a completely genetically determined trait.

Adoption studies provide a second familial aggregation strategy for estimating the influence of genes on multifactorial traits. The strategy consists of comparing disease rates among the adopted offspring of affected parents with the rates among adopted offspring of unaffected parents. Certain biases can influence these studies, namely, (1) parental environment could have long-lasting effects on an adopted child, (2) adoption agencies attempt to match the adoptive parents with the natural parents in terms of socioeconomic status, and (3) children might be several years old when adopted, introducing the potential for many environmental confounds. Moreover, these studies are reasonably good at estimating additive genetic effects that are not age-specific, but poor at estimating nonadditive genetic effects or genetic effects that are expressed differently across the age span.

There are many other methods to detect and estimate familial aggregation using family data. For example, the recurrence risk is often used to determine the strength

of familial aggregation for a discrete trait. The recurrence risk is the probability that a relative of an affected individual is also affected. The most commonly used measure is the sibling recurrence risk, i.e., the probability that a sibling of an affected individual is also affected. The ratio of the sibling recurrence risk and the overall disease prevalence is called a *sibling relative risk*. It is one of the measures of the magnitude of the genetic contribution to susceptibility for a dichotomous trait (affected vs. unaffected). Examination of relative recurrence risk values for various classes of relatives could suggest that the trait is influenced by multiple loci [15]. For a single ascertainment scheme, the sibling recurrence risk can be calculated from sibling data as follows [16]:

$$K_s = \frac{\sum_{s=1} \sum_{a=1} (a-1) n_{s(a)}}{\sum_{s=1} \sum_{a=1} (s-1) n_{s(a)}}$$

where a = number of affected sibs in a sibship, s = number of siblings in the sibship, and $n_{s(a)}$ = number of sibships of size s with a affected sibs.

Note that the aforementioned familial aggregation methods use only trait information from the sample. Owing to the availability of genome-wide single-nucleotide polymorphisms (SNP) data, it is now feasible to calculate the heritability using genome-wide SNP markers. One such method was proposed by Visscher et al. [17], who used genome-wide identity of descent (IBD) sharing probability between full sibs using genome-wide SNP data.

11.2.2 Segregation Analysis

Once a genetic basis of the trait has been established, the next step has traditionally been to determine the genetic models that explain the segregation of a phenotype (continuous, dichotomous, or ordinal) in a given familial data set via segregation analysis. Segregation analysis requires phenotypic data on related individuals and does not require any molecular data. Segregation analysis is the statistical methodology to determine whether a model with one or more major genes and/or polygenes (i.e., a set of genes, each with a small quantitative effect, that together produce a phenotype) is consistent with the observed pattern of phenotypic inheritance, and to estimate the parameters of the best-fitting genetic model. It entails determining the mode of inheritance (additive, recessive, or dominant), estimating “disease” allele frequency, and

estimating penetrance (probability of being affected given genotype). At one time, segregation analysis was one of the most important tools for genetic analysis of familial data. In the late 1980s, large numbers of DNA markers became available, thus rendering segregation analysis less popular.

If the trait is monogenic and thus due to a single major gene effect, segregation analysis has proven to be a very effective tool in determining the parameters for mode of inheritance. Subsequently these parameters have been used in model-based linkage analysis (see later for more detail) to find the location of putative disease-causing genes. This paradigm has been used successfully for the simple Mendelian traits, in which only one gene is segregating. For multifactorial traits, which may be due to the effects of many genes and environmental effects, estimation of the genetic model may be virtually impossible using segregation analysis.

To determine the parameters of the genetic model using segregation analysis, the likelihood of a particular mode of inheritance can be formulated using three types of probability functions. First, there is the probability distribution for segregation of genotypes among the founders (individuals whose parents are not included in the observed pedigree data), where genotypes of the founders are independently drawn from the population based on the prevalence of the disease and mode of inheritance. Second, there is the probability distribution of the segregating genotypes of the nonfounders (individuals with both parents in the pedigree) conditional on their parental genotypes. Third, there are penetrance functions (probability of being affected given a particular genotype). To test whether there is a segregation of a single gene, the likelihood under the assumed genetic model is compared with the likelihood under the null model of segregation with no genetic effect. The more complicated or general model could be included for testing a particular mode of inheritance, including polygenic or multifactorial components in modeling the disease; however, the number of possible genetic models with a given mode of inheritance may be too large to make any meaningful inference about the disease model.

11.2.3 Linkage Analysis

Genetic linkage analysis is based on the observation that any two loci that are close to each other on the same chromosome tend to cosegregate among related individuals

more often than two random loci in the genome. Thus, the affected individuals sharing a genotype at a putative disease locus would be more likely to share a genotype at linked marker loci. In the absence of linkage, the recombination fraction (i.e., θ = the proportion of gametes in which two genes on the same parental chromosome are separated; for more details, see chapter [Chromosomal Basis of Inheritance](#)), is $\frac{1}{2}$; however, if there is linkage, the recombination fraction is less than $\frac{1}{2}$. If we cover the entire human genome using evenly spaced markers across the chromosomes, it will be possible to find marker loci associated with a given trait of interest.

The methods of linkage analysis can be divided into two broad classes: *model-based* (parametric) and *model-free* (sometimes referred to as *nonparametric*) linkage analysis.

11.2.3.1 Model-Based Analysis

Model-based methods assume a specific genetic model underlying the trait, and the statistical evidence in favor of linkage with a marker locus is summarized by the maximum value of the *lod score* [18]. The lod score is the logarithm of the likelihood ratio of observing a particular set of family data under a specific alternative hypothesis of linkage relative to the null hypothesis of no linkage between disease and marker loci ($H_0: \theta = \frac{1}{2}$). For details of how the likelihood is formulated, see Elston and Stewart [19], Elston and Rao [20], and Lander and Green [21].

This approach has been remarkably successful in identifying disease genes for Mendelian disorders. To calculate such likelihood for families we must specify a probability model. Several assumptions are usually made in calculating lod scores: the mode of inheritance underlying the marker and the trait is known, the parameters such as penetrances and allele frequencies at both marker and trait loci are known without error, and all founders are unrelated to one another. Misspecification of any of these assumptions can affect the validity or power of the analysis and can result in an inconsistent estimate of the recombination fraction. Thus, the models used in model-based analysis must adequately approximate the complexity of the disease being investigated. It is noteworthy that incorrect specification of a legitimate model-based linkage test may reduce power but generally does not lead to an inflated type I error rate (false positive rate).

11.2.3.2 Model-Free Linkage Analysis

The genetic mechanism underlying a complex disease is often unknown, and it is impossible to specify the correct genetic parameters, such as mode of inheritance, disease allele frequency, and penetrance, in complex diseases. Under these circumstances, model-free linkage analysis, which makes no assumption about the mode of inheritance of the trait, is usually preferred. If a disease susceptibility locus and a marker locus are linked and, by definition, cosegregating in a family, pairs of relatives who are concordant for the disease (i.e., both affected or neither affected) should share more alleles identical-by-descent (IBD⁵) than an average pair of relatives with the same degree of kinship. Similarly, discordant (affected–unaffected) relative pairs should share fewer alleles IBD than an average pair of relatives with the same degree of kinship at the disease locus or marker linked to the disease locus. Model-free methods were first derived for sibpairs [22,23], but were extended to other relative pairs [24,25]. The Haseman–Elston method consists of regressing the squared phenotypic difference among siblings within sibpairs on the estimated proportion of alleles the sibs share IBD at a marker locus. A negative slope suggests linkage because it indicates that greater similarity at a trait locus tends to occur with greater similarity at a marker locus. There have been a number of extensions to the Haseman–Elston method to increase its power [26–35]. In 2003, *Human Heredity* published a special topic issue titled “Recent Advances in the Analysis of Genetic Traits” celebrating the 30th anniversary of the seminal paper by Haseman and Elston for quantitative trait linkage analysis (*Human Heredity* 2003;55:2–3). Several software programs are available to perform linkage analyses, including S.A.G.E. ([36], <http://darwin.cwru.edu/sage/>); GENEHUNTER ([37], <http://archive.broadinstitute.org/ftp/distribution/software/genehunter/>); Mx (<https://mx.vcu.edu/>); MERLIN ([38], <http://csg.sph.umich.edu/abecasis/merlin/download/>); and FASTLINK (<https://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html>). A complete list of available linkage programs can be accessed through

the OMIC Tools website (<https://omictools.com/linkage-analysis-category>).

The aforementioned model-free methods mostly utilize relative pairs such as sibpairs. However, large extended families could provide more linkage information than these relative pairs. Methods based on the variance components framework have become a popular choice for linkage analysis because of the ease of modeling covariates and gene×gene or gene×environment interactions. These methods can also utilize large extended families in the model-free environment [39–43]. There are several programs available to perform linkage analyses based on variance components methodology, such as ACT [40], SOLAR [39], and Mx. Mx uses a structural equation modeling approach, which is equivalent to the other types of variance components approaches under most circumstances. There are many more statistical genetics packages freely available for public use, and a complete list along with download information for each package can be found at <http://gaow.github.io/genetic-analysis-software/> and at <http://www.soph.uab.edu/ssg/linkage/lddac>.

11.2.4 Transmission Disequilibrium Test and Association Analysis Using Familial Data

In classical genetic studies, the identification of a chromosomal region with linkage analysis is the first step in the gene mapping process. Since linkage analysis provides information of a genomic region, a typical quantitative trait locus (QTL) may cover several millions of basepairs and may contain hundreds of genes. The initial detection of the QTL is followed by addition of more markers within the QTL to narrow down the region as much as possible. Once the resolution limit of the linkage approach is reached, the most commonly employed follow-up is to fine map using association analyses with SNPs.

The rationale for association analyses is to confirm the involvement of a putative allele involved in a trait of interest. The rationale of fine mapping is that the greater number of SNPs and the greater sensitivity of the association tests provide more detailed information of the target region. The SNPs located close to a disease locus may cosegregate because of LD, i.e., allelic association due to linkage. The allelic association forms the theoretical basis for association mapping. Allelic association that is not due to LD is of no interest in mapping disease genes. The simplest way to test for association is to perform a case–control study, in which the cases are the individuals with the disease and the controls are without the disease.

⁵“Identity of descent (IBD) alleles in an individual or in two people are defined as alleles that are identical because they have both been inherited from the same common ancestor, as opposed to identity by state” (<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hmg>).

Association is then tested by ascertaining whether a particular marker allele is more frequent among the cases than the controls. A significant result will be observed if the marker is in LD with the disease locus, or from a variety of confounding reasons such as population stratification. Case-control association studies done without controlling for stratification are prone to false positive results with no biological significance. For this reason, association studies were not popular until the mid-1990s, when methods to account for stratification were established using familial data [44]. To control for population stratification, Spielman et al. [44] proposed the *transmission disequilibrium test* (TDT). TDT is a family-based association test in the presence of linkage that controls for population stratification by comparing the allele frequencies among alleles transmitted to an affected offspring with those that are not transmitted to an affected offspring from informative parental matings (i.e., matings with at least one heterozygous parent). This study design requires the collection of family trios that include two parents and an affected offspring. More than 225 extensions and variations of the original TDT have been proposed (see the exhaustive review of TDT procedures in Tiwari et al. [45]). There are a number of software programs available for TDT and/or association analyses using family data, such as FBAT (family-based association test; <http://www.biostat.harvard.edu/~fbat/fbat.htm>), ASSOC (<http://darwin.cwru.edu/sage/>), and GASSOC (<http://www.mayo.edu/research/labs/statistical-genetics-genetic-epidemiology/software>). A complete list of association programs can be found at <http://gaow.github.io/genetic-analysis-software/> and <http://www.soph.uab.edu/ssg/linkage/associationtdt>.

Once the results from the association analyses are deemed adequate, the next step is to screen the candidate genes for DNA sequence variation by direct sequencing. The relevance of the detected mutations is confirmed with additional association studies in the original and other populations, as well as functional assays in vitro (expression studies in different cell lines) and in vivo (transgenic and knockout animal models) [46].

11.3 THE INTERNATIONAL HAPMAP PROJECT

In the context presented earlier, studies progress from estimates of heritability, to segregation analysis, to linkage, and then finally to familial association analysis to

determine candidate genes for a trait of interest. However, recently this paradigm has changed. With the advent of high-dimensional genotyping technologies using microarrays, the approach for discovering new genetic variants of a disease or trait has changed drastically. In 1996, Lander proposed the “common disease, common variant” (CDCV) hypothesis [47]. The CDCV is based on the idea that the genetic component of common diseases is attributable in part to common allelic variants (i.e., alleles with frequency at least 5%). The HapMap project was initiated to create a dense set of genetic markers to test the CDCV hypothesis. The draft of the complete human genome sequence was completed in 2001 and had a strong effect on advances in genome sequencing technology [48]. The International HapMap Project was an international partnership that was formed in 2002 to help researchers find genes associated with human disease by providing a public database of common genome-wide human variation across populations [49–51]. The first stage of the HapMap project focused on four diverse populations: 30 trios (two parents and one adult child) from the Yoruba people in Ibadan, Nigeria; 30 trios from the Centre d’Etude du Polymorphisme Humain (CEPH) collection of Utah residents of northern and western European ancestry, 45 unrelated individuals from the Han Chinese in Beijing, and 45 unrelated individuals from the Japanese in Tokyo. This project genotyped over 1 million SNPs in phase I and an additional 2.1 million in phase II in the HapMap samples [50,51]. It helped initiate advances in SNP array technologies to make genome-wide association studies (GWAS) feasible and affordable. Affymetrix and Illumina SNP arrays became available to researchers, who initially surveyed approximately 100,000 SNPs, and who now have surveyed 2.5 million SNPs. During the most recent phase, HapMap 3, 1184 individuals representing 11 global populations were genotyped for approximately 1.6 million common SNPs [52] (<http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html>).

As a complement, the 1000 Genomes Project, which ran from 2008 to 2015, was initiated to provide a catalog of low-frequency SNPs and structural and sequence variants in the human genome [53]. The final data set contains information for 2504 individuals from 26 populations (over 88 million variants), using a combination of low-coverage whole-genome sequencing, dense microarray genotyping, and deep exome sequencing.

The results have furthered the understanding of the processes that shape genetic diversity and disease biology, while enabling genotype imputation, array design, and cataloging of variants [54–57]. The International Genome Sample Resource (<http://www.internationalgenome.org>) provides ongoing support for the 1000 Genomes Project data and incorporates published genomic data, including those from new populations.

11.4 GENOME-WIDE ASSOCIATION STUDIES

GWAS are an approach that involves scanning thousands to a few million SNPs across the whole genome on many individuals to find association with a disease or trait. As mentioned earlier, GWAS became a popular choice of genetic studies to detect putative loci associated with a disease or trait because of the availability of high-throughput SNP arrays, decreased cost of genotyping, and methods to correct for population stratification (i.e., systematic differences in allele frequencies between subpopulations in a given population possibly due to different ancestry). Before the HapMap era, investigators were reluctant to conduct association studies using population data because of concerns about population stratification. For example, in case–control studies, we usually test association of a particular SNP by comparing allele frequency between cases and controls. Allele frequencies are known to vary within and between populations depending on genetic ancestry [50,58]. Genetic ancestry becomes a confounding variable leading to spurious associations if allele frequencies are different within or between race/ethnic groups. Methods for correcting population substructure are described later.

11.4.1 Study Designs

Any type of data set, such as pedigree data, case–control data, or population data, are all appropriate choices for GWAS. Analysis must adjust for familial correlations in pedigree data and population stratification in population or case–control data sets to control for the confounding due to relatedness or population substructure. Case–control and population data have been commonly used for GWAS because of their availability and convenience of ascertainment; however, there are some issues associated with the case–control design. If the disease is heterogeneous, extra attention should be paid to minimizing heterogeneity in case selection, e.g., selecting the

most extreme cases or selecting individuals from a familial disease cohort. There has been controversy regarding the selection of optimal controls. Usually, controls from the same population and residing in the same geographic area are preferred, but can be difficult to ascertain. The Wellcome Trust Case Control Consortium used 3000 UK controls and 2000 cases from each of seven different diseases to show that using common controls was effective, had minimal effect on genotypic distributions, and did not lead to excess false positives [59,60]; however, misclassification error in control selection could affect the power of the association analysis. Specifically, this is true for late-onset diseases because controls have not yet reached the age to develop the disease. This issue can be resolved by increasing the sample size [59]. Population stratification and cryptic relatedness (i.e., relatedness among individuals in the study that is not known to the investigator) can also increase the false positive findings, as previously discussed. The family-based association studies are robust to population stratification, but it is difficult to ascertain all pedigree members, which leads to missing data within families and loss of power compared with case–control designs [61]. There are some other issues with study design selection, and an excellent review is provided by McCarthy et al. [59].

11.4.2 Quality Control

The first step of GWAS analysis is the quality control (QC) of the genotypic and phenotypic data. There are a number of procedures needed to ensure the quality of genotype data both at the genotyping laboratory and after calling genotypes using statistical approaches. Here we will assume that the genotyping laboratory has used best practices to remove technical variation, and we present only statistical methods that are used after completion of the genotyping. The QC and association analysis of GWAS data can be performed using the robust, freely available, and open-source software PLINK developed by Purcell et al. [62]. Anderson et al. provide step-by-step PLINK and R commands to implement most of the procedures [63]. In addition, two publications provide excellent reviews of the QC protocol for GWAS data [64,65]. Here, we provide a few important steps of the QC in GWAS using guidelines similar to those in Laurie et al. [64] and Turner et al. [65]. Note that the current genotyping technology is very reliable, but there are still some possibilities of errors when genotyping large number of SNPs.

11.4.3 Sex Inconsistency

It is possible that self-reported sex of the individual is incorrect. Sex inconsistency can be checked by comparing the reported sex of each individual with predicted sex by using X-chromosome markers' heterozygosity to determine the sex of the individual empirically.

11.4.4 Relatedness and Mendelian Errors

Another kind of error that can occur in genotyping is due to sample mix-up, cryptic relatedness, duplications, and pedigree errors, such as self-reported relationships that are not accurate. To detect sample relatedness, one can calculate three IBD probabilities of sharing 0, 1, and 2 alleles that are IBD for each pair of individuals using software such as PLINK and a kinship coefficient matrix. Individuals sharing zero alleles at every locus are unrelated, individuals sharing one allele IBD at every locus are parent-offspring pairs, individuals sharing two alleles IBD at every locus are MZ twins or a duplicated sample, and on average sibpairs share zero, one, and two alleles IBD with sharing probabilities 0.25, 0.5, and 0.25, respectively. The relationship errors can be corrected by consulting with the self-reported relationships and/or using inferred genetic relationships. Cryptic relatedness can inflate the variance of the test statistic (e.g., if the test statistic is the difference in the overall allele counts between case and control samples in a trend test [66]). The presence of cryptic relatedness in case-control studies increases the false positives in association analysis. Devlin and Roeder provided a method to correct for the variance inflation (see [66] and [67], for details).

11.4.5 Batch Effects

For GWAS, samples are processed together for genotyping in a batch. The size and composition of the sample batch depends on the type of the commercial array; for example, an Affymetrix array can genotype up to 96 samples, and an Illumina array can genotype up to 24 samples. To minimize batch effects, samples with different phenotypes, sex, race, and ethnicity should be randomly assigned to plates. The downstream association study can be confounded by the batch effects. There are several methods available to detect any batch effects. The most commonly used method is to compare the average minor allele frequencies and average genotyping call rates across all SNPs for each plate. Most genotyping laboratories perform batch effect detection and usually

regentype the data if there is a batch effect or a large amount of missing data.

11.4.6 Marker and Sample Genotyping Efficiency or Call Rate

Marker genotyping efficiency is defined as the proportion of samples with a genotype call for each marker. If large numbers of samples are not called for a particular marker, that is an indication of a poor assay, and the marker should be removed from further analysis. The threshold for removing markers varies from study to study depending on the sample size of the study. Usual recommended call rates are approximately 98%–99%. If the quality of the DNA sample is poor, it leads to a low call rate of genotypes for the individual; i.e., the number of missing genotypes will be large and the sample should be excluded from further analysis. Before performing the association analysis, one should filter out the samples and markers using some threshold for marker and sample call rates.

11.4.7 Population Stratification

There are a number of methods proposed to correct for population substructure. Three commonly used methods to correct for the underlying variation in allele frequencies that induces confounding include genomic control [4,66–74], structured association testing [75–77], and principal components (PCs) [78,79]. The genomic control method estimates an inflation factor (ratio of the variance of the test statistic and the variance under the null hypothesis) and adjusts the test statistics for all markers in GWAS downward by the inflation factor. Usually, the inflation factor is calculated using a few hundred loci. Structure association testing [75,76] estimates the ancestry proportions of each individual from the founding population using markers with different allele frequencies in the founder population and then uses these proportions to cluster individuals to create homogeneous groups with similar ancestry profiles for the association analysis. Principal components analysis (PCA) uses thousands of markers to detect population stratification with a program such as Eigenstrat ([78,79], <https://data.broadinstitute.org/alkesgroup/EIGENSOFT/>). The PCs are entered as covariates into the association model [78,79]. There are two issues with using PCA: how many SNPs to use and how many PCs should be included as covariates in the association analysis.

11.4.8 Marker Allele Frequency and Hardy–Weinberg Equilibrium Filter

The Hardy–Weinberg equilibrium (HWE) test compares the observed genotypic proportion at the marker versus the expected proportion. Deviation from HWE at a marker locus can be due to population stratification, inbreeding, selection, nonrandom mating, genotyping error, actual association to the disease or trait under study, or a deletion or duplication polymorphism. However, HWE is typically used to detect genotyping errors. SNPs that do not meet HWE at a certain threshold of significance are usually excluded from further association analysis. It is also important to discard SNPs based on minor allele frequency (MAF). Most GWAS are powered to detect a disease association with common SNPs ($MAF \geq 0.05$). The rare SNPs may lead to spurious results due to the small number of homozygotes for the minor allele, genotyping errors, or population stratification.

11.5 IMPUTATION

Genotype imputation is the process of predicting genotypes that are not directly assayed in a sample of individuals by using a reference panel of haplotypes. Imputation methods work by identifying sharing between the reference haplotypes and the underlying haplotypes of the unrelated study subjects using local IBD patterns. Imputation provides finemapping of genomic regions, which increases the ability (power) to find causal SNPs, especially for harder-to-tag rare SNPs. Imputed SNPs that exhibit large associations may become candidates for replication studies. In meta-analyses, when different genotyping chips are used for different cohorts, imputation can equate the set of SNPs across studies. Imputation can also correct genotyping errors and extend to non-SNP variation (i.e., copy number variants, insertions/deletions). Owing to uncertainty in the imputation process, a probability distribution is calculated over all three genotypes, and this should be incorporated into any downstream analysis. The study population, properties of the reference panel and genotyping chip, and allele frequencies will affect genotyping accuracy. Error rates increase as MAF decreases because of the difficulty in tagging rare SNPs. Using a combination of populations can increase accuracy, especially at rare SNPs. Error rates also decrease as reference panel size increases [80,81]. Use of the 1000 Genomes Project

reference panel instead of HapMap Project data has been shown to improve imputation accuracy, increase the strength of known associations, and identify novel loci [82,83]; however, not all of the variants present in HapMap data are in 1000 Genomes data [84]. Therefore, the choice of reference panel is an important consideration. Formatted reference panels can be downloaded from software websites.

Li et al. [80] and Marchini and Howie [81] provide excellent reviews of different imputation methods and software. Major programs include IMPUTE2 ([85,86], http://mathgen.stats.ox.ac.uk/impute/impute_v2.html), MaCH ([87], <http://csg.sph.umich.edu/abecasis/MaCH/download/>), Minimac([88,89], <http://genome.sph.umich.edu/wiki/Minimac#Reference>), and BEAGLE ([90,91], <https://faculty.washington.edu/browning/beagle/beagle.html>). The first step of any imputation process should be standard QC of sample data. Next, the data need to be in the correct format (i.e., MERLIN pedigree and data files, one per chromosome) and on the appropriate build (the latest 1000 Genomes reference panel uses NCBI Build 37/HG 19). The University of California at Santa Cruz (UCSC) provides an online liftOver tool for converting genome positions between builds. Markers should be stored by chromosome position, in the forward strand, and encoded as “A”, “C”, “G”, or “T”. Minimac imputation uses a two-step approach. First, the samples must be phased into a series of estimated haplotypes (i.e., using MaCH). Second, Minimac performs imputation with these phased haplotypes.

Since the imputation procedure can take weeks to complete, time savings can be achieved by imputing chromosomes in chunks. The ChunkChromosome tool can be used to divide each chromosome into smaller chunks (i.e., 2500-marker chunks, with 500-marker overhang) (<http://genome.sph.umich.edu/wiki/ChunkChromosome>). After imputation has been completed, imputation quality evaluation should be performed to remove SNPs with poor quality. Different programs provide different-quality statistics, which are not directly comparable. Minimac provides three helpful measurements: (1) looRSQ, the estimated R^2 for each SNP; (2) empRSQ, the true R^2 comparing imputed and actual genotypes; and (3) empR, the empirical correlation between actual and imputed genotypes (note that negative values indicate the alleles are most likely flipped). The size of the final sample (genotyped + imputed

SNPs) depends on the number of SNPs that pass pre-determined thresholds [88]. Tutorials to impute 1000 Genomes SNPs using Minimac or IMPUTE2 can be found at http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook and http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook, respectively.

11.6 ASSOCIATION METHODS/ STATISTICAL ANALYSIS

11.6.1 Power and Sample Size Calculations

Statistical power is the probability of rejecting the null hypothesis of no association when a true association is present. Power calculation requires a specified effect size and variance, sample size, and significance level (type I error, i.e., $\alpha=0.05$). A sample size that ensures sufficient power (i.e., 80% power) is critical to the detection of causal variants in complex diseases. GWAS generally require larger sample sizes because association analyses may compare millions of SNPs, increasing false positive rates. In addition, causal SNPs may have small effect sizes that are difficult to find in small samples [92]. Since Mendelian diseases arise from single-gene mutations, the causal mutations have an enormous impact on disease risk, and the large effects can frequently be detected with modest sample sizes. In contrast, the genetic structure underlying common diseases involves multiple risk loci and environmental factors. Doubling the number of subjects generally doubles the number of SNPs passing the 5×10^{-8} significance threshold, but the rate of increase and the minimum sample size depend on the disease complexity. In Crohn disease and ulcerative colitis, increasing beyond ~5000 cases was required to see a doubling effect, whereas in type 2 diabetes and breast cancer, the number was ~30,000 [93,94].

The power to detect significant associations at the *P*-value threshold is affected by allele frequency, LD, inheritance model, disease prevalence, effect size, and phenotype variation [95–97]. Although most of these are inherent characteristics of the study population, there are factors the investigator can change to maximize power, such as choice of study subjects and methods of phenotype and genotype measurements, data QC, and statistical analyses. In population cohort studies, using a liability threshold model (in which significance is affected by the proportion of variance in liability

explained by the locus) can simplify calculations. In case-control studies, it is easier to use an assumed odds ratio and the allele frequency of the putative risk variant. Since many GWAS rely on LD between typed and untyped variants to increase coverage, the extent of LD has a large effect on power. When a proxy SNP that has correlation R with the causal SNP is used, the sample size required to obtain the same amount of power is increased by a factor of $\sim 1/R^2$ [94,95]. Allele frequency is an important consideration because (1) locus allele frequencies determine heterozygosity, which is directly proportional to the phenotypic variance explained by the locus, and (2) the SNPs chosen for commercial SNP arrays represent common variations; rare variants are less likely to have a high correlation (R^2) with included SNPs. These factors are relevant when imputation is used because the statistical power for untyped SNPs depends on imputation quality, which is determined by the level of LD between typed and untyped SNPs. Longitudinal studies or multivariate outcomes can help increase power in studies of phenotypes that are not distinctive, rare, stable, and/or highly familial (i.e., hypertension and depression). Assuming the correct effect model is also important since, for example, a test that assumes additive effects has more power than one that assumes dominant effects, if the true effects are indeed additive. Last, appropriate data QC, inclusion of covariates (except perhaps in logistic regression analysis of case-control studies) [98], and correction for confounding factors will increase statistical power [94]. Sham and Purcell provide an excellent review of statistical power and significance testing in large-scale genetic studies, including more detail about Mendelian and complex diseases, tests for rare variants, and exome-sequencing studies [94].

Owing to the complicated nature of LD in the human genome, analytical power calculations are difficult and simulations are frequently used [99]. In situations in which simulations require extensive time and computational resources, Tiwari et al. provided a quick method for power estimation that is applicable to a wide range of genomic studies [100]. This method is based on Elston's Excellent Estimator formula and uses data similar to a user's data, but differs in sample size and/or α level ([100]; download at <http://www.ssg.uab.edu/eee-power/>).

There are several freely available GWAS power analysis programs. Genetic Power Calculator provides power analysis for common tests, including variance

components QTL linkage and association tests, TDT, case-control designs, and sibships [101]. The GAS and CaTS power calculators are for one- or two-stage GWAS [102]. PGA performs power analysis for case-control association studies of candidate genes, fine-mapping studies, and whole-genome scans [103]. R packages include *gap* (genetic analysis package for population- and family-based designs) [104] and *powerpkg* (power analyses for the affected sib pair and the TDT design) [105], download links for these programs at <http://gwa-testdriver.ssg.uab.edu/software.jsp>. *GWApower* is another R package that calculates power for GWAS using commercially available genotyping chips and simulation results generated from the HAPGEN program [99], <http://www.well.ox.ac.uk/software>. Lee et al. provided derivation of sample size/power for rare variants [106], <https://www.hsph.harvard.edu/skat/download/>. GCTA (genome-wide complex trait analysis) provides power calculation for estimating genetic variance or correlation using genome-wide SNPs in GREML analysis [107,108], <http://cnsgenomics.com/shiny/gctaPower/>. Finally, QUANTO may be used for genetic epidemiology studies [109,110], <http://biostats.usc.edu/Quant.html>.

11.6.2 Discovery Phase of the Genome-Wide Association Study

The choice of the statistical test for association in discovery phase depends on the study design and the phenotype under consideration. In a case-control design, the goal is to compare the allele or genotypic frequencies between cases (affected) and controls (normals). This can be tested with Pearson's χ^2 test, Fisher's exact test, or the Cochran-Armitage test. Pearson's χ^2 tests the null hypothesis of no association between rows and columns of the 2×3 contingency table consisting of the counts of the three genotypes among cases and controls [111]. Fisher's exact test is similar to Pearson's test, but the deviation from the null hypothesis is calculated exactly from all possible permutations of the data and, thus, does not assume the asymptotic χ^2 distribution [112]. The Cochran-Armitage test for trend is a test of proportions of cases versus controls [113–115] and assumes an additive mode of inheritance that is a linear trend. However, there is a loss of power if the trend is not linear. Freidlin et al. recommended using a maximum of the test statistics obtained from additive, dominant, or recessive effects models [116]. Note that in these

statistical procedures, one cannot model covariates such as sex, age, race, age of onset, PCs (from admixture), etc. To accommodate any relevant covariates in the analysis, one can use logistic regression. Logistic regression is more flexible in that it can model covariates, multiple SNPs as main effects, SNP-by-SNP interactions, SNP-by-environment interactions, etc. If the phenotype is continuous, analysis of variance (ANOVA) and general linear model approaches can be employed. One can also use a linear regression framework if extremes of the distribution are used to define case and control status. Huang and Lin have given an efficient association method using extreme phenotypes [117].

The analysis of familial data requires correcting for the dependency of observations. The notable methods include linear mixed model [118–126], *GEMMA* (genome-wide efficient mixed model analysis for association studies [122,123], <http://www.xzlab.org/software.html>), FBAT (see review by Laird and Lange [124]), ASSOC (a module of the S.A.G.E. software suite [36]), the R packages *GenABEL* (GWA analysis for quantitative, binary, and time-to-event traits [118]) and *ProbABEL* (GWA analysis of imputed genetic data [121], <http://www.genabel.org/packages/ProbABEL>), *lme4* [125], and *GWAF* [126]. After scanning ~ 2 – 9 million imputed SNPs (i.e., ~ 2 – 9 million statistical tests) to determine significant associations, appropriate multiple testing correction is required to control for false positives and in choosing SNPs for follow-up studies. The guideline for significant association for GWAS is generally a *P* value of $\sim 5 \times 10^{-8}$ [127]. However, it is common practice to use a higher *P* value threshold for follow-up study or replication. Balding [128] provides a comprehensive discussion of the advantages and disadvantages of these methods pertaining to GWAS.

11.6.3 Postanalysis Quality Control

After association testing has been done, QQ plots, which plot observed *P* values versus *P* values expected by chance, should be used to determine if there are still issues with the data. For example, there may still be confounding factors (i.e., population stratification, batch effects, or other systematic group differences) that are creating falsepositive results. Traditionally, $-\log_{10}$ (*P* values) are used to spread the points out and facilitate detection of unusual results. The *inflation factor*, λ , measures the deviation of the observed results from expected results. If λ is greater than 1, the QQ plot is

inflated, most likely from residual confounding factors. If λ is less than 1, the QQ plot is deflated, most likely from overcorrection in the association analysis (i.e., too many covariates). In either case, test statistics should be adjusted. If most of the *P* values occur on the identity ($y=x$) line, with some falling above, the results indicate that these SNPs may have significant associations with the phenotype. Next, Manhattan plots can be used to display small *P* values by genomic position [129,130]. The interactive program ManhattanPlotter (<http://www.biologiaevolutiva.org/~cmorcillo/tools/ManhattanPlotter/ManhattanPlotter.htm>) and the R packages *Manhattanly* and *qqman* provide functions to create QQ and Manhattan plots [131,132].

LocusZoom plots enable you to focus on a region and view the strength of association signals, local LD, recombination patterns, and nearby genes (Fig. 11.5). LocusZoom is a program that uses LD information from HapMap phase II or 1000 Genomes and gene information from the UCSC browser ([133], <https://statgen.sph.umich.edu/locuszoom/download/>). LocusExplorer is another program useful for visualization and exploration of genetic association data ([134], <http://www.oncogenetics.icr.ac.uk/LocusExplorer/>). Last, cluster plots are used to examine SNPs with significant departure from HWE or a high Mendelian error rate. Well-defined clusters indicate

high-quality SNPs, whereas poorly defined clusters may indicate a poor call rate or other genotyping issue. The Bioconductor package *GWASTools* has functions that execute most of the GWAS QC procedures described here [135,136].

11.6.4 Validation and Replication Phase

Some investigators have recommended reanalysis of the original discovery phase GWAS data using a different genotype platform for validation, which has been termed “technical validation” [59]. Technical validation allows the detection of technical errors in genotyping that might give rise to spurious association signals or false positives; however, given the limitations of the resources available to investigators, it may not be feasible. The replication phase or follow-up study is one of the most challenging aspects of the GWAS and is required to control for false positives. Replication in an independent data set with similar genetic background and phenotype is warranted. Usually several hundred or a few thousand SNPs are tested in a replication set depending on the threshold used for significant association *P* value. The statistical methods are the same as in the discovery phase depending on the study design and type of phenotype. SNPs for replication could be selected based on several criteria, including strength of

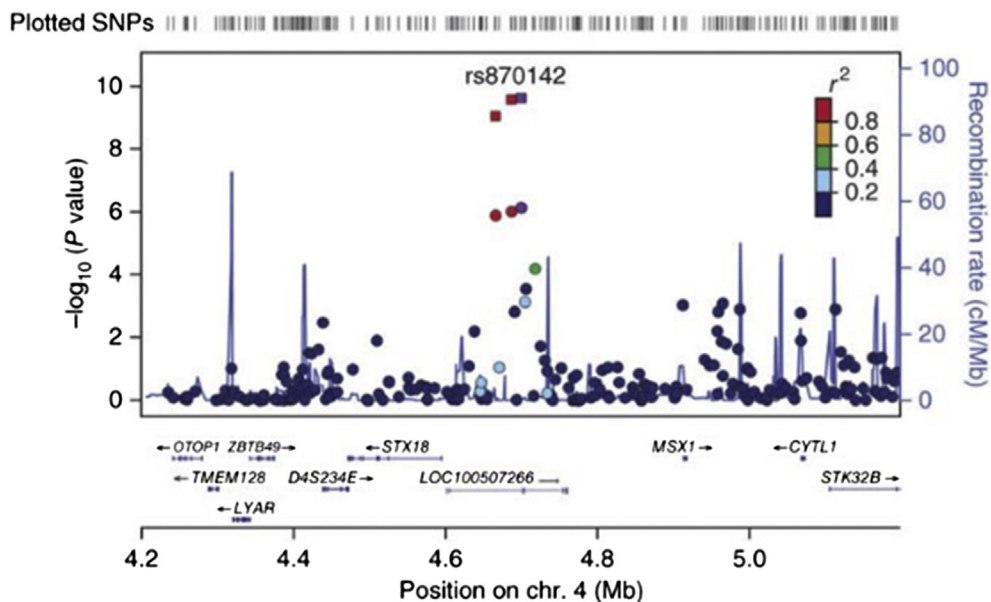


Figure 11.5 LocusZoom plot displaying the HDL cholesterol-associated region near the *MMAB* gene. (Taken from Kathiresan et al. Nat Genet 2009;41(1):56–65.)

association in GWAS (e.g., $P < 10^{-8}$), size of potential regions and/or candidate genes involved, degree of LD in these regions, presence of coding SNPs of reasonable frequency ($>1\%$), potential biological relevance to the disease/trait and previously identified association with disease/trait with lower threshold (with $P < 10^{-6}$), significant SNPs close to transcription binding sites or close to microRNAs with lower threshold (with $P < 10^{-6}$), and SNPs in disease/trait relevant pathways. Note that the P value thresholds as given for selecting potential SNPs for replication are arbitrary and should be decided by an investigator(s) depending on how many SNPs are to be replicated.

11.7 ANALYSIS OF RARE VARIANTS USING NEW TECHNOLOGIES

Introduction of the HapMap project and large-scale GWAS studies was driven by the CDCV hypothesis, which was first introduced in the 1990s [47,137]. The CDCV is based on the idea that the genetic components of common diseases are attributable in part to allelic variants that are present in more than 5% of the population. An extension of this hypothesis is that the same variants will be responsible for the disease across multiple populations [138]. The early success of GWAS (i.e., in age-related macular degeneration) seemed to support the theory that a large proportion of the genetic variants underlying complex disease could be explained by the CDCV. It is now becoming apparent that many common variants confer only a small portion of risk individually and explain only a small portion of heritability of common complex diseases [139]. While GWAS have been successful in many ways, identifying hundreds of variants for a large number of traits (<http://www.ebi.ac.uk/gwas>), there still remains a large proportion of heritability that has yet to be explained.

When the CDCV hypothesis was first introduced, it was not without contention [140]. One of the strongest counterarguments was based on the hypothesis of “common disease, rare variant,” which is in essence the antithesis of the CDCV hypothesis [80,141,142]. The rare variant hypothesis proposes that common complex diseases are due to the combined effects of multiple rare variants with moderate to low individual risk. Unlike CDCV, it is generally thought that, due to population history, these rare variants will be population-specific [143]. It is only recently, with the availability

of affordable large-scale sequencing technology and advances in analytical methods (discussed later), that scientists have gained the ability to address the role of CDRV in human disease. It is likely that the genetic basis of complex disease is somewhere between the two extremes, with multiple genes interacting together with a variety of common and rare variants and other genetic and environmental factors [144].

As new high-throughput, massively parallel sequencing technologies emerged in 2005, direct sequencing became commonly used to directly interrogate whole genomic sequences for association with disease without prior specification of SNPs currently available on commercial SNP chips [145]. Such technologies overcome some of the shortcomings of GWAS methods, such as ascertainment bias in the set of currently available SNPs and the ability to assay rare or private variants. In addition, greater flexibility exists in the search for variants other than SNPs, such as copy number variants; insertions or deletions, or indels; inversions, etc. Whole-exome sequencing, in which only the sequences of exons are assayed, has been used to discover causal mutations in a number of Mendelian disorders, such as Miller syndrome [146] and hereditary spastic paraparesis [147].

Because of the enormous number of variants introduced from new sequencing technologies, and the small sample sizes typically present, new bioinformatic and statistical methods have been developed to reduce the dimensionality and improve the probability of detection of causal variants. Prior bioinformatic processing may include filtering by IBD methods, if family data are present [148], or filtering based on the expected mode of inheritance [146]. In addition, if only rare variants are desired, then common variants can be filtered out using the SNP database dbSNP [149]. Predicted functional variants (nonsense, missense, splice site variants, indels, frameshift mutations, etc.) can be discerned using tools such as SIFT and PolyPhen [150,151]. Once likely nonfunctional variants have been filtered out, new statistical methods for summarizing the effects of multiple rare variants at a single gene can be applied. Some examples of these methods include the *cohort allelic sums test* method, which compares the number of individuals with mutations within a gene between cases and controls [152]; the *combined multivariate and collapsing* method, which collapses multiple rare variants in conjunction with common variants using multivariate analysis [153]; the *sequence kernel association test*, which

tests for association between common or rare variants and a continuous or dichotomous trait while adjusting for covariates [154]; methods that weight the counts of each variant using the estimated standard deviation of the total number of mutations [155,156]; or a method that models these weights in a flexible Bayesian framework [157]. A review of rare variant analysis methods is given in [158]. Whole-exome and whole-genome studies (in which the contribution of noncoding regions to disease can be assayed) are underway as of this writing for complex (multifactorial) diseases, and the next couple of years will show if these technologies can help to fill in the gaps from GWAS, termed “missing heritability” [139], in identifying causal variants underlying multifactorial diseases. In addition, new sequencing technologies offer opportunities for functional characterization studies, such as gene expression profiling using next-generation sequencing [159] and epigenetic profiling [160], and in identifying somatic mutations occurring in cancer [161–163].

11.8 STATISTICAL FINE MAPPING OF GENOME-WIDE ASSOCIATION STUDIES DATA SETS TO DETERMINE CAUSAL VARIANTS

Fine mapping to determine the causal variants using GWAS has been investigated by integrating functional annotations with GWAS SNPs in a Bayesian framework [164,165]. Statistical fine mapping involves several steps. (1) Select all SNPs with LD ($r^2 > 0.3$), with lead SNPs as potentially causal. (2) Determine the functional annotations using ANNOVAR [166], which can be incorporated into the analysis pipeline. Since the majority of the loci observed in GWAS are located in noncoding regions outside of protein-coding regions [167–169], noncoding annotation is crucial to make sense of GWAS findings. The noncoding annotations can be found using ENCODE [170], the NIH Roadmap Epigenomics Consortium [171], and FANTOM5 [172] projects. (3) Predict coding and noncoding deleteriousness of the variants (strategy is given later). (4) Integrate the functional data probabilistically to prioritize causal variants in statistical fine mapping using a Bayesian framework for both coding and noncoding regions, for example, using a hierarchical model for jointly analyzing GWAS and genomic annotations using a Bayes factor [168]. There are several

other approaches using Bayesian frameworks that have been proposed, namely fGWAS [173–176], PAINTOR [177], PICS [178], CAVIAR [179,180], and GoShifter [181]. There is no gold standard method for fine mapping using functional variant information to determine the causal variants. For example, one could use PAINTOR for fine mapping, which is suitable for transethnic fine-mapping studies, and then validate the findings from PAINTOR with CAVIAR software as well as methodology proposed by van de Bunt using an approximate Bayesian approach [182].

Coding and noncoding regions functional scores derivation: Accurate deleteriousness prediction for nonsynonymous variants is important for distinguishing causal mutations from background polymorphisms. Although many deleteriousness prediction methods are developed, their prediction results are sometimes inconsistent with one another and their merits become unclear in practical downstream variant association-based analysis. For example, rs334, a nonsynonymous exonic variant, is well known as a surrogate variant for sickle cell anemia, but the CADD score is 7.277, which is much lower than the recommended score (>10) for conserved or damaging variants. To address this issue, a comprehensive approach of integrating annotation information from different categories of prediction approaches, such as (1) function prediction scores like SIFT [150], PolyPhen2 [151], and MutationTaster (<http://www.mutationtaster.org/>); (2) conservation scores like GERP++ [183]; and (3) Ensembl scores like CADD [184] and CONDEL (<http://bg.upf.edu/fannsd-b/help/condel.html>), should be used for the annotation of protein consequence, to identify potential nonsynonymous mutations with deleterious consequence on the protein. In addition, functional scores of noncoding variants (GWAVA [185]) and the recently developed programs EIGEN [186], for prediction of functional impact for coding and noncoding variants, and PINES [187], which is designed specifically for prediction of functional impact for noncoding variants, should be used, since most of the GWAS significant SNPs may be in noncoding regions.

11.9 TRANSETHNIC META-ANALYSIS

Transethnic meta-analysis can be performed using MANTRA [188] or PAINTOR [189]. MANTRA is a meta-analysis method developed for transethnic fine

mapping, which maximizes homogeneity between closely related populations while allowing for heterogeneity between more diverse groups. Recently, transethnic meta-analysis methods have been used to discover genetic variants associated with a number of complex traits [190–195].

11.10 OTHER DATA TYPES AND THEIR ANALYSIS METHODS

11.10.1 Gene Expression and RNA-Seq Data

RNA-Seq analysis has become a standard method for global gene expression analysis. Although microarray gene expression statistical methods are well established, there does not exist a gold standard pipeline for analyzing RNA-Seq data. Unlike the microarrays, RNA-Seq does not operate on predetermined selection of cDNA probes, thereby offering a perfect proxy not only for expressed transcript abundance for known genes, but also for the detection and quantification of (1) splice isoforms, (2) novel transcripts, and (3) protein–RNA binding sites. Once RNA abundance is quantified, comparison of gene expression differences and changes between samples representing different treatment/biological conditions has become a well-established application of RNA-Seq. The major steps involved in RNA-Seq data analysis are (1) experimental design consideration, (2) QC, (3) read alignment, (4) quantification of gene and transcript levels, (5) visualization, (6) differential gene expression, (7) alternative splicing, (8) functional analysis, and (9) gene fusion detection. The experimental design involves randomization and selection of the library type, sequencing depth, and number of replicates. The number of replicates and sequencing depth are crucial to detect significant differences in a transcript or gene between experimental groups. The statistical power to detect differential expression varies with desired effect size, sequencing depth, and number of replicates. A number of power calculation tools are available according to experimental design; for example, the Bioconductor package *RNASeqPower* could be used to calculate power assuming effect size, sequencing depth, and number of replicates ([196], <https://bioconductor.org/packages/release/bioc/html/RNASeqPower.html>). Also, *RNAseqPS* ([197] with web interface at <https://cqs.mc.vanderbilt.edu/shiny/RnaSeqSampleSize/>) is a

convenient tool to calculate power and sample size, and Scotty (<http://scotty.genetics.utah.edu/>) calculates the power based on the number of biological replicates, sequencing depth, and cost per replicate and per million reads sequenced. Illumina software generates *FASTQ* files (sequence and quality score). The QC pipeline is crucial in obtaining the highest quality data for all subsequent analyses. QC metrics include sequence quality, GC content, presence of adapters, overrepresentation of *k*-mers and duplicated reads, amount of ribosomal RNA remaining after poly(A) selection, quantification of 3' end bias, detection of viral RNAs through alignment of sequencing reads to a viral genome database, and percentage of reads that align to the genome and transcriptome. Percentage of mapped reads is a global indicator of overall sequencing accuracy and the presence of contamination. Another important metric is the uniformity of read coverage on exons and the mapped strand. *FASTQC* (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) is a commonly used tool to perform QC for Illumina RNA-Seq data. The next step involves alignment of reads to a reference genome or mapping to an annotated transcriptome. *TopHat2* [198]/*STAR2* [199]/*Bowtie2* [200] are popular for mapping to gene and transcriptome. The transcript-specific annotation can be done using *GenomicFeatures* ([201], <http://www.bioconductor.org/packages/2.12/bioc/html/GenomicFeatures.html>). The next step is to estimate gene and transcriptome expression quantification. Programs such as *HTSeq* or *featureCounts* can provide a table of aggregate raw counts of mapped reads [202]. However, the raw read counts are affected by factors such as transcript length, total number of reads, and sequencing biases. Since longer transcripts and deeper sequencing give more reads, the initial solution was proposed to calculate the reads per kilobase per million mapped reads (RPKM) [203]. Subsequent measures were proposed, such as FPKM (fragments per kilobase of exon per million fragments mapped) or TPM (transcripts per kilobase million) [204]. Another issue is that *HTSeq* discards reads mapping to multiple locations, so an alternate method, *RSEM* (RNA-Seq by expectation maximization), that assigns the reads to different locations can be used to quantify the expression from the transcriptome [205]. In addition, the *RSEM* algorithm returns the TPM values for downstream differential gene expression analysis. RPKM, FPKM, and TPM normalize away the most important

factor for comparing samples for differential expression analysis. There are a number of normalization methods that have been proposed for adjusting length and total reads, such as *TMM* [206], *DESeq* [207], *UpperQuartile* [208], etc. A good review of comparisons of these approaches is given in Dillies et al. [209]. *NOISeq* is a useful R package that contains a variety of diagnostic plots to identify sources of biases in RNA-Seq data and then applies appropriate normalization procedures in each case [210]. Since RNA-Seq data are read count data, they must be modeled with a Poisson, negative binomial, or zero-inflated negative binomial distribution. One proposed method is a Bayesian hierarchical model that uses a Poisson distribution, conditional on baseline expression and posttreatment expression level fold change [211]. Another method uses a negative binomial model that assumes common dispersion across all genes and executes an exact test for differential expression [212,213]. The negative binomial model was later extended to generalized linear models, making the method applicable to general experiments [214]. Last, the Bioconductor R package *edgeR* [215,216] and *DESeq2* [217] provide several tools for differential gene expression analysis. *TopHat2/Bowtie2* and *STAR2* aligners also offer built-in options to quantify alternate splicing patterns and gene fusion detections. To identify molecular pathways that are differentially expressed between the biological conditions, the top differentially expressed genes are further investigated for their coexpression pattern in several biological, process, and metabolic pathways. As part of downstream functional analysis, tools like *Ingenuity Pathway Analysis* framework [218], *WebGestalt* [219], and *DAVID* [220–223] can be used to conduct pathway-level analysis.

11.10.2 Epigenome and Methylation Data

For multifactorial diseases/traits, there often exists the problem of missing heritability [139]. The study of epigenetics may provide vital information in a better understanding of this phenotypic variability among individuals, since epigenetic modifications made to the genome regulate the induction and silencing of gene expression. Methylation of CpG's in CpG islands of promoters, which are typically unmethylated for most genes, is frequently associated with gene silencing. DNA methylation, a type of epigenetic modification, occurs when a methyl group is added to the cytosine base,

primarily within the context of CpG dinucleotides. There are two major categories for methylation data creation, namely, microarray-based methods that require hybridization and next-generation sequencing methods. Array hybridization utilizes fluorescence signals to detect methylation levels that can be noisy. The Illumina Infinium methylation arrays have been the most common way to investigate the role of methylation across the human genome for diseases or traits as of this writing. Bisulfite sequencing will become more common as the cost goes down. Here, we will provide a short introduction for both types of data and available software packages starting with array data and then bisulfite sequencing data.

11.10.3 Methylation Data From Arrays

Methylation data from arrays in the human genome are measured by *B* values, which are the ratios of methylated to overall probe intensities, and *M* values, which are the \log_2 ratios of methylated probe to unmethylated probe intensities. The analysis of methylation data by arrays includes probe filtering (detection *P* value, bead count, and SNPs), background correction, adjustment for type II chemistry bias, normalization, cell composition correction, batch effect analysis, batch effect correction, poor performing probes filtering, and differential methylation analysis. There are several R packages available for analyzing methylation data, including *wateRmelon* [224], *methylumi* ([225], <https://www.bioconductor.org/packages/release/bioc/html/methylumi.html>), *minify* [226,227], *ChAMP* [228], and *RnBeads* [229], to name a few. All of these packages allow the user to import raw IDAT files or tabular methylation values; however, *methylumi* and *wateRmelon* do not provide a complete pipeline from raw data to identification of differentially methylated regions (DMRs). Some packages, such as *minifi* and *RnBeads*, provide a complete pipeline for QC to statistical testing for DMRs. The latest version of the *ChAMP* package offers additional features like detection of differentially methylated blocks, quantification of copy number events from 450k/EPIC arrays, and *gene enrichment set analysis* modules.

Sequencing approaches for DNA methylation fall into two main categories, capture-based or enrichment-based sequencing and bisulfite conversion-based sequencing (BS-Seq). Capture-based sequencing uses methyl-binding proteins or antibodies to capture methylated DNA followed by sequencing. Capture-based

sequencing falls into three categories: methods based on methylated DNA immunoprecipitation (such as *MeDIP-Seq* [230]), restriction enzymes sensitive to DNA methylation (such as *MRE-Seq* [231]), and methyl-binding domains to enrich for methylated DNA (such as *methylCap-Seq* [231], *MBD-Seq* [232], and *MIRA-Seq* [233–235]). The drawback of capture-based sequencing is that it has low resolution for methylation detection over 100- to 1000-bp regions. In BS-Seq, bisulfite treatment converts unmethylated cytosines (C's) to thymine (T's) (via uracil), while methylated C's remain unchanged. In BS-Seq, there is no enrichment for the methylated DNA since the whole genome is treated with bisulfite, captured, and then sequenced, providing single-base-pair resolution and methylation status of each CpG site. Therefore, BS-Seq follows fewer steps: treat fragmented DNA with bisulfite (unmethylated C will be converted to U and amplified as T, and methylated C will be protected and remain C, and no change to other bases), amplify the treated DNA, sequence the DNA sequence, align sequence reads to the genome, and then analyze the data. Another affordable alternative to genome-wide methylation sequencing is reduced representation bisulfite sequencing (RRBS) [236,237]. While RRBS is enriched for CpG-rich regions of the genome and hence its high-resolution coverage offerings, its counterpart, whole-genome bisulfite sequencing (WGBS), is a better option for comprehensive cytosine modification profiling. Targeted bisulfite sequencing uses the best of bisulfite sequencing with high-throughput sequencing, but works on a predetermined set of genomic regions of interest that revolve around regions of differentially methylated gene-regulatory elements. Understanding the intricate interactions between gene transcription factor binding sites (TBFS) and DNA methylation can also be well studied using both targeted and WGBS approaches. Packages like *MeDReaders* [238] offer a platform for exploring already cataloged interactions between TBFS and DNA methylation as well as overlapping methylation findings from one's own research. There is no gold standard for analysis of the BS-Seq data. There are several tools available for data preprocessing (for experimental design, read QC, bisulfite conversion rate estimation, base calling, and adapter rimming), data processing (read alignment, methylation scoring, QC, quantitative score assessment), data analysis (DMR detection, transcript binding prediction, interpretation, etc.), and data visualization

(<https://omictools.com/bs-seq-category>). The two most common software systems used for alignment of BS-Seq data are *BISMAR*K [239] and *BSMAP* [239]. The goal of the experiments is to identify methylated regions or loci, so the approaches that borrow information across sites are preferred. There are two common approaches used, namely, the smoothing approach and the Bayesian hierarchical models. A useful software package, *bsseq*, includes smoothing, smoothed *t*-test, Fisher's exact test, DMR identification, and a tool for visualization of results [240]. In contrast, the Bayesian hierarchical technique models biological and technical variation separately using a Betabinomial hierarchical model [241]. Robinson et al. have provided a very nice review of methylation methods for both array-based and sequencing-based methods [242]. Betabinomial distributions have been used to model methylated reads in several software packages, such as *DSS* [241], *BiSeq* [243,244], *MOABS* [245], *RADMeth* [246], and *MethylSig* [247], to identify differentially methylated loci or DMRs. A broader outlook on the involved methylation technologies and their analysis tools is given in Kurdyukov and Bullock [248]. Understanding the coverage of methylation loci, regions, and blocks on a genomic landscape to infer their distribution across gene, promoter, regulatory, intronic, and exonic regions involves the integration of genomic annotation datasets. Software packages like *Genomation* [249] and *ChIPpeakAnno* [250] offer annotation datasets that both RRBS and WGBS analysis modules can benefit from in their cytosine modification findings.

11.10.4 Proteome and Protein Data

Proteomics includes “the systematic identification and quantification of the complete complement of proteins (the proteome) of a biological system (cell, tissue, organ, biological fluid, or organism) at a specific point in time.”⁶ Processing and analysis of proteome data is a complex, multistep process that involves liquid chromatography (LC) coupled to mass spectrometry (MS). The two most common approaches are shotgun MS, which achieves a deep coverage of the proteome, or targeted MS, which uses a defined set of peptides [251]. The Seattle Proteome Center provides the Trans-Proteomic Pipeline, a group of tools for MS-based proteomics, including statistical validation, quantitation, visualization, and

⁶<https://www.nature.com/subjects/proteomic-analysis>.

data conversion. First, MS spectra are assigned to a database search (i.e., X!Tandem, SpectraST, SEQUEST, or Mascot). *PeptideProphet* then validates the peptide assignments and computes a probability that each is correct [252]. Further peptide-level validation may be performed with *iProphet* [253]. In the case of quantitation experiments, *ASAPRatio* measures the relative expression levels of peptides and proteins from isotopically labeled samples [254]. *ProteinProphet* then groups peptides by their corresponding protein(s) to calculate probabilities that those proteins were present in the original sample [255]. *Pep3D* is a useful tool for visualizing LC–MS data and results from *PeptideProphet* [256]. Finally, *reSpect* may identify more peptides from existing spectra without collecting further data [257]. Information on the use and download of these tools can be found at [258] http://tools.proteomecenter.org/wiki/index.php?title=TPP_Tutorial. A bioinformatic analysis of the proteomic data pipeline can be accessed in Schmidt et al. [254].

Functional analysis of a large protein list begins by connecting the protein name to a unique identifier (i.e., Entrez Gene, Unigene, UniProt, etc.) and its associated Gene Ontology (GO) term (part of a biological process, molecular function, or cellular component) ([259], <http://www.geneontology.org>). *GO-term enrichment analysis* (DAVID [220–223]), *Babelomics* 5 ([260], <http://babelomics.bioinfo.cipf.es/>), and *GSEA* ([261,262], <http://software.broadinstitute.org/gsea/index.jsp>) then compare the proportion of specific GO terms in the sample with the natural abundance in the reference dataset [263]. This is followed by pathway analysis, protein–protein interaction analysis, and protein domain and motif analysis [251]. A comprehensive list of biological pathway and molecular interaction resources can be found at *Pathguide* ([264], <http://www.pathguide.org/>). For proteins of unknown function, a *BLAST* search against the database of known protein sequences can determine if proteins with similar amino acid sequences have been described in other organisms ([265], <https://blast.ncbi.nlm.nih.gov/Blast.cgi>). The Proteome Exchange project (<http://www.proteomeexchange.org>) operates the databases PRIDE, Proteome Commons, and Peptide Atlas, which provide stored proteomic datasets [266–268]. An extensive list of gene and protein analysis programs and databases can also be found at <http://www.humgen.nl/programs.html>.

11.10.5 Metabolome and Metabolite Data

Metabolomics is the study of the metabolite composition of a cell type, tissue, or biological fluid. Metabolomics is the fastest growing area of research. Metabolites represent intermediate phenotypes that lead to clinical phenotypes. Metabolomic analyses consist of several steps, namely designing the experiments, sample preparation (grinding, freeze-drying, dilution, etc.), metabolite extraction (targeted or untargeted), derivatization (used only for gas chromatography [GC]), metabolite separation (using LC, GC, or capillary electrophoresis), detection (mass spectroscopy, near-infrared spectroscopy, or nuclear magnetic resonance [NMR]), and data processing and statistical analysis. The most common techniques used in data acquisition are mass spectroscopy and NMR. The aim of the data processing is to extract biologically relevant metabolite features from the data. A good understanding of extraction methods is crucial in minimizing the risk of false positive results in downstream statistical analysis. Thus, processing consists of noise filtering, peak detection, peak alignment, normalization, and deconvolution. A feature is typically a peak or signal that represents a metabolite compound. Once the metabolite features are quantified, then the downstream statistical analysis could follow with univariate or multivariate statistical analyses. Commonly used univariate methods include Student's *t*-test and ANOVA to assess differences between two or more groups (case–control or treatments assigned to different groups). These methods assume that metabolite features are normally distributed and there is no heteroscedasticity (i.e., variability among different groups is unequal), so they may not be appropriate if the statistical assumptions are not met. We can test for these assumptions using the Kolmogorov–Smirnov normality test or Bartlett's homogeneity of variances test. Another option is to use nonparametric tests, such as Mann–Whitney *U* test or Kruskal–Wallis one-way ANOVA, that do not depend on the normality or homogeneity of variance. If the outcome of interest is continuous, linear regression models can be used. Univariate analysis tests each feature separately, leading to the multiple testing problem. Several well-known approaches have been used to adjust for multiple testing, namely Bonferroni correction or false discovery rate (FDR) [269–272]. The Bonferroni correction assumes the independence of each hypothesis test, which is not true in general for untargeted metabolomics studies.

Also, Bonferroni correction is the most conservative. Thus, FDR has been favored over Bonferroni correction for metabolomics studies. In contrast to univariate methods, multivariate methods do not assume independence of metabolites and model a group of metabolites clustered together, resulting in reduction of multidimensionality. Multivariate methods are classified into supervised and unsupervised methods. Supervised methods, which use both metabolite and trait information, include support vector machines, neural networks, linear discriminant analysis, regression analysis, logistic regression analysis, regression trees, naïve Bayes, inductive logic programming, etc. Partial least squares (PLS) [273], a supervised method, is most commonly used in identifying the metabolomics patterns associated with the variable of interest. The advantage of PLS is that it can be used as a regression analysis for continuous outcomes or discriminant analysis [274], for binary outcomes. The drawback of the PLS is that some metabolites that are not associated with the outcome variable can influence the results. To alleviate the problem, orthogonal PLS can be used; since data variance is factorized into two orthogonal components, one is correlated with the outcome of interest and the other is uncorrelated [275]. The most commonly used unsupervised method in metabolomics is PCA. PCA creates new variables (PCs) by linear combinations of metabolites while maximizing the variance and minimizing the covariance between components as well as reducing the data dimension. This new variable can be correlated with the outcome of interest. Other unsupervised methods, such as *hierarchical clustering analysis*, have been used to analyze metabolites. Software to perform the processing of the metabolite data sets is provided by individual instrument manufactures. There are some R packages that can be used to analyze metabolomics data, such as *XCMS* (available online as well as command line level using R), a very powerful software for the analysis of metabolomics data sets written in R ([276], <https://xcmsonline.scripps.edu>), and *MetaboAnalyst* ([277], www.metaboanalyst.ca).

11.11 FUTURE DIRECTIONS/INTEGRATION

Integrating information from the genome, epigenome, transcriptome, proteome, metabolome, phenome, etc., to understand the biological mechanisms of multifactorial traits is challenging. Integration of different omic

data sets faces issues such as computational burden, high dimensionality of data, connectivity, data heterogeneity and comparability, visualization, multiple testing, and downstream analysis with different pathways and networks. For example, in expression quantitative trait (eQTL) analysis, associations between genotypes from GWAS and transcriptome profiles are sought to unravel the genetic basis of multifactorial diseases. Similarly, association of genotypes can be tested with methylation (meQTL), protein (pQTL), and metabolites (mQTL) (see Fig. 11.6).

MatrixEQTL is a computationally efficient software and can be used for eQTL, meQTL, pQTL, and mQTL studies [278]. Several tools have been developed to help integrate multiple omic data sets based on pathway enrichment analysis from gene or protein expression and metabolomics data (*IMPALA* [279]); transcriptome, proteome, and metabolome data sets (*iPEAP* [280]); and transcriptome and metabolomic data (*MetaboAnalyst* [277]). Pathway-based approaches rely on known pathways, which may not represent the complexity of biological systems. Another set of tools has been developed based on biological network analysis; for example, *SAMNetWeb* [281] and *pwOmics* [282] model transcriptomics and proteomics data sets, *MetScape* [283] models gene expression and metabolite data, and *Grinn* [284] uses genomic, proteomic, and metabolomic data sets.

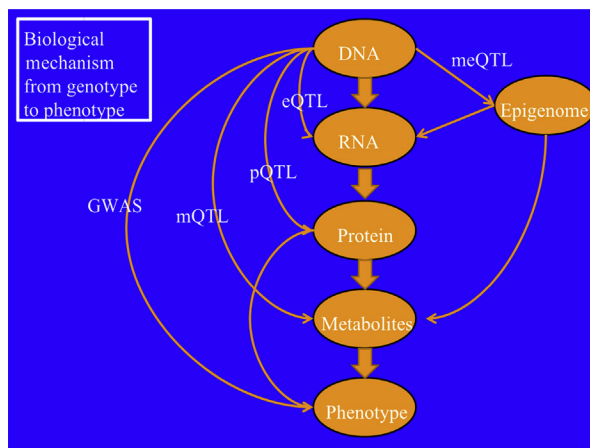


Figure 11.6 Relationships between various types of genetic data and the molecular biology pathway. *eQTL*, expression quantitative trait loci; *GWAS*, genome-wide association studies; *meQTL*, methylation QTL; *mQTL*, metabolites QTL; *pQTL*, protein QTL.

Another notable software to analyze multivariate mQTL data in a GWAS framework, *xMWAS*, integrates metabolomics, transcriptomic data, and network analysis using multilevel sparse PLS regression [285]. Integration of omics is a rapidly evolving field of research and a holy grail of systems biology.

11.12 CONCLUSIONS

Genetic modeling is a challenging art and science. Advances in molecular technology and statistical methodology⁷ and increasing availability of large samples allow many new investigations to be undertaken on unprecedented scales. Interpretation of the resulting findings remains both difficult and one of the more exciting challenges facing today's biomedical researchers.

REFERENCES

- [1] Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet* October 2002;3(10):779–89.
- [2] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 1918;52:399–433.
- [3] Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet* August 1965;29:51–76.
- [4] Reich T, James JW, Morris CA. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. *Ann Hum Genet* November 1972;36(2):163–84.
- [5] Chakraborty R. The inheritance of pyloric stenosis explained by a multifactorial threshold model with sex dimorphism for liability. *Genet Epidemiol* 1986;3(1): 1–15.
- [6] Dronamraju KR, Bixler D, Majumder PP. Fetal mortality associated with cleft lip and cleft palate. *Johns Hopkins Med J* December 1982;151(6):287–9.
- [7] Dronamraju KR, Bixler D. Fetal mortality in oral cleft families (IV): the “doubling effect”. *Clin Genet* July 1983;24(1):22–5.
- [8] Elston RC, Boklage CE. An examination of fundamental assumptions of the twin method. *Prog Clin Biol Res* 1978;24A:189–99.
- [9] Hopper JL. Twin concordance. In: *Encyclopedia of biostatistics*, vol. 6. New York: John Wiley; 1998. p. 4626–9.
- [10] Karlin S, Cameron EC, Williams PT. Sibling and parent–offspring correlation estimation with variable family size. *Proc Natl Acad Sci U S A* May 1981;78(5):2664–8.
- [11] Neale MC. Adoption studies. In: *Encyclopedia of biostatistics*, vol. 1. New York: John Wiley; 1998. p. 77–81.
- [12] Neale MC, Cardon LR. *Methodology for genetic studies of twins and families*. London: Kluwer; 1992.
- [13] Carey G. Sibling imitation and contrast effects. *Behav Genet* May 1986;16(3):319–41.
- [14] Lykken DT, McGue M, Tellegen A, Bouchard Jr TJ. Emergenesis. Genetic traits that may not run in families. *Am Psychol* December 1992;47(12):1565–77.
- [15] Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* February 1990;46(2):222–8.
- [16] Olson JM, Cordell HJ. Ascertainment bias in the estimation of sibling genetic risk parameters. *Genet Epidemiol* March 2000;18(3):217–35.
- [17] Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* March 2006;2(3):e41. [Epub 2006 Mar 24].
- [18] Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* September 1955;7(3):277–318.
- [19] Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Hum Hered* 1971;21(6):523–42.
- [20] Elston RC, Rao DC. Statistical modeling and analysis in human genetics. *Annu Rev Biophys Bioeng* 1978;7:253–86. [Review].
- [21] Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* April 1987;84(8):2363–7.
- [22] Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* March 1972;2(1):3–19.
- [23] Penrose LS. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann Eugen (London)* 1935;6:133–8.
- [24] Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE. A more powerful robust sib-pair test of linkage for quantitative traits. *Genet Epidemiol* 1989;6(3):435–49.
- [25] Olson JM, Wijsman EM. Linkage between quantitative trait and marker loci: methods using all relative pairs. *Genet Epidemiol* 1993;10(2):87–102.

⁷ A steady stream of videos offering tutelage on these advances can be freely seen at <http://www.soph.uab.edu/ssg/courses/>.

- [26] Drigalenko E. How sib pairs reveal linkage. *Am J Hum Genet* October 1998;63(4):1242–5.
- [27] Forrest WF. Weighting improves the “new Haseman-Elston” method. *Hum Hered* 2001;52(1):47–54.
- [28] Gerhard D, Hothorn LA. Rank transformation in Haseman-Elston regression using scores for location-scale alternatives. *Hum Hered* 2010;69(3):143–51.
- [29] Sham PC, Purcell S. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet* June 2001;68(6):1527–32.
- [30] Sham PC, Purcell S, Cherny SS, Abecasis GR. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet* August 2002;71(2):238–53.
- [31] Shete S, Jacobs KB, Elston RC. Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. *Hum Hered* 2003;55(2–3):79–85.
- [32] Visscher PM, Hopper JL. Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Ann Hum Genet* November 2001;65(Pt 6):583–601.
- [33] Wang T, Elston RC. A modified revisited Haseman-Elston method to further improve power. *Hum Hered* 2004;57(2):109–16.
- [34] Wright FA. The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet* March 1997;60(3):740–2.
- [35] Xu X, Weiss S, Xu X, Wei LJ. A unified Haseman-Elston method for testing linkage with quantitative traits. *Am J Hum Genet* October 2000;67(4):1025–8.
- [36] S.A.G.E. 6.x. Statistical analysis for genetic epidemiology. 2010. <http://darwin.cwru.edu/sage/>.
- [37] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* March 1996;58:1347–63.
- [38] Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* January 2002;30(1):97–101.
- [39] Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* May 1998;62(5):1198–211.
- [40] Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* March 1994;54(3):535–43.
- [41] Amos CI, Zhu DK, Boerwinkle E. Assessing genetic linkage and association with robust components of variance approaches. *Ann Hum Genet* March 1996;60(Pt 2):143–60.
- [42] Goldgar DE. Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* December 1990;47(6):957–67.
- [43] Schork NJ. Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am J Hum Genet* December 1993;53(6):1306–19.
- [44] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* March 1993;52(3):506–16.
- [45] Tiwari HK, Barnholtz-Sloan J, Wineinger N, Padilla MA, Vaughan LK, Allison DB. Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered* 2008;66(2):67–86.
- [46] Page GP, George V, Go RC, Page PZ, Allison DB. “Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet* October 2003;73(4):711–9. [Review].
- [47] Lander ES. The new genomics: global views of biology. *Science* October 25, 1996;274(5287):536–9.
- [48] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissensbach J, Heilig R, Saurin W, Artiguenave F, Brottier

- P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. International human genome sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* February 15, 2001;409(6822):860–921.
- [49] International HapMap Consortium. The International HapMap project. *Nature* December 18, 2003;426(6968):789–96.
- [50] International HapMap Consortium. A haplotype map of the human genome. *Nature* October 27, 2005;437(7063):1299–320.
- [51] International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Wayne MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Peér I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, NkwoDIMMAH C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. A second generation human haplotype map of over 3.1 million SNPs. *Nature* October 18, 2007;449(7164):851–61.
- [52] International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF,

- Zhang Q, Ghorri MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. *Nature* September 2, 2010;467(7311):52–8.
- [53] Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* October 28, 2010;467(7319):1061–73. [Erratum in: *Nature* May 26, 2011;473(7348):544].
- [54] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* November 2012;491:56–65.
- [55] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* October 2015;526:68–74.
- [56] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Korbel JO. An integrated map of structural variation in 2,504 human genomes. *Nature* October 2015;526(7571):75–81.
- [57] Birney E, Soranzo N. Human genomics: the end of the start for population sequencing. *Nature* October 2015;526:52–3.
- [58] Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* July 20, 2001;293(5529):489–93.
- [59] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* May 2008;9(5):356–69. [Review].
- [60] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* June 7, 2007;447(7145):661–78.
- [61] Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* September 13, 1996;273(5281):1516–7.
- [62] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* September 2007;81(3):559–75.
- [63] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* September 2010;5(9):1564–73. <https://doi.org/10.1038/nprot.2010.116>. [Epub 2010 Aug 26].
- [64] Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS. GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* September 2010;34(6):591–602.
- [65] Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 2011 Jan; Chapter 1:Unit1.19;68:1.19.1–1.19.18. <https://doi.org/10.1002/0471142905.hg0119s68>.
- [66] Devlin B, Roeder K. Genomic control for association studies. *Biometrics* December 1999;55(4):997–1004.
- [67] Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet* September 2005;1(3):e32.
- [68] Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* November 2001;60(3):155–66. [Review].
- [69] Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* June 2000;66(6):1933–44.
- [70] Dadd T, Weale ME, Lewis CM. A critical evaluation of genomic control methods for genetic association studies. *Genet Epidemiol* May 2009;33(4):290–8. [Review].
- [71] Devlin B, Bacanu SA, Roeder K. Genomic control to the extreme. *Nat Genet* November 2004;36(11):1129–30.
- [72] Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* January 2001;20(1):4–16. [Review].
- [73] Zheng G, Freidlin B, Li Z, Gastwirth JL. Genomic control for association studies under various genetic models. *Biometrics* March 2005;61(1):186–92.
- [74] Zheng G, Freidlin B, Gastwirth JL. Robust genomic control for association studies. *Am J Hum Genet* February 2006;78(2):350–6.

- [75] Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* July 1999;65(1):220–8.
- [76] Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* July 2000;67(1):170–81.
- [77] Redden D, Divers J, Vaughan L, Tiwari H, Beasley T, Fernandez J, Kimberly R, Feng R, Padilla M, Lui N, Miller M, Allison D. Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. *PLoS Genet* August 25, 2006;2(8):e137.
- [78] Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* December 2006;2(12):e190.
- [79] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* August 2006;38(8):904–9.
- [80] Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genom Hum Genet* 2009;10:387–406.
- [81] Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* July 2010;11:499–511.
- [82] Zheng H-F, Rong J-J, Liu M, Han F, Zhang XW, Richards JB, Wang L. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One* January 2015;10(1):e0116487.
- [83] Wood AR, Perry JRB, Tanaka T, Hernandez DG, Zheng HF, Melzer D, Frayling TM. Imputation of variants from the 1000 genomes project modestly improves known associations and can identify low-frequency variant - phenotype associations undetected by HapMap based imputation. *PLoS One* May 2013;8(5):e64343.
- [84] Buchanan CC, Torstenson ES, Bush WS, Ritchie MD. A comparison of cataloged variation between international HapMap Consortium and 1000 genomes project data. *J Am Med Inf Assoc* 2012 Mar-Apr;19(2):28994.
- [85] Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906–13.
- [86] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;5:e1000529.
- [87] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* December 2010;34(8):816–34.
- [88] Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* July 2012;44(8):955–9.
- [89] Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics* March 2015;31(5):782–4.
- [90] Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007;81:1084–97.
- [91] Browning SR, Browning BL. Genotype imputation with millions of reference samples. *Am J Hum Genet* 2016;98:116–26.
- [92] Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genom Inform* 2012;10(2):117–22.
- [93] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet* January 2012;90:7–24.
- [94] Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* April 2014;15:335–46.
- [95] Scherag A, Müller HH, Dempfle A, Hebebrand J, Schäfer H. Data adaptive interim modification of sample sizes for candidate-gene association studies. *Hum Hered* 2003;56:56–62.
- [96] Gordon D, Levenstien MA, Finch SJ, Ott J. Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. *Pac Symp Biocomput* 2003:490–501.
- [97] Pfeiffer RM, Gail MH. Sample size calculations for population and family-based case-control association studies on marker genotypes. *Genet Epidemiol* 2003;25:136–48.
- [98] Pirinen M, Donnelly P, Spencer CCA. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet* July 2012;44:848–51.
- [99] Spencer C, Su Z, Donnelly P, Marchini J. Designing Genome-Wide Association Studies: sample size, power, and the choice of genotyping chip. *PLoS Genet* May 2009;5(5):e1000477.
- [100] Tiwari HK, Birkner T, Moondan A, Zhang S, Page GP, Patki A. Accurate and flexible power calculations on the spot: applications to genomic research. *Stat Interface* 2011;4(3):353–8.
- [101] Purcell S, Cherny SS, Sham PC. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003;19(1):149–50.

- [102] Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–13.
- [103] Menashe I, Rosenberg PS, Chen BE. PGA: power calculator for case-control genetic association analyses. *BMC Genet* May 2008;9:36.
- [104] Zhao JH. Gap: genetic analysis package. R package version 1. 2017. p. 1–17. <https://CRAN.R-project.org/package=gap>.
- [105] Weeks DE. powerpkg: power analyses for the affected sib pair and the TDT design. R package version 1. 2012. p. 5. <https://CRAN.R-project.org/package=pow-erpkg>.
- [106] Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* September 2012;13:762–75.
- [107] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* January 2011;88(1):76–82.
- [108] Visscher PM, Hemani G, Vinkhuyzen AAE, Chen GB, Lee SH, Wray NR, Goddard ME, Yang J. Statistical power to detect genetic (Co)Variance of complex traits using SNP data in unrelated samples. *PLoS Genet* 2014;10(4):e1004269.
- [109] Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med* January 2002;21(1):35–50.
- [110] Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* March 2002;155(5):478–84.
- [111] Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil Mag* 1900;50(302):157–75. [Series 5].
- [112] Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J Roy Stat Soc* 1922;85(1):87–94.
- [113] Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11(3):375–86.
- [114] Cochran WG. Some methods for strengthening the common chi-square tests. *Biometrics* 1954;10(4):417–51.
- [115] Sasieni P. From genotypes to genes: doubling the sample size. *Biometrics* December 1997;53(4):1253–61.
- [116] Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 2002;53(3):146–52.
- [117] Huang BE, Lin DY. Efficient association mapping of quantitative trait loci with selective genotyping. *Am J Hum Genet* 2007;80:567–76.
- [118] Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* September 2007;177(1):577–85.
- [119] Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. *Genetics* March 2008;178(3):1709–23.
- [120] Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* April 2010;42(4):355–60.
- [121] Aulchenko YS, Struchalin MV, van Duijn CM. ProBABEL package for genome-wide association analysis of imputed data. *BMC Bioinf* 2010;11:1345.
- [122] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* June 17, 2012;44(7):821–4. <https://doi.org/10.1038/ng.2310>.
- [123] Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* April 2014;11(4):407–9. <https://doi.org/10.1038/nmeth.2848>. [Epub 2014 Feb 16].
- [124] Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* May 2006;7(5):385–94. [Review].
- [125] Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Software* 2015;67(1):1–48. <https://doi.org/10.18637/jss.v067.i01>.
- [126] Chen MH, Yang Q. GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics* February 15, 2010;26(4):580–1. <https://doi.org/10.1093/bioinformatics/btp710>. [Epub 2009 Dec 29].
- [127] Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol* February 2008;32(2):179–85.
- [128] Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* October 2006;7(10):781–91. [Review].
- [129] Corvin A, Craddock N, Sullivan PF. Genome-wide association studies: a primer. *Psychol Med* July 2010;40(7):1063–77.
- [130] Hinrichs AL, Larkin EK, Suarez BK. Population stratification and patterns of linkage disequilibrium. *Genet Epidemiol* 2009;33(Suppl. 1):S88–92.
- [131] Bhatnagar S. Interactive Q-Q and Manhattan plots using Plotly.js. 2016 Nov. <http://sahirbhatnagar.com/manhattanly/>.

- [132] Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. 2014. <https://doi.org/10.1101/005165>. *bioRxiv*.
- [133] Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* September 2010;26(18):2336–7.
- [134] Dadev T, Leongamornlert DA, Saunders EJ, Eeles R, Kote-Jarai Z. LocusExplorer: a user-friendly tool for integrated visualization of human genetic association data and biological annotations. *Bioinformatics* 2016;32(6):949–51.
- [135] Schillert A, Schwarz DF, Vens M, Szymczak S, König IR, Ziegler A. ACPA: automated cluster plot analysis of genotype data. *BMC Proc* 2009;3(Suppl. 7):S58.
- [136] Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, Zheng X, Crosslin DR, Levine D, Lumley T, Nelson SC, Rice K, Shen J, Swarnkar R, Weir BS, Laurie CC. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* December 2012;28(24):3329–31.
- [137] Chakravarti A. Population genetics—making sense out of sequence. *Nat Genet* January 1999;21(1 Suppl.):56–60. [Review].
- [138] Lohmueller KE, Mauney MM, Reich D, Braverman JM. Variants associated with common disease are not unusually differentiated in frequency across populations. *Am J Hum Genet* January 2006;78(1):130–6.
- [139] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature* October 8, 2009;461(7265):747–53. [Review].
- [140] Terwilliger JD, Hiekkalinna T. An utter refutation of the “fundamental theorem of the HapMap”. *Eur J Hum Genet* April 2006;14(4):426–37.
- [141] Terwilliger JD, Göring HH. Update to Terwilliger and Göring’s “Gene mapping in the 20th and 21st centuries” (2000): gene mapping when rare variants are common and common variants are rare. *Hum Biol* December 2009;81(5–6):729–33.
- [142] Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant... or not? *Hum Mol Genet* October 1, 2002;11(20):2417–23.
- [143] Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* June 2008;40(6):695–701.
- [144] Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* February 2004;5(2):89–100.
- [145] Mardis ER. A decade’s perspective on DNA sequencing technology. *Nature* February 10, 2011;470(7333):198–203.
- [146] Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* January 2010;42(1):30–5.
- [147] Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res* May 2011;21(5):658–64.
- [148] Rödelberger C, Krawitz P, Bauer S, Hecht J, Bigham AW, Bamshad M, de Condor BJ, Schweiger MR, Robinson PN. Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. *Bioinformatics* March 15, 2011;27(6):829–36.
- [149] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* September 10, 2009;461(7261):272–6.
- [150] Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc* January 2016;11(1):1–9. <https://doi.org/10.1038/mprof.2015.123>. [Epub 2015 Dec 3].
- [151] Sunyaev S, Ramensky V, Koch I, Lathe 3rd W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet* March 15, 2001;10(6):591–7.
- [152] Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* February 3, 2007;615(1–2):28–56.
- [153] Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* September 2008;83(3):311–21.
- [154] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* July 2011;89(1):82–93.
- [155] Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* February 2009;5(2):e1000384.
- [156] Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* June 11, 2010;86(6):832–8.
- [157] Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* January 2011;35(1):57–69.

- [158] Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* November 2010;11(11):773–85.
- [159] Ansorge WJ. Next-generation DNA sequencing techniques. *Nat Biotechnol* April 2009;25(4):195–203.
- [160] Hirst M, Marra MA. Next generation sequencing based approaches to epigenomics. *Brief Funct Genom* December 2010;9(5–6):455–65. [Review].
- [161] Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* October 2010;11(10):685–96. [Review].
- [162] Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, Wunderlich A, Barmeyer C, Seemann P, Koenig J, Lappe M, Kuss AW, Garshasbi M, Bertram L, Trappe K, Werber M, Herrmann BG, Zatloukal K, Lehrach H, Schweiger MR. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PLoS One* December 22, 2010;5(12):e15661.
- [163] Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, NISC Comparative Sequencing Program, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, Rosenberg SA, Samuels Y. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* May 2011;43(5):442–6. [Epub 2011 Apr 15].
- [164] Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JM, Auton A, Myers S, Morris A, Pirinen M, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Craddock N, Hurles M, Ouwehand W, Parkes M, Rahman N, Duncanson A, Todd JA, Kwiatkowski DP, Samani NJ, Gough SC, McCarthy MI, Deloukas P, Donnelly P. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* December 2012;44(12):1294–301. <https://doi.org/10.1038/ng.2435>. [Epub 2012 Oct 28].
- [165] Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Mägi R, Reschen ME, Mahajan A, Locke A, Rayner NW, Robertson N, Scott RA, Prokopenko I, Scott LJ, Green T, Sparso T, Thuillier D, Yengo L, Grallert H, Wahl S, Frånberg M, Strawbridge RJ, Kestler H, Chheda H, Eisele L, Gustafsson S, Steinthorsdottir V, Thorleifsson G, Qi L, Karssen LC, van Leeuwen EM, Willems SM, Li M, Chen H, Fuchsberger C, Kwan P, Ma C, Linderman M, Lu Y, Thomsen SK, Rundel JK, Beer NL, van de Bunt M, Chalisey A, Kang HM, Voight BF, Abecasis GR, Almgren P, Baldassarre D, Balkau B, Benediktsson R, Blüher M, Boeing H, Bonnycastle LL, Bottinger EP, Burtt NP, Carey J, Charpentier G, Chines PS, Cornelis MC, Couper DJ, Crenshaw AT, van Dam RM, Doney AS, Dorkhan M, Edkins S, Eriksson JG, Esko T, Eury E, Fadista J, Flannick J, Fontanillas P, Fox C, Franks PW, Gertow K, Gieger C, Gigante B, Gottesman O, Grant GB, Grarup N, Groves CJ, Hassinen M, Have CT, Herder C, Holmen OL, Hreidarsson AB, Humphries SE, Hunter DJ, Jackson AU, Jonsson A, Jørgensen ME, Jørgensen T, Kao WH, Kerrison ND, Kinnunen L, Klopp N, Kong A, Kovacs P, Kraft P, Kravic J, Langford C, Leander K, Liang L, Lichtner P, Lindgren CM, Lindholm E, Linneberg A, Liu CT, Lobbens S, Luan J, Lyssenko V, Männistö S, McLeod O, Meyer J, Mihailov E, Mirza G, Mühleisen TW, Müller-Nurasyid M, Navarro C, Nöthen MM, Oskolkov NN, Owen KR, Palli D, Pechlivanis S, Peltonen L, Perry JR, Platou CG, Roden M, Ruderfer D, Rybin D, van der Schouw YT, Sennblad B, Sigurðsson G, Stančáková A, Steinbach G, Storm P, Strauch K, Stringham HM, Sun Q, Thorand B, Tikkanen E, Tonjes A, Trakalo J, Tremoli E, Tuomi T, Wennauer R, Wiltshire S, Wood AR, Zeggini E, Dunham I, Birney E, Pasquali L, Ferrer J, Loos RJ, Dupuis J, Florez JC, Boerwinkle E, Pankow JS, van Duijn C, Sijbrands E, Meigs JB, Hu FB, Thorsteinsdottir U, Stefansson K, Lakka TA, Rauramaa R, Stumvoll M, Pedersen NL, Lind L, Keinanen-Kiukkaanniemi SM, Korpi-Hyövälti E, Saaristo TE, Saltevo J, Kuusisto J, Laakso M, Metspalu A, Erbel R, Jöcke KH, Moebus S, Ripatti S, Salomaa V, Ingelsson E, Boehm BO, Bergman RN, Collins FS, Mohlke KL, Koistinen H, Tuomilehto J, Hveem K, Njølstad I, Deloukas P, Donnelly PJ, Frayling TM, Hattersley AT, de Faire U, Hamsten A, Illig T, Peters A, Cauchi S, Sladek R, Froguel P, Hansen T, Pedersen O, Morris AD, Palmer CN, Kathiresan S, Melander O, Nilsson PM, Groop LC, Barroso I, Langenberg C, Wareham NJ, O'Callaghan CA, Gloyn AL, Altshuler D, Boehnke M, Teslovich TM, McCarthy MI, Morris AP, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* December 2015;47(12):1415–25. <https://doi.org/10.1038/ng.3437>. [Epub 2015 Nov 9].
- [166] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* September 2010;38(16):e164. <https://doi.org/10.1093/nar/gkq603>. [Epub 2010 Jul 3].
- [167] Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* June 9, 2009;106(23):9362–7. <https://doi.org/10.1073/pnas.0903103106>. [Epub 2009 May 27].

- [168] Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* April 3, 2014;94(4):559–73. <https://doi.org/10.1016/j.ajhg.2014.03.004>. [Erratum in: *Am J Hum Genet* July 3, 2014;95(1):126].
- [169] Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium, Kähler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* November 6, 2014;95(5):535–52. <https://doi.org/10.1016/j.ajhg.2014.10.004>. [Epub 2014 Nov 6].
- [170] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* September 6, 2012;489(7414):57–74. <https://doi.org/10.1038/nature11247>.
- [171] Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, Gaasterland T, Gissi C, King B, Kochiwa H, Kuehl P, Lewis S, Matsuo Y, Nikaido I, Pesole G, Quackenbush J, Schriml LM, Staubli F, Suzuki R, Tomita M, Wagner L, Washio T, Sakai K, Okido T, Furuno M, Aono H, Baldarelli R, Barsh G, Blake J, Boffelli D, Bojunga N, Carninci P, de Bonaldo MF, Brownstein MJ, Bult C, Fletcher C, Fujita M, Gariboldi M, Gustincich S, Hill D, Hofmann M, Hume DA, Kamiya M, Lee NH, Lyons P, Marchionni L, Mashima J, Mazzarelli J, Mombaerts P, Nordone P, Ring B, Ringwald M, Rodriguez I, Sakamoto N, Sasaki H, Sato K, Schönbach C, Seya T, Shibata Y, Storch KF, Suzuki H, Toyo-oka K, Wang KH, Weitz C, Whittaker C, Wilming L, Wynshaw-Boris A, Yoshida K, Hasegawa Y, Kawaji H, Kohtsuki S, Kohtsuki S, Hayashizaki Y, RIKEN Genome Exploration Research Group Phase II Team, the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature* February 8, 2001;409(6821):685–90.
- [172] Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: roadmap for regulation. *Nature* February 19, 2015;518(7539):314–6. <https://doi.org/10.1038/518314a>. [No abstract available].
- [173] Jiangtao L, Arthur B, Kwangi A, Kiranmoy D, Jiahan L, Zhong W, Yao L, Rongling W. Functional genome-wide association studies of longitudinal traits. In: Chow SC, editor. *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. London, UK: Wiley; 2010.
- [174] Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. *Bioinformatics* February 15, 2011;27(4):516–23. <https://doi.org/10.1093/bioinformatics/btq688>. [Epub 2010 Dec 14].
- [175] Das K, Li J, Wang Z, Tong C, Fu G, Li Y, Xu M, Ahn K, Mauger D, Li R, Wu R. A dynamic model for genome-wide association studies. *Hum Genet* June 2011;129(6):629–39. <https://doi.org/10.1007/s00439-011-0960-6>. [Epub 2011 Feb 4].
- [176] Li J, Wang Z, Li R, Wu R. Bayesian group lasso for non-parametric varying-coefficient models with application to functional genome-wide association studies. *Ann Appl Stat* June 2015;9(2):640–64.
- [177] Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* October 30, 2014;10(10):e1004722. <https://doi.org/10.1371/journal.pgen.1004722>. [eCollection 2014 Oct].
- [178] Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shores N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* February 19, 2015;518(7539):337–43. <https://doi.org/10.1038/nature13835>. [Epub 2014 Oct 29].
- [179] Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. *Genetics* October 2014;198(2):497–508. <https://doi.org/10.1534/genetics.114.167908>. [Epub 2014 Aug 7].
- [180] Hormozdiari F, Kichaev G, Yang WY, Pasaniuc B, Eskin E. Identification of causal genes for complex traits. *Bioinformatics* June 15, 2015;31(12):i206–13. <https://doi.org/10.1093/bioinformatics/btv240>.
- [181] Trynka G, Westra HJ, Slowikowski K, Hu X, Xu H, Stranger BE, Klein RJ, Han B, Raychaudhuri S. Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am J Hum Genet* July 2, 2015;97(1):139–52. <https://doi.org/10.1016/j.ajhg.2015.05.016>.
- [182] van de Bunt M, Cortes A, IGAS Consortium, Brown MA, Morris AP, McCarthy MI. Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS Genet* September 25,

- 2015;11(9):e1005535. <https://doi.org/10.1371/journal.pgen.1005535>. [eCollection 2015].
- [183] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* December 2, 2010;6(12):e1001025. <https://doi.org/10.1371/journal.pcbi.1001025>.
- [184] Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* February 2, 2014. <https://doi.org/10.1038/ng.2892>.
- [185] Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nat Methods* March 2014;11(3):294–6. <https://doi.org/10.1038/nmeth.2832>. [Epub 2014 Feb 2].
- [186] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* February 2016;48(2):214–20. <https://doi.org/10.1038/ng.3477>.
- [187] Bodea CA, Mitchell AA, Runz H, Sunyaev SR. Phenotype-specific information improves prediction of functional impact for noncoding variants. *bioRxiv*. <https://doi.org/10.1101/083642>.
- [188] Morris AP. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* December 2011;35(8):809–22. <https://doi.org/10.1002/gepi.20630>.
- [189] Kichaev G, Pasaniuc B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am J Hum Genet* August 6, 2015;97(2):260–71. <https://doi.org/10.1016/j.ajhg.2015.06.007>. [Epub 2015 Jul 16. Erratum in: *Am J Hum Genet* Aug 6, 2015;97(2):353].
- [190] Keller MF, Reiner AP, Okada Y, van Rooij FJ, Johnson AD, Chen MH, Smith AV, Morris AP, Tanaka T, Ferrucci L, Zonderman AB, Lettre G, Harris T, Garcia M, Bandinelli S, Qayyum R, Yanek LR, Becker DM, Becker LC, Kooperberg C, Keating B, Reis J, Tang H, Boerwinkle E, Kamatani Y, Matsuda K, Kamatani N, Nakamura Y, Kubo M, Liu S, Dehghan A, Felix JF, Hofman A, Uitterlinden AG, van Duijn CM, Franco OH, Longo DL, Singleton AB, Psaty BM, Evans MK, Cupples LA, Rotter JI, O’Donnell CJ, Takahashi A, Wilson JG, Ganesh SK, Nalls MA. Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum Mol Genet* December 20, 2014;23(25):6944–60.
- [191] Ng MC, Shriner D, Chen BH, Li J, Chen WM, Guo X, Liu J, Bielinski SJ, Yanek LR, Nalls MA, Comeau ME, Rasmussen-Torvik LJ, Jensen RA, Evans DS, Sun YV, An P, Patel SR, Lu Y, Long J, Armstrong LL, Wagenknecht L, Yang L, Snively BM, Palmer ND, Mudgal P, Langeveld CD, Keene KL, Freedman BI, Mychaleckyj JC, Nayak U, Raffel LJ, Goodarzi MO, Chen YD, Taylor Jr HA, Correa A, Sims M, Couper D, Pankow JS, Boerwinkle E, Adeyemo A, Doumatey A, Chen G, Mathias RA, Vaidya D, Singleton AB, Zonderman AB, Igo Jr RP, Sedor JR, Kabagambe EK, Siscovick DS, McKnight B, Rice K, Liu Y, Hsueh WC, Zhao W, Bielak LF, Kraja A, Province MA, Bottinger EP, Gottesman O, Cai Q, Zheng W, Blot WJ, Lowe WL, Pacheco JA, Crawford DC, Grundberg E, Rich SS, Hayes MG, Shu XO, Loos RJ, Borecki IB, Peyser PA, Cummings SR, Psaty BM, Fornage M, Iyengar SK, Evans MK, Becker DM, Kao WH, Wilson JG, Rotter JI, Sale MM, Liu S, Rotimi CN, Bowden DW. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* August 2014;10(8):e1004517.
- [192] Cornelis MC, Byrne EM, Esko T, Nalls MA, Ganna A, Paynter N, Monda KL, Amin N, Fischer K, Renstrom F, Ngwa JS, Huikari V, Cavadino A, Nolte IM, Teumer A, Yu K, Marques-Vidal P, Rawal R, Manichaikul A, Wojczynski MK, Vink JM, Zhao JH, Burlutsky G, Lahti J, Mikkila V, Lemaitre RN, Eriksson J, Musani SK, Tanaka T, Geller F, Luan J, Hui J, Magi R, Dimitriou M, Garcia ME, Ho WK, Wright MJ, Rose LM, Magnusson PK, Pedersen NL, Couper D, Oostra BA, Hofman A, Ikram MA, Tiemeier HW, Uitterlinden AG, van Rooij FJ, Barroso I, Johansson I, Xue L, Kaakinen M, Milani L, Power C, Snieder H, Stolk RP, Baumeister SE, Biffar R, Gu F, Bastardot F, Kutalik Z, Jacobs Jr DR, Forouhi NG, Mihailov E, Lind L, Lindgren C, Michaelsson K, Morris A, Jensen M, Khaw KT, Luben RN, Wang JJ, Mannisto S, Perala MM, Kahonen M, Lehtimäki T, Viikari J, Mozaffarian D, Mukamal K, Psaty BM, Doring A, Heath AC, Montgomery GW, Dahmen N, Carithers T, Tucker KL, Ferrucci L, Boyd HA, Melbye M, Treur JL, Mellstrom D, Hottenga JJ, Prokopenko I, Tonjes A, Deloukas P, Kanoni S, Lorentzon M, Houston DK, Liu Y, Danesh J, Rasheed A, Mason MA, Zonderman AB, Franke L, Kristal BS, Karjalainen J, Reed DR, Westra HJ, Evans MK, Saleheen D, Harris TB, Dedoussis G, Curhan G, Stumvoll M, Beilby J, Pasquale LR, Feenstra B, Bandinelli S, Ordoñas JM, Chan AT, Peters U, Ohlsson C, Gieger C, Martin NG, Waldenberger M, Siscovick DS, Raitakari O, Eriksson JG, Mitchell P, Hunter DJ, Kraft P, Rimm EB, Boomsma DI, Borecki IB, Loos RJ, Wareham NJ, Vollenweider P, Caporaso N, Grabe HJ, Neuhauser ML, Wolfenbutter LH, Hu FB, Hyppönen E, Jarvelin MR, Cupples LA, Banks PW, Ridker PM, van Duijn CM, Heiss G, Metspalu A, North KE, Ingelsson E, Nettleton JA, van Dam RM, Chasman DI. Genome-wide

- meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol Psychiatr* May 2015;20(5):647–56.
- [193] Mahajan A, Rodan AR, Le TH, Gaulton KJ, Haessler J, Stilp AM, Kamatani Y, Zhu G, Sofer T, Puri S, Schellinger JN, Chu PL, Cechova S, van ZN, Arnlöv J, Flessner MF, Giedraitis V, Heath AC, Kubo M, Larsson A, Lindgren CM, Madden PA, Montgomery GW, Papanicolaou GJ, Reiner AP, Sundstrom J, Thornton TA, Lind L, Ingelsson E, Cai J, Martin NG, Kooperberg C, Matsuda K, Whitfield JB, Okada Y, Laurie CC, Morris AP, Franceschini N. Trans-ethnic fine mapping highlights kidney-function genes linked to salt sensitivity. *Am J Hum Genet* September 1, 2016;99(3):636–46.
- [194] Hu Y, Tanaka T, Zhu J, Guan W, Wu JHY, Psaty BM, McKnight B, King IB, Sun Q, Richard M, Manichai-kul A, Frazier-Wood AC, Kabagambe EK, Hopkins PN, Ordovas JM, Ferrucci L, Bandinelli S, Arnett DK, Chen YI, Liang S, Siscovick DS, Tsai MY, Rich SS, Fornage M, Hu FB, Rimm EB, Jensen MK, Lemaitre RN, Mozaffarian D, Steffen LM, Morris AP, Li H, Lin X. Discovery and fine-mapping of loci associated with MUFAs through trans-ethnic meta-analysis in Chinese and European populations. *J Lipid Res* May 2017;58(5):974–81.
- [195] Liu CT, Raghavan S, Maruthur N, Kabagambe EK, Hong J, Ng MC, Hivert KF, Lu Y, An P, Bentley AR, Drolet AM, Gaulton KJ, Guo X, Armstrong LL, Irvin MR, Li M, Lipovich L, Rybin DV, Taylor KD, Agyemang C, Palmer ND, Cade BE, Chen WM, Dauriz M, Delaney JA, Edwards TL, Evans DS, Evans MK, Lange LA, Leong A, Liu J, Liu Y, Nayak U, Patel SR, Porneala BC, Rasmussen-Torvik LJ, Snijder MB, Stallings SC, Tanaka T, Yanek LR, Zhao W, Becker DM, Bielak LF, Biggs ML, Bottinger EP, Bowden DW, Chen G, Correa A, Couper DJ, Crawford DC, Cushman M, Eicher JD, Fornage M, Franceschini N, Fu YP, Goodarzi MO, Gottesman O, Hara K, Harris TB, Jensen RA, Johnson AD, Jhun MA, Karter AJ, Keller MF, Kho AN, Kizer JR, Krauss RM, Langefeld CD, Li X, Liang J, Liu S, Lowe Jr WL, Mosley TH, North KE, Pacheco JA, Peyser PA, Patrick AL, Rice KM, Selvin E, Sims M, Smith JA, Tajuddin SM, Vaidya D, Wren MP, Yao J, Zhu X, Ziegler JT, Zmuda JM, Zonderman AB, Zwinderman AH, AAAG Consortium, CARE Consortium, COGENT-BP Consortium, eMERGE Consortium, MEDIA Consortium, Adeyemo A, Boerwinkle E, Ferrucci L, Hayes MG, Kardia SL, Miljkovic I, Pankow JS, Rotimi CN, Sale MM, Wagenknecht LE, Arnett DK, Chen YD, Nalls MA, MAGIC Consortium, Province MA, Kao WH, Siscovick DS, Psaty BM, Wil-son JG, Loos RJ, Dupuis J, Rich SS, Florez JC, Rotter JJ, Morris AP, Meigs JB. Trans-ethnic meta-analysis and functional annotation illuminates the genetic architecture of fasting glucose and insulin. *Am J Hum Genet* July 7, 2016;99(1):56–75. <https://doi.org/10.1016/j.ajhg.2016.05.006>. [Epub 2016 Jun 16].
- [196] Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. *J Comput Biol* December 2013;20(12):970–8. <https://doi.org/10.1089/cmb.2012.0283>. [Epub 2013 Aug 20].
- [197] Guo Y, Zhao S, Li CI, Sheng Q, Shyr Y. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Canc Inf* October 13, 2014;13(Suppl. 6):1–5. <https://doi.org/10.4137/CIN.S17688>. [eCollection 2014. Review].
- [198] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* April 25, 2013;14(4):R36. <https://doi.org/10.1186/gb-2013-14-4-r36>.
- [199] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* January 1, 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>. [Epub 2012 Oct 25].
- [200] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10(3):R25. <https://doi.org/10.1186/gb-2009-10-3-r25>. [Epub 2009 Mar 4].
- [201] Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;9(8):e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>. [Epub 2013 Aug 8].
- [202] Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* January 15, 2015;31(2):166–9. <https://doi.org/10.1093/bioinformatics/btu638>. [Epub 2014 Sep. 25].
- [203] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* July 2008;5(7):621–8. <https://doi.org/10.1038/nmeth.1226>. [Epub 2008 May 30].
- [204] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol* January 26, 2016;17:13. <https://doi.org/10.1186/s13059-016-0881-8>. [Erratum in: *Genome Biol*. 2016;17(1):181].

- [205] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf* August 4, 2011;12:323. <https://doi.org/10.1186/1471-2105-12-323>.
- [206] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;11(3):R25. <https://doi.org/10.1186/gb-2010-11-3-r25>. [Epub 2010 Mar 2].
- [207] Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;11(10):R106. <https://doi.org/10.1186/gb-2010-11-10-r106>. [Epub 2010 Oct 27].
- [208] Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf* February 18, 2010;11:94. <https://doi.org/10.1186/1471-2105-11-94>.
- [209] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrézic F, French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNAsequencing data analysis. *Briefings Bioinf* November 2013;14(6):671–83. <https://doi.org/10.1093/bib/bbs046>. [Epub 2012 Sep. 17].
- [210] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. *BMC Bioinf* December 17, 2011;12:480. <https://doi.org/10.1186/1471-2105-12-480>.
- [211] Chung LM, Ferguson JP, Zheng W, Qian F, Bruno V, Montgomery RR, Zhao H. Differential expression analysis for paired RNA-Seq data. *BMC Bioinf* March 27, 2013;14:110. <https://doi.org/10.1186/1471-2105-14-110>.
- [212] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* November 1, 2007;23(21):2881–7. [Epub 2007 Sep. 19].
- [213] Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* April 2008;9(2):321–32. [Epub 2007 Aug 29].
- [214] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* May 2012;40(10):4288–97. <https://doi.org/10.1093/nar/gks042>. [Epub 2012 Jan 28].
- [215] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* January 1, 2010;26(1):139–40. <https://doi.org/10.1093/bioinformatics/btp616>. [Epub 2009 Nov 11].
- [216] Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* September 2013;8(9):1765–86. <https://doi.org/10.1038/nprot.2013.099>. [Epub 2013 Aug 22].
- [217] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.
- [218] Krämer A, Green J, Pollard Jr J, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* February 15, 2014;30(4):523–30. <https://doi.org/10.1093/bioinformatics/btt703>. [Epub 2013 Dec 13].
- [219] Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* July 3, 2017;45(W1):W130–7. <https://doi.org/10.1093/nar/gkx356>.
- [220] Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* July 2007;35(Web Server issue):W169–75. [Epub 2007 Jun 18].
- [221] Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 2007;8(9):R183.
- [222] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4(1):44–57. <https://doi.org/10.1038/nprot.2008.211>.
- [223] Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* January 2009;37(1):1–13. <https://doi.org/10.1093/nar/gkn923>. [Epub 2008 Nov 25].
- [224] Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genom* May 1, 2013;14:293. <https://doi.org/10.1186/1471-2164-14-293>.
- [225] Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinf* November 30, 2010;11:587. <https://doi.org/10.1186/1471-2105-11-587>.

- [226] Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* May 15, 2014;30(10):1363–9. <https://doi.org/10.1093/bioinformatics/btu049>. [Epub 2014 Jan 28].
- [227] Fortin JP, Triche Jr TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina Human-MethylationEPIC array with minfi. *Bioinformatics* February 15, 2017;33(4):558–60. <https://doi.org/10.1093/bioinformatics/btw691>.
- [228] Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* February 1, 2014;30(3):428–30. <https://doi.org/10.1093/bioinformatics/btt684>. [Epub 2013 Dec 12].
- [229] Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* November 2014;11(11):1138–40. <https://doi.org/10.1038/nmeth.3115>. [Epub 2014 Sep. 28].
- [230] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* August 2005;37(8):853–62. [Epub 2005 Jul 10].
- [231] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiesen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeecx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* July 8, 2010;466(7303):253–7. <https://doi.org/10.1038/nature09165>.
- [232] Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res* January 2010;38(2):391–9. <https://doi.org/10.1093/nar/gkp992>. [Epub 2009 Nov 11].
- [233] Rauch TA, Pfeifer GP. The MIRA method for DNA methylation analysis. *Meth Mol Biol* 2009;507:65–75.
- [234] Rauch T, Pfeifer GP. Methods for assessing genome-wide DNA methylation. In: Tollefsbol T, editor. *Handbook of Epigenetics*. Amsterdam, The Netherlands: Elsevier; 2010. p. 135–47.
- [235] Jung M, Kadam S, Xiong W, Rauch TA, Jin SG, Pfeifer GP. MIRA-seq for DNA methylation analysis of CpG islands. *Epigenomics* August 2015;7(5):695–706. <https://doi.org/10.2217/epi.15.33>. [Epub 2015 Apr 17].
- [236] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* October 13, 2005;33(18):5868–77.
- [237] Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* April 2011;6(4):468–81. <https://doi.org/10.1038/nprot.2010.190>. [Epub 2011 Mar 18].
- [238] Wang G, Luo X, Wang J, Wan J, Xia S, Zhu H, Qian J, Wang Y. MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res* January 4, 2018;46(D1):D146–51. <https://doi.org/10.1093/nar/gkx1096>.
- [239] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* June 1, 2011;27(11):1571–2. <https://doi.org/10.1093/bioinformatics/btr167>. [Epub 2011 Apr 14].
- [240] Hansen KD, Langmead B, Irizarry RA. BSsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* October 3, 2012;13(10):R83. <https://doi.org/10.1186/gb-2012-13-10-r83>.
- [241] Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* April 2014;42(8):e69. <https://doi.org/10.1093/nar/gku154>. [Epub 2014 Feb 22].
- [242] Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X. Statistical methods for detecting differentially methylated loci and regions. *Front Genet* September 16, 2014;5:324. <https://doi.org/10.3389/fgene.2014.00324>. [eCollection 2014. Review].
- [243] Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* July 1, 2013;29(13):1647–53. <https://doi.org/10.1093/bioinformatics/btt263>. [Epub 2013 May 8].
- [244] Hebestreit K, Klein H. BiSeq: processing and analyzing bisulfite sequencing data. 2015. R package version 1.18.0.
- [245] Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol* February 24, 2014;15(2):R38. <https://doi.org/10.1186/gb-2014-15-2-r38>.
- [246] Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing

- experiments. *BMC Bioinf* June 24, 2014;15:215. <https://doi.org/10.1186/1471-2105-15-215>.
- [247] Park Y, Figueroa ME, Rozek LS, Sartor MA. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* September 1, 2014;30(17):2414–22. <https://doi.org/10.1093/bioinformatics/btu339>. [Epub 2014 May 16].
- [248] Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology* January 6, 2016;5(1). <https://doi.org/10.3390/biology5010003>. pii: E3 [Review].
- [249] Akalin A, Franke V, Vlahovick K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* April 1, 2015;31(7):1127–9. <https://doi.org/10.1093/bioinformatics/btu775>. [Epub 2014 Nov 21].
- [250] Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinf* May 11, 2010;11:237. <https://doi.org/10.1186/1471-2105-11-237>.
- [251] Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteome data. *BMC Syst Biol* 2014;8(Suppl. 2):S3.
- [252] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* October 2002;74(20):5383–92.
- [253] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* December 2011;10(12):M111. 007690.
- [254] Li XJ, Zhang H, Ranish JA, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem* December 2003;75(23):6648–57.
- [255] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* September 2003;75(17):4646–58.
- [256] Li X-J, Pedrioli PGA, Eng J, Martin D, Yi EC, Lee H, Aebersold R. A tool to visualize and evaluate data obtained by liquid chromatography/electrospray ionization/mass spectrometry. *Anal Chem* July 2004;76(13):3856–60.
- [257] Shteynberg D, Mendoza L, Hoopmann MR, Sun Z, Schmidt F, Deutsch EW, Moritz RL. reSpect: software for identification of high and low abundance ion species in chimeric tandem mass spectra. *J Am Soc Mass Spectrom* November 2015;26(11):1837–47.
- [258] Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii A, Aebersold R. A guided tour of the trans-proteomic pipeline. *Proteomics* March 2010;10(6):1150–9.
- [259] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25–9.
- [260] Alonso R, Salavert F, Garcia-Garcia F, Carbonell-Caballero J, Bleda M, Garcia-Alonso L, Sanchis-Juan A, Perez-Gil D, Marin-Garcia P, Sanchez R, Cubuk C, Hidalgo MR, Amadoz A, Hernansaiz-Ballesteros RD, Alemán A, Tarraga J, Montaner D, Medina I, Dopazo J. Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic Acids Res* 2015;43(W1):W117–21.
- [261] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics* December 1, 2007;23(23):3251–3.
- [262] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* October 2005;102(43):15545–50.
- [263] Malik R, Dulla K, Nigg E, Körner R. From proteome lists to biological impact - tools and strategies for the analysis of large MS data sets. *Proteomics* 2010;10:1270–83.
- [264] Bader G, Cary M, Sander C. Pathguide: a pathway resource list. *Nucleic Acids Res* 2006;34(Database):D504–6.
- [265] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215(3):403–10.
- [266] Desiere F, Deutsch EW, Nesvizhskii AI, Mallik P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* 2004;6.
- [267] Falkner JA, Ulintz PJ, Andrews PC. A code and data archival and dissemination tool for the proteomics community. *Am Biotechnol Lab* 2006.
- [268] Vizcaíno AJ, Côté RG, Csordas A, Dienes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 2012;41(D1):D1063–9.

- [269] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 1995;57(No. 1):289–300.
- [270] Benjamini Y, Daniel Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* Aug., 2001;29(No. 4):1165–88.
- [271] Storey JD. The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann Stat* 2003;31(Num-ber 6):2013–35.
- [272] Storey JD. False discovery rate. In: Lovric M, editor. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer; 2011.
- [273] Fonville JM, Richards SE, Barton RH, Boulange CL, Ebbels TMD, Nicholson JK, Holms E, Dumas ME. The evolution of partial least squares models and related chemometric approaches in metabolomics and metabolic phenotyping. *J Chemometr* 2010;24(11–12):636–49. <https://doi.org/10.1002/cem.1359>.
- [274] Barker M, Rayens W. Partial least squares for discrimination. *J Chemometr* 2003;17:166–73. <https://doi.org/10.1002/chem.785>.
- [275] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemom* 2002;16:119–28. <https://doi.org/10.1002/cem.695>.
- [276] Smith C, Want EJ, O'Maille G, Abagyan R, Siuzdak GE. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 2006;78:779.
- [277] Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. *Nucleic Acids Res* 2015;43:W251.
- [278] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* May 15, 2012;28(10):1353–8. <https://doi.org/10.1093/bioinformatics/bts163>. [Epub 2012 Apr 6].
- [279] Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* October 15, 2011;27(20):2917–8. <https://doi.org/10.1093/bioinformatics/btr499>. [Epub 2011 Sep. 4].
- [280] Sun H, Wang H, Zhu R, Tang K, Gong Q, Cui J, Cao Z, Liu Q. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics* March 1, 2014;30(5):737–9. <https://doi.org/10.1093/bioinformatics/btt576>. [Epub 2013 Oct 3].
- [281] Gosline SJ, Oh C, Fraenkel E. SAMNetWeb: identifying condition-specific networks linking signaling and transcription. *Bioinformatics* April 1, 2015;31(7):1124–6. <https://doi.org/10.1093/bioinformatics/btu748>. [Epub 2014 Nov 19].
- [282] Wachter A, Beißbarth T. pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics* September 15, 2015;31(18):3072–4. <https://doi.org/10.1093/bioinformatics/btv323>. [Epub 2015 May 21].
- [283] Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, Sartor MA, Stringer KA, Jagadish HV, Burant C, Athey B, Omenn GS. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* February 1, 2012;28(3):373–80. <https://doi.org/10.1093/bioinformatics/btr661>. [Epub 2011 Nov 30].
- [284] Wanichthanarak K, Fahrman JF, Grapov D. Genomic, proteomic, and metabolomic data integration strategies. *Biomark Insights* 2015;10:1–6. <https://doi.org/10.4137/BMI.S29511>.
- [285] Uppal K, Go Y-M, Jones DP. xMWAS: an R package for data-driven integration and differential network analysis. Mar. 30, 2017. <https://doi.org/10.1101/122432>. bioRxiv.

FURTHER READING

- Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* November 2010;52(3):232–6. <https://doi.org/10.1016/j.jymeth.2010.06.012>. [Epub 2010 Jun 11].
- Sham PC, Cherny SS, Purcell S, Hewitt JK. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* May 2000;66:1616–30.
- Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinf* July 27, 2009;10:232. <https://doi.org/10.1186/1471-2105-10-232>.

Population Genetics

*H. Richard Johnston¹, Bronya J.B. Keats²,
Stephanie L. Sherman¹*

¹Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, United States

²Department of Genetics (Emeritus), Louisiana State University Health Sciences Center, New Orleans, LA, United States

Population genetics is the study of genetic variation within and among populations and the evolutionary factors that explain this variation. Its foundation is the Hardy–Weinberg law, which is maintained as long as the population size is large, mating is at random, and mutation, selection, and migration are negligible. If these assumptions are violated, allele frequencies and genotype frequencies may change from one generation to the next. Ethnic variation in allele frequencies is found throughout the genome, and by examining this genetic diversity, evolutionary patterns can be inferred, and variants contributing to common diseases can be identified. As a result of major international initiatives, extensive databases containing millions of genetic variants are available. Together with automated technology for genotyping, sequencing, and bioinformatic analysis, these datasets provide the population geneticist with a huge set of densely mapped polymorphisms for reconciling genome variation with population histories of bottlenecks, admixture, and migration, for revealing evidence of natural selection, and for advancing the understanding of many diseases.

12.1 INTRODUCTION

With the monumental scientific advances that have resulted from the Human Genome Project, the genetic composition of populations can now be examined in detail. Thousands of rare alleles are known to be disease-causing variants and for most of these diseases

[e.g., cystic fibrosis, Tay–Sachs, phenylketonuria (PKU), hemophilia A, and familial hypercholesterolemia], many different pathogenic variants are found within the same gene. In general, the frequencies of these rare disease alleles differ among populations, as do the frequencies of common alleles (>1%), many of which are associated with common diseases such as Crohn's disease, diabetes, coronary disease, celiac disease, multiple sclerosis, and macular degeneration. The principles of population genetics attempt to explain the genetic diversity in present populations and the changes in allele and genotype frequencies over time. Population genetic studies facilitate the identification of alleles associated with disease risk and provide insight into the effect of medical intervention on the population frequency of a disease. Allele and genotype frequencies depend on factors such as mating patterns, population size and distribution, mutation, migration, and selection. By making specific assumptions about these factors, the Hardy–Weinberg law, a fundamental principle of population genetics, provides a model for calculating genotype frequencies from allele frequencies for a random mating population in equilibrium.

12.2 HARDY–WEINBERG LAW

The allele frequencies at a locus can always be calculated from the genotype frequencies, but the converse is not necessarily true. The Hardy–Weinberg law states that for a single autosomal locus in a large population in which

(1) mating takes place at random with respect to genotype, (2) allele frequencies are the same in males and females, and (3) mutation, selection, and migration are negligible, genotype frequencies can be calculated from allele frequencies after one generation, regardless of the allele and genotype frequencies in the initial population. This is not true for a single X-linked locus or for any set of loci considered jointly; for these loci, the establishment of this relationship between allele and genotype frequencies takes more than one generation.

12.2.1 Autosomal Locus

Consider a locus with two alleles, A_1 and A_2 , and suppose the population frequencies of the three genotypes A_1A_1 , A_1A_2 , A_2A_2 are p_{11} , p_{12} , p_{22} , respectively, where $p_{11} + p_{12} + p_{22} = 1$, then, in this initial population, the frequency of A_1 is $p_{11} + \frac{1}{2}p_{12}$ and the frequency of A_2 is $p_{22} + \frac{1}{2}p_{12}$. Random mating is approximately equivalent to random union of gametes. Thus, random mating within this initial population results in the following genotype frequencies in the next generation:

$$\text{Frequency of } A_1A_1 = (p_{11} + \frac{1}{2}p_{12})^2$$

$$\text{Frequency of } A_1A_2 = 2(p_{11} + \frac{1}{2}p_{12})(p_{22} + \frac{1}{2}p_{12})$$

$$\text{Frequency of } A_2A_2 = (p_{22} + \frac{1}{2}p_{12})^2$$

The genotype frequencies in this second generation may be different from those in the first generation. However, calculation of the allele frequencies from the genotype frequencies in the second generation gives

$$\begin{aligned} \text{Frequency of } A_1 &= (p_{11} + \frac{1}{2}p_{12})^2 + (p_{11} + \frac{1}{2}p_{12})(p_{22} + \frac{1}{2}p_{12}) \\ &= p_{11}(p_{11} + p_{12} + p_{22}) + \frac{1}{2}p_{12} \\ &= p_{11} + \frac{1}{2}p_{12} \end{aligned}$$

Similarly, the frequency of A_2 is $p_{22} + \frac{1}{2}p_{12}$, which is equal to $1 - (p_{11} + \frac{1}{2}p_{12})$. These allele frequencies are identical to those in the first generation. In other words, if the allele frequencies are $p = p_{11} + \frac{1}{2}p_{12}$ and $q = 1 - p = p_{22} + \frac{1}{2}p_{12}$, then after one generation of random mating, the genotype frequencies are p^2 , $2pq$, and q^2 . These frequencies are the Hardy–Weinberg proportions, and the population is said to be in Hardy–Weinberg equilibrium.

TABLE 12.1 Table Establishment of Equilibrium in One Generation for an Autosomal Locus

	Mating Frequency	OFFSPRING GENOTYPE FREQUENCIES		
		A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	$(0.1)^2$	$(0.1)^2$	0	0
$A_1A_1 \times A_1A_2$	$2(0.1)(0.2)$	$(0.1)(0.2)$	$(0.1)(0.2)$	0
$A_1A_1 \times A_2A_2$	$2(0.1)(0.7)$	0	$2(0.1)(0.7)$	0
$A_1A_2 \times A_1A_2$	$(0.2)^2$	$1/4(0.2)^2$	$1/2(0.2)^2$	$1/4(0.2)^2$
$A_1A_2 \times A_2A_2$	$2(0.2)(0.7)$	0	$(0.2)(0.7)$	$(0.2)(0.7)$
$A_2A_2 \times A_2A_2$	$(0.7)^2$	0	0	$(0.7)^2$
Total	1	0.04	0.32	0.64

Table 12.1 presents a numerical example, in which the initial population comprises 20, 40, and 140 individuals with genotypes A_1A_1 , A_1A_2 , and A_2A_2 , respectively. The genotype frequencies are

$$\text{Frequency } A_1A_1 = p_{11} = 20/200 = 0.10$$

$$\text{Frequency } A_1A_2 = p_{12} = 40/200 = 0.20$$

$$\text{Frequency } A_2A_2 = p_{22} = 140/200 = 0.70$$

and, therefore, the allele frequencies are

$$\text{Frequency } A_1 = 0.10 + \frac{1}{2}(0.20) = 0.2$$

$$\text{Frequency } A_2 = 0.70 + \frac{1}{2}(0.20) = 0.8$$

Random union of gametes results in the following genotype frequencies in the next generation:

$$\text{Frequency } A_1A_1 = 0.2^2 = 0.04$$

$$\text{Frequency } A_1A_2 = 2(0.2)(0.8) = 0.32$$

$$\text{Frequency } A_2A_2 = 0.8^2 = 0.64$$

Note that these genotype frequencies are different from those in the initial population. To confirm that these results are correct, Table 12.1 shows the genotype frequencies in the offspring that result from each of the

six possible mating types. For example, all the offspring of the mating $A_1A_1 \times A_1A_1$ must be A_1A_1 , while for the mating type $A_1A_1 \times A_1A_2$, each of the two offspring genotypes, A_1A_1 and A_1A_2 , has a probability of one half (Table 12.1). Summing-up the columns in Table 12.1 gives the frequencies of each of the genotypes in the second generation. These frequencies are the same as those obtained by random union of gametes, and the allele frequencies calculated from these genotype frequencies are:

$$\text{Frequency } A_1 = 0.04 + \frac{1}{2} (0.32) = 0.2$$

$$\text{Frequency } A_2 = 0.64 + \frac{1}{2} (0.32) = 0.8$$

Repeating these steps will give identical genotype and allele frequencies in the third generation to those in the second generation. Note that only the genotype frequencies change in the establishment of equilibrium; the allele frequencies in the initial population remain the same in subsequent generations.

The chi-square goodness-of-fit test may be used to determine whether the observed numbers of each genotype are significantly different from those expected under Hardy–Weinberg equilibrium. The total number of individuals is 200, so the expected numbers for the three genotypes are $(0.2)^2 200 = 8$, $2(0.2)(0.8) 200 = 64$, and $(0.8)^2 200 = 128$, compared with the observed numbers of 20, 40, and 140. The test value is $(20 - 8)^2/8 + (40 - 64)^2/64 + (140 - 128)^2/128 = 28$. This value is compared to the chi-square distribution with one degree of freedom. (There are three classes, but the total number of individuals is known and also the allele frequencies are known. Therefore, there is only one independent class and one degree of freedom.) In general, the number of degrees of freedom is equal to the number of genotypes minus the number of alleles. The 99.9th percentile of the chi-square distribution with one degree of freedom is 10.83.

Thus, the observed numbers of each genotype in the initial population are significantly different at the 1% level from those expected under Hardy–Weinberg equilibrium. However, after one generation of random mating, the observed and expected numbers are the same.

Calculation of allele frequencies from genotype frequencies is straightforward when all three genotypes are observable, but, in the case of recessive diseases, such as cystic fibrosis, only two phenotype classes are observed. If, however, equilibrium is assumed, the frequency of affected

individuals is q^2 ; thus, the square root of this frequency is the frequency of the disease allele. The frequency of heterozygotes (carriers) is $2(1 - q)q$, and the proportion of carriers among unaffected individuals in the population is

$$[2(1 - q)q] / (1 - q^2) = 2q / (1 + q)$$

For example, in populations of European ancestry, the frequency of cystic fibrosis is estimated to be 1/2000; thus, the frequency of the abnormal allele is 0.022 and the normal allele is 0.978. The frequency of heterozygotes is therefore $2 \times 0.022 \times 0.978$, which is about 1/23. That is, approximately 4% of the populations are carriers, but less than 0.1% are affected. Several different pathogenic variants have been described in the cystic fibrosis gene. Each one of these is a disease allele; thus, the frequency of 0.022 is actually the sum of the frequencies of all the disease alleles in the cystic fibrosis gene.

The Hardy–Weinberg principle may be extended to more than two alleles. In general, for n alleles, A_1, A_2, \dots, A_n , with frequencies p_1, p_2, \dots, p_n , the genotype frequencies are p_i^2 for homozygotes A_iA_i and $2p_i p_j$ for heterozygotes A_iA_j . The heterozygosity value (H) for a locus is the total frequency of heterozygotes, and it may be written as

$$H = \sum 2p_i p_j = 1 - \sum p_i^2$$

For two alleles, the maximum heterozygosity is 0.5, for five alleles it is 0.8, and for 10 alleles it is 0.9. In other words, for a locus to have a heterozygosity of 80%, it must have at least five alleles. (The maximum heterozygosity is reached when the alleles have equal frequencies.)

Example

Suppose a locus has five alleles (designated 1, 2, 3, 4, 5) with frequencies 0.5, 0.3, 0.1, 0.08, 0.02. What are the genotype frequencies when Hardy–Weinberg equilibrium is established? What is the heterozygosity value (H) at this locus?

With n alleles, there are $n(n + 1)/2$ genotypes. Thus, for five alleles there are 15 genotypes. The frequencies of the five homozygotes, 1–1, 2–2, 3–3, 4–4, 5–5, are 0.25, 0.09, 0.01, 0.0064, 0.0004, respectively. The frequencies of the 10 heterozygotes, 1–2, 1–3, 1–4, 1–5, 2–3, 2–4, 2–5, 3–4, 3–5, 4–5, are 0.3, 0.1, 0.08, 0.02, 0.06, 0.048, 0.012, 0.016, 0.004, 0.0032, respectively.

$$\begin{aligned} \text{Heterozygosity } (H) &= \\ 1 - (0.25 + 0.09 + 0.01 + 0.0064 + 0.0004) &= 0.6432 \end{aligned}$$

TABLE 12.2 Approach to Equilibrium for a Locus on the X Chromosome

Generation	p_m	p_f
0	0.33	0.57
1	0.57	0.45
2	0.45	0.51
3	0.51	0.48
4	0.48	0.495
5	0.495	0.4875
6	0.4875	0.49125
7	0.49125	0.489375
8	0.489375	0.4903125
9	0.4903125	0.48984375
10	0.48984375	0.490078125
11	0.490078125	0.4899609375
12	0.4899609375	0.49001953125
—		
—		
—		
Equilibrium	0.49	0.49

12.2.2 X-Linked Locus

The genotype frequencies at a locus on the X chromosome differ in the two sexes because males have only one X chromosome, whereas females have two X chromosomes. Thus, in males, the genotype frequency is equal to the allele frequency. For Hardy–Weinberg equilibrium, the allele frequencies in males must be equal to those in females. Suppose the frequency of the A_1 allele is p_m in males and p_f in females in the first generation. By the principles of X-linked inheritance, the frequency of this allele in males in the second generation must be p_f because males get their X chromosomes from their mothers. By contrast, for females in the second generation the frequency of the A_1 allele is $\frac{1}{2}(p_m + p_f)$ because females get one X chromosome from each parent. The difference between the male and female frequencies in this generation is $p_f - \frac{1}{2}(p_m + p_f) = \frac{1}{2}(p_f - p_m)$, which is one half of the difference in the first generation. Similarly, in the third generation, the male allele frequency is $\frac{1}{2}(p_m + p_f)$, while the female frequency is $\frac{1}{4}p_m + \frac{3}{4}p_f$ and the difference is $\frac{1}{4}(p_f - p_m)$. With each generation, the difference between the male and female frequencies becomes smaller, and equilibrium is reached when they are the same. The equilibrium allele frequency of A_1 is equal to $\frac{2}{3}p_f + \frac{1}{3}p_m$ in both sexes. Table 12.2 shows the approach to equilibrium for an X-linked locus

TABLE 12.3 Genotype Equilibrium Frequencies for an X-Linked Locus

OFFSPRING GENOTYPE FREQUENCIES						
	Mating Frequency	MALE		FEMALE		
		A_1	A_2	A_1A_1	A_1A_2	A_2A_2
$A_1 \times A_1A_1$	p^3	p^3	0	p^3	0	0
$A_1 \times A_1A_2$	$2p^2q$	p^2q	p^2q	p^2q	p^2q	0
$A_1 \times A_2A_2$	pq^2	0	pq^2	0	pq^2	0
$A_2 \times A_1A_1$	p^2q	p^2q	0	0	p^2q	0
$A_2 \times A_1A_2$	$2pq^2$	pq^2	0	0	pq^2	pq^2
$A_2 \times A_2A_2$	q^3	0	q^3	0	0	q^3
Total		p	q	p^2	$2pq$	q^2

when the initial allele frequencies are 0.33 in males and 0.57 in females. With each generation, the difference between the frequencies in males and females is reduced, and they approach the equilibrium frequency of $\frac{1}{3}(0.33) + \frac{2}{3}(0.57) = 0.49$. If the frequencies of the two alleles at the locus (A_1 and A_2) are $p = \frac{1}{3}p_m + \frac{2}{3}p_f$ and $q = 1 - p$, the equilibrium genotype frequencies are p and q in males and p^2 , $2pq$, and q^2 in females. Table 12.3 gives the frequency of each possible mating type and the expected offspring genotype frequencies for males and females. Summing-up these genotype frequencies shows that the equilibrium frequencies are maintained in the next generation (Table 12.3).

Example

Suppose the frequency of an allele at an X-linked locus is 0.03 in males and 0.06 in females. What is the equilibrium allele frequency? Furthermore, suppose that this allele is responsible for a recessive trait. What are the equilibrium frequencies of this trait in males and females, and what is the frequency of heterozygous (carrier) females?

The equilibrium frequency of the allele is $\frac{1}{3}(0.03) + \frac{2}{3}(0.06) = 0.05$. Thus, the frequency of males with the trait is 0.05 and the frequency of females with the trait is $(0.05)^2 = 0.0025$. The frequency of carrier females is $2(0.05)(0.95) = 0.095$.

12.2.3 Two Loci

Equilibrium is reached after one generation of random mating for a single autosomal locus and over several generations for an X-linked locus. However, the approach to equilibrium may be much longer for two loci considered

TABLE 12.4 Joint Genotype Frequencies for Two Loci

Genotype	Frequency	Equilibrium Frequency
$A_1A_1B_1B_1$	g_{11}^2	$p_1^2 q_1^2$
$A_1A_1B_1B_2$	$2g_{11}g_{12}$	$2p_1^2 q_1 q_2$
$A_1A_1B_2B_2$	g_{12}^2	$p_1^2 q_2^2$
$A_1A_2B_1B_1$	$2g_{11}g_{21}$	$2p_1 p_2 q_1^2$
$A_1A_2B_1B_2$	$2g_{11}g_{22} + 2g_{12}g_{21}$	$4p_1 p_2 q_1 q_2$
$A_1A_2B_2B_2$	$2g_{12}g_{22}$	$2p_1 p_2 q_2^2$
$A_2A_2B_1B_1$	g_{21}^2	$p_2^2 q_1^2$
$A_2A_2B_1B_2$	$2g_{21}g_{22}$	$2p_2^2 q_1 q_2$
$A_2A_2B_2B_2$	g_{22}^2	$p_2^2 q_2^2$

jointly, and the number of generations depends on the recombination fraction. Suppose the first locus has alleles A_1 and A_2 , with frequencies p_1 and p_2 , and the second locus has alleles B_1 and B_2 , with frequencies q_1 and q_2 , respectively. The four possible gametes are A_1B_1 , A_1B_2 , A_2B_1 , A_2B_2 ; let their frequencies in the population be g_{11} , g_{12} , g_{21} , g_{22} , where $p_1 = g_{11} + g_{12}$, $p_2 = g_{21} + g_{22}$, $q_1 = g_{11} + g_{21}$, and $q_2 = g_{12} + g_{22}$. Allowing these gametes to unite at random gives the genotype frequencies in the next generation (Table 12.4). Now consider the gametic output of this population. In doing so, we must take into account the fact that the frequency of gametes produced by the double heterozygote ($A_1A_2B_1B_2$) depends on the recombination fraction, θ (Table 12.4). If the phase is A_1B_1/A_2B_2 , then A_1B_1 and A_2B_2 are nonrecombinants, and A_1B_2 and A_2B_1 are recombinants. Conversely, if phase is A_1B_2/A_2B_1 , then A_1B_2 and A_2B_1 are nonrecombinants, and A_1B_1 and A_2B_2 are recombinants. Therefore, the frequency of A_1B_1 gametes from double heterozygotes is $g_{11}g_{22}(1 - \theta) + g_{12}g_{21}\theta$. In addition, all the gametes produced by individuals with the genotype $A_1A_1B_1B_1$, and one half of those produced by individuals with the genotypes $A_1A_1B_1B_2$ and $A_1A_2B_1B_1$ will be A_1B_1 . Thus, the total frequency of A_1B_1 gametes in this generation is $g_{11}^2 + g_{11}g_{12} + g_{11}g_{21} + g_{11}g_{22}(1 - \theta) + g_{12}g_{21}\theta$, which may be written as $g_{11} - \theta D$, where $D = g_{11}g_{22} - g_{12}g_{21}$. D is called the coefficient of linkage disequilibrium (LD) and is a measure of allelic association. Similar calculations may be done for each of the gametic types, and the

frequencies obtained are $g_{12} + \theta D$, $g_{21} + \theta D$, and $g_{22} - \theta D$ for A_1B_2 , A_2B_1 , and A_2B_2 , respectively.

If the loci are unlinked, $\theta = 1/2$, and the change in gametic frequency from one generation to the next is $1/2 D$. For linked loci the change is θD . Thus, the more closely two loci are linked, the slower is the approach to equilibrium. The coefficient of LD after t generations may be written as

$$D_t = (1 - \theta) D_{t-1} = (1 - \theta)^t D_0$$

which approaches zero as t trends to infinity.

At equilibrium, D is equal to zero and the genotype and gametic frequencies are products of the allele frequencies (see Table 12.4). The gametic frequencies may be written as $g_{11} = p_1 q_1 + D$, $g_{12} = p_1 q_2 - D$, $g_{21} = p_2 q_1 - D$, and $g_{22} = p_2 q_2 + D$. Each of these gametic frequencies must be greater than or equal to zero. Thus, D must be greater than or equal to both $-p_1 q_1$ and $-p_2 q_2$, and D must be less than or equal to both $p_1 q_2$ and $p_2 q_1$. These results may be written as

$$D_{\min} = \max(-p_1 q_1, -p_2 q_2)$$

$$D_{\max} = \min(p_1 q_2, p_2 q_1)$$

For two loci each with two alleles, D must lie between -0.25 and 0.25 , and it can reach these extreme values only if the frequencies of the four alleles are 0.5 . Thus, the value of D is dependent on allele frequencies, meaning that D values for different pairs of loci are not comparable. The value of the standardized measure, $D' = D/D_{\text{extreme}}$, where $D_{\text{extreme}} = -D_{\min}$ if $D < 0$ and D_{\max} if $D > 0$, is less dependent on the allele frequencies and lies between -1 and 1 .

The statistic, δ , is another measure of LD that is useful for estimating the location of a disease locus if a single mutation is likely. The formula is

$$\delta = (p_D - p_N) / (1 - p_N)$$

where p_D is the frequency of the associated allele on disease chromosomes and p_N is the frequency of this allele on normal chromosomes. This value represents an estimate of the proportion of disease chromosomes bearing the original associated allele. If there is a single mutation, the proportion of chromosomes carrying this mutation is the same for all marker loci, so differences in δ across loci should largely represent effects of recombination. Thus, δ can be used to determine the most likely location of the disease locus among a set of tightly linked marker loci.

The effectiveness of LD in locating disease mutations was demonstrated with the identification of the cystic fibrosis F508del mutation in 1989. Over the next decade, it was used successfully in the identification of ethnic-specific disease mutations in the Ashkenazi Jewish, Finnish, Acadian, Roma, and other isolated and endogamous populations.

Note that LD is one possible explanation for the association between a phenotype and a marker allele in a population. In this case, the disease locus is tightly linked to the marker locus. However, population association does not necessarily mean tight linkage, and vice versa. Other possible reasons for association are pleiotropy (multiple effects of the same gene), such as the association between stomach cancer and the A allele of the ABO blood group, and departures from random mating due to events such as racial admixture, stratification, inbreeding, and assortative mating.

Examples

1. Suppose the frequencies of the gametes A_1B_1 , A_1B_2 , A_2B_1 , A_2B_2 are 0.5, 0.1, 0.3, 0.1, respectively. What is the value of D after one generation of random mating if (1) the two loci are unlinked and (2) the recombination fraction between the two loci is 0.01?
The value of D in the original population is $(0.5)(0.1) - (0.1)(0.3) = 0.02$. After one generation, $D = (1 - 0.5)(0.02) = 0.01$ if the two loci are unlinked, and $D = (1 - 0.01)(0.02) = 0.0198$ if the recombination fraction is 0.01.
2. How many generations are required for the value of D to be one-half its initial value?
 $D_t/D_0 = (1 - \theta)^t = 1/2$; therefore, $t = \log(1/2)/\log(1 - \theta)$. Thus, for θ equal to 0.3, 0.1, 0.01, and 0.001, the numbers of generations required are approximately 2, 7, 69, and 693, respectively. Note that for unlinked loci, D is halved in one generation, as seen in Example 1.

12.3 FACTORS THAT AFFECT HARDY–WEINBERG EQUILIBRIUM

The assumption of a large, random mating population is fundamental to Hardy–Weinberg equilibrium. If mating is not at random, the allele frequencies at a locus (say, p and q) in the population do not change from one generation to the next, but the genotype frequencies are not p^2 , $2pq$, and q^2 . Evolutionary forces such as random genetic drift, mutation, selection, and migration, however, will change allele frequencies (and consequently genotype frequencies) from one generation to the next.

12.3.1 Factors That Affect Genotype Frequencies but Not Allele Frequencies

Random mating has been assumed so far in all the derivations. If gametes do not unite at random, the genotype frequencies are not in Hardy–Weinberg proportions and cannot be derived simply from allele frequencies. Consanguinity (inbreeding), assortative mating, and stratification (e.g., ethnic subgroups within a population) are examples of nonrandom mating. In these situations, the frequency of homozygotes is increased at the expense of heterozygotes, and the genotype frequencies may be significantly different from Hardy–Weinberg expectations. Allele frequencies, however, do not change.

12.3.1.1 Consanguinity and Inbreeding

Individuals who are related genetically are termed consanguineous, and the offspring of mating between such individuals are said to be inbred. Inbreeding increases the frequency of homozygous genotypes and decreases the frequency of heterozygous genotypes in the population. The offspring of consanguineous marriages have an increased risk of having recessive disorders over that of the general population. The increase in risk depends on the population frequency of the disease allele and the degree of relationship between the parents. In cultures in which uncle–niece and first- and second-cousin marriages are encouraged, recessive disorders that are rare in most randomly mating populations may be relatively common. The coefficient of inbreeding (F) for a child of a consanguineous marriage is the probability that the child receives two alleles at a given locus that are both from the same ancestor and are, thus, identical by descent (autozygous). For example, half first cousins share a grandparent in common. The probability that a child of half first cousins is homozygous by descent at a locus is $F = (1/2)^5 = 1/32$. In general, for autosomal loci, the inbreeding coefficient for an individual is $F = (1/2)^{(n_1 + n_2 + 1)}$, where n_1 and n_2 are the numbers of generations separating the individuals in the consanguineous mating from their common ancestor. (This formula assumes that the common ancestor is not inbred.) Half first cousins are separated from their common grandparent by two generations. Thus, the exponent is $2 + 2 + 1 = 5$. Table 12.5 gives the estimated proportion of alleles shared by consanguineous individuals that are identical by descent as well as the coefficient of inbreeding for the offspring of these consanguineous matings

TABLE 12.5 Proportion of Alleles Shared by Related Individuals That Are Identical by Descent and the Inbreeding Coefficient (F) in the Offspring of Various Types of Consanguineous Mating

Type of Mating	Proportion of Shared Alleles	F
Parent–offspring	1/2	1/4
Brother–sister	1/2	1/4
Half sibs	1/4	1/8
Uncle–niece, aunt–nephew	1/4	1/8
First cousins	1/8	1/16
Double first cousins	1/4	1/8
Half first cousins	1/16	1/32
First cousins once removed	1/16	1/32
Second cousins	1/32	1/64
Second cousins once removed	1/64	1/128
Third cousins	1/128	1/256

(Table 12.5). If a child is inbred through more than one line of descent, the total coefficient of inbreeding is the sum of each of the separate coefficients. For example, first cousins are related through two grandparents. Thus, the inbreeding coefficient for the offspring of first cousins is $F = (\frac{1}{2})^5 + (\frac{1}{2})^5 = (\frac{1}{2})^4 = 1/16$. The coefficient of inbreeding is also an estimate of the proportion of loci at which an individual is autozygous.

The coefficient of inbreeding for X-linked loci depends on the number of males in the lines of descent and is always zero for male offspring, because they have only one X chromosome. In order to calculate the inbreeding coefficient for daughters of first cousins, four possibilities need to be considered for the first cousins: their fathers are brothers, their mothers are sisters, the father of the male cousin and the mother of the female cousin are siblings, or vice versa. If the fathers are brothers, the first cousins cannot share any X-linked alleles in common because the male first cousin did not inherit an X chromosome from his father. Thus, female offspring of this type of first-cousin mating are not inbred for X-linked loci and have an inbreeding coefficient of zero. Similarly, if the first cousins are offspring of a brother and a sister with the father being the son of the brother and the mother being the daughter of the sister, the inbreeding coefficient for their daughters is zero because the first cousins cannot share any X-linked alleles in common.

On the other hand, if the mothers of the first cousins are sisters, then the inbreeding coefficient for X-linked loci in their daughters is greater than that for autosomal loci because a male transmits the X chromosome he received from his mother to all his daughters. Thus, the inbreeding coefficient in this situation is $(\frac{1}{2})^3 + (\frac{1}{2})^4 = 3/16$. The fourth possibility is that the first cousins are offspring of a brother and sister, with the sister being the mother of the male and the brother being the father of the female. In this case, the inbreeding coefficient for X-linked loci in female offspring is $(\frac{1}{2})^3 = 1/8$.

Genotype frequencies in inbred populations cannot be calculated from the allele frequencies alone, but they can be obtained if the average inbreeding coefficient in the population is known. The amount of inbreeding in the population may be measured in terms of the decrease in heterozygosity relative to a random mating population. If the allele frequencies at a locus are p and q , then under random mating the frequency of heterozygotes is $2pq$. Suppose the frequency of heterozygotes in the inbred population is H . Then the inbreeding coefficient for the population is $F = (2pq - H)/2pq$. Therefore, $H = 2pq - 2pqF$. The frequencies of the two types of homozygotes in the inbred population can then be calculated to be $p^2 + pqF$ and $q^2 + pqF$. If the inbreeding coefficient is zero (i.e., random mating), the genotype frequencies are those expected for Hardy–Weinberg equilibrium. On the other hand, if there is complete inbreeding ($F = 1$), the frequency of heterozygotes is zero, and the population consists only of homozygotes with frequencies of p and q . However, note that the allele frequencies will not change from one generation to the next, regardless of the value of the inbreeding coefficient in the population.

Example

Suppose the frequency of an autosomal recessive disease is 1/40,000 in the general population. What is the expected frequency of the disease among the offspring of first cousins?

The frequency of the deleterious allele is 1/200, the square root of the frequency of the disease. The inbreeding coefficient for offspring of first-cousin marriages is 1/16. Thus, the frequency of the disease among the offspring of first cousins is $1/40,000 + (199/200)(1/200)(1/16) = 1/2977$.

12.3.1.2 Assortative Mating

Assortative mating is the tendency for people to choose mates who are more similar (positive) or dissimilar

(negative) to themselves in phenotype characteristics than would be expected by chance. If these characteristics are genetically determined, positive assortative mating may increase homozygosity in the population. An important difference between inbreeding and positive assortative mating is that inbreeding affects all loci, while assortative mating affects only those that play a role in the phenotype characteristics that are similar. Clinical examples of positive assortative mating are those between individuals who are profoundly hearing impaired or blind, which in some cases may be attributable to the same genotypes.

12.3.1.3 Stratification

A stratified population is one in which mating occurs within subgroups, and thus mating is not random in the population (even though random mating may occur in each subgroup). Suppose there are two subgroups (A_1 and A_2) with allele frequencies of p_1 and q_1 , and p_2 and q_2 . Then the frequency of heterozygotes in the combined population ($A_1 + A_2$) is $p_1q_1 + p_2q_2$. If there was random mating in $A_1 + A_2$, then after one generation the frequency of heterozygotes would be $p_1q_1 + p_2q_2 + \frac{1}{2}(p_1 - p_2)^2$, which is always greater than $p_1q_1 + p_2q_2$, while the decrease in each of the two homozygote frequencies is $\frac{1}{4}(p_1 - p_2)^2$. Thus, if A_1 and A_2 remain as separate subgroups, the frequency of heterozygotes in the combined population will always be lower (and the frequency of homozygotes greater) than expected under Hardy–Weinberg equilibrium.

12.3.2 Factors That Affect Allele Frequencies

Evolutionary forces such as random genetic drift, mutation, selection, and migration change the allele frequencies in a population. Important examples of each of these forces have been documented in human populations, and their effects are becoming better understood as knowledge of the genetic structure of populations at the DNA level increases.

12.3.2.1 Random Genetic Drift

The Hardy–Weinberg principle assumes that population size is large, and this assumption is probably valid for many present-day populations. However, if the population size is small, allele frequencies may change from one generation to the next by chance alone. This change is a consequence of sampling in small populations and is called random genetic drift. The sample is the set of

gametes that contributes to the next generation. Suppose this sample consists of $2N$ gametes (N individuals) and consider a locus with two alleles, A_1 and A_2 . The $2N + 1$ possible values of the frequency of A_1 are

$$0, 1/2N, 2/2N, 3/2N, \dots, (2N-1)/2N, 2N/2N.$$

The probability that the number of A_1 alleles in the population is k ($0 \leq k \leq 2N$) depends on the population size and the frequencies, p and q , of A_1 and A_2 , respectively, in the previous generation. It may be written as

$$Pr(k) = \binom{2N}{k} p^k q^{2N-k}$$

Thus, if N and p are known, the probability of a particular frequency of A_1 in the next generation may be calculated. For example, if $N=50$ and $P=.5$, the probability that the frequency of A_1 in the next generation is less than 0.4 or greater than 0.6 is 0.023, while the probability that it is between 0.45 and 0.55 is 0.682. The probability that A_1 will either be lost or become fixed in the population in the next generation is extremely small, but is greater than zero. If $N=50$ and $P=.01$, the probability that A_1 will be lost in the next generation is 0.37, and the probability that it will have a frequency of greater than 0.05 is 0.002. The precise change in allele frequency from one generation to the next cannot be predicted because drift is a random process. However, over a number of generations, drift can lead to the loss of some alleles from the population, with others becoming fixed. If a large number of populations are considered, the average behavior of allele frequencies can be predicted. The probability that a new allele in a population will eventually become fixed is $\frac{1}{2N}$, the frequency of the allele in the population at the time it arose. If the allele is to become fixed in the population, the average time to fixation is approximately $4N$ generations. After a large enough number of generations of random genetic drift, every allele in a population can be traced back to a single allele in the initial ancestral population. All other alleles in the initial population will have been lost. This concept is known as coalescence, and it has been used to model DNA sequence variation in populations.

Random genetic drift in a population is similar to inbreeding and stratification, in that its effect on the population is a reduction in the number of heterozygotes and an increase in the number of homozygotes. When the population size is drastically reduced (a bottleneck), the genetic drift is known as a founder effect. Examples of this effect (e.g., new colonization by a small

subset of a population or environmental disasters such as plague and famine) abound in history. The founder effect is likely to explain the relatively high frequency of certain diseases in some ethnic groups (e.g., Tay-Sachs disease in the Ashkenazi Jewish population).

Example

Suppose a new mutation arises in a population of size 500. What is the probability that this allele will be lost in the next generation? What is the probability that it will eventually become fixed in the population?

The total number of gametes in the population is 1000. Thus, the frequency of the new allele is 0.001, and the probability that it will be lost in the next generation is $[1000!/(0!)(999!)](0.001)^{1000}(0.999)^0 = 0.001$. The probability that this new allele will eventually become fixed in the population is 1/1000.

12.3.2.2 Mutation

When mutations occur in the germ cells, they may be passed on to the next generation. The change in the DNA may be a single nucleotide substitution or it may involve many nucleotides, such as in the case of an insertion or deletion. Many hemoglobinopathies are due to point mutations that cause the replacement of an amino acid (missense) and consequently an abnormal protein product. The most common mutation causing Tay-Sachs disease is a 4-base-pair (bp) insertion (frameshift), while the F508del mutation in the cystic fibrosis gene is a 3-bp deletion.

The source of genetic variation in a population is mutation. Mutation rates in humans have been estimated to be on the order of 10^{-4} to 10^{-6} per gene per generation. The rate of nucleotide substitutions is estimated to be 1 in 10^8 per generation, implying that 30 nucleotide mutations would be expected in each human gamete.

Most new mutations are lost due to chance. However, new mutations arise in each generation, and some become established in the population. Suppose μ is the mutation rate from A_1 to A_2 per generation. If the frequencies of A_1 and A_2 are p_t and q_t , respectively, in generation t , then in the $(t+1)$ th generation the frequency of A_2 is

$$q_{t+1} = q_t + \mu p_t = q_t + \mu(1 - q_t) = \mu + (1 - \mu)q_t$$

assuming no back mutation.

Similarly, $q_t = \mu + (1 - \mu)q_{t-1}$, $q_{t-1} = \mu + (1 - \mu)q_{t-2}$, and so forth. By substitution, q_t may be written in terms of q_0 , the frequency of A_2 in the initial generation:

$$q_t = 1 - (1 - \mu)^t (1 - q_0)$$

$$\text{or } (1 - \mu)^t = (1 - q_t) / (1 - q_0) = p_t / p_0$$

Because μ is very small, $(1 - \mu)^t$ is approximately equal to $e^{-t\mu}$. Thus, the number of generations required to change the frequency of A_2 from q_0 to q_t is inversely proportional to the mutation rate. Also note that as t gets larger and larger, q_t gets closer and closer to 1. In other words, if mutation from A_1 to A_2 is the only force acting to change the allele frequencies, then A_2 will eventually become fixed in the population. The change in allele frequency from one generation to the next is $q_{t+1} - q_t = \mu(1 - q_t)$, meaning that the change in allele frequency is greater for smaller frequencies of A_2 .

So far we have considered mutation in only one direction. Now suppose the mutation rate from A_1 to A_2 is μ and the reverse rate from A_2 to A_1 is ν . Then the change in the frequency of A_2 per generation is $\mu p - \nu q$, and equilibrium is reached when this change is equal to zero. Thus, the equilibrium frequencies are $p = \nu / (\mu + \nu)$ and $q = \mu / (\mu + \nu)$. This equilibrium is stable, meaning that if the frequencies are disturbed, they will eventually return to their equilibrium values as long as no other forces are affecting them.

Mutation rates have been estimated for a number of autosomal dominant disorders, such as neurofibromatosis type I, which has the high rate of 10^{-4} , and tuberous sclerosis, with a rate of about 10^{-5} . Some of these disorders (e.g., achondroplasia, for which the mutation rate is estimated to be 10^{-5}) have reduced fitness, which is discussed in the next section.

Examples

1. How many generations will be required to change the frequency of A_2 (1) from 0.1 to 0.2, (2) from 0.8 to 0.9, if the mutation rate from A_1 to A_2 is 10^{-4} ?

The number of generations is

$$\begin{aligned} t &= 1/\mu [\ln(1 - q_0) - \ln(1 - q_t)] \\ &= 1/\mu [\ln(0.9) - \ln(0.8)] \\ &= 1/\mu (0.1178) \end{aligned}$$

Therefore, for a mutation rate of 10^{-4} , 1178 generations are required, whereas for a mutation rate of 10^{-5} , 11,780 generations are required to change the frequency of A_2 from 0.1 to 0.2. On the other hand, to change the frequency from 0.8 to 0.9 requires 6932

TABLE 12.6 Selection Against the A_2A_2 Genotype at an Autosomal Locus

Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency before selection	p^2	$2pq$	q^2
Relative fitness	1	1	$1 - s$
Frequency after selection	p^2	$2pq$	$q^2(1 - s)$
After one generation of selection			
Frequency $A_1 = (p^2 + pq)/[p^2 + 2pq + q^2(1 - s)] = p/(1 - sq^2)$			
Frequency $A_2 = [pq + (1 - s)q^2]/(1 - sq^2)$			
If $s = 1$ (i.e., complete selection against the A_2A_2 genotype), then			
After one generation of selection			
Frequency $A_1 = (p^2 + pq)/(p^2 + 2pq) = 1/(1 + q)$			
Frequency $A_2 = pq/(p^2 + 2pq) = q/(1 + q)$			
After t generations of selection			
Frequency $A_1 = [1 + (t - 1)q]/(1 + tq)$			
Frequency $A_2 = q/(1 + tq)$			

generations if the mutation rate is 10^{-4} and 69,315 generations if the mutation rate is 10^{-5} .

- Suppose the mutation rate from A_1 to A_2 is 10^{-4} and the reverse rate is 10^{-5} . What is the equilibrium frequency of A_1 ?

The equilibrium frequency of A_1 is $10^{-5}/(10^{-4} + 10^{-5}) = 0.091$. However, to reach this equilibrium frequency may take tens of thousands of generations, depending on the initial allele frequencies.

12.3.2.3 Selection

The fitness of an individual is defined as an ability to survive and reproduce. The process by which the frequencies of genotypes in individuals with greater fitness increase in the population is natural selection. It acts to decrease the frequencies of the less-fit genotypes. The relative fitness is defined as $1 - s$, where s is the selection coefficient against the deleterious genotype. Thus, the most-fit genotype has a relative fitness of 1 (and a selection coefficient of 0).

Consider the situation where there are three genotypes, A_1A_1 , A_1A_2 , A_2A_2 , at a locus with relative fitnesses of 1, 1, $1 - s$, respectively. That is, there is selection against the A_2A_2 homozygote. (If $s = 1$, the selection is complete, meaning that individuals with the A_2A_2 genotype do not reproduce.) Table 12.6 shows the change in allele frequencies from one generation to the next. In the case in which $s = 1$,

the frequencies after t generations can be written in terms of the initial allele frequencies (Table 12.6). Substituting in the formula given in Table 12.6 shows that when the A_2A_2 homozygote does not reproduce, the number of generations required to reduce the frequency of A_2 to one half its initial value is equal to the reciprocal of its initial value. Thus, if the frequency of A_2 is 0.01, it will take 100 generations of complete selection against the A_2A_2 homozygote to reduce the frequency of A_2 to 0.005. In other words, lack of reproduction of individuals with a rare recessive disease does not lead to a rapid reduction in the frequency of the deleterious allele from one generation to the next.

Now consider the situation in which there is partial selection against the A_2A_2 genotype. The allele frequencies after t generations cannot be written in terms of the initial frequencies, but the decrease in the frequency of the A_2 allele from one generation to the next can be calculated. This decrease is equal to $sq^2(1 - q)/(1 - sq^2)$, and the number of generations required to change the frequency of A_2 from its initial value to a new value can be approximated. For example, if $s = 0.001$ and the initial frequency of A_2 is 0.01, more than 100,000 generations will be required to reduce the frequency to 0.005. This example makes the point that even if the selective disadvantage of a genotype is very small, the allele frequencies in the population will gradually change. For the same selection coefficient ($s = 0.001$), 11,665 generations are required to reduce the frequency of A_2 from 0.7 to 0.1. If there is selection against the heterozygous genotype (A_1A_2) as well as the A_2A_2 genotype, with $s = 0.001$ for A_2A_2 and $s = 0.0005$ for A_1A_2 , then 6156 generations are required to reduce the frequency of A_2 from 0.7 to 0.1.

In the case in which there is selection favoring the heterozygote over both homozygotes, an equilibrium state is reached for the allele frequencies. Table 12.7 shows the change in allele frequencies from one generation to the next. At equilibrium, $s_1p = s_2q$, so that $p = s_2/(s_1 + s_2)$ and $q = s_1/(s_1 + s_2)$ (Table 12.7).

This equilibrium is stable and is called a balanced polymorphism. This type of selection is known as overdominance. If, on the other hand, selection is against the heterozygote, the equilibrium is unstable, and the selection is known as underdominance. The equilibrium frequencies are the same, but if a disturbance occurs such that $q > s_1/(s_1 + s_2)$, q will increase further rather than returning to its equilibrium value. The reverse is also true, so eventually one allele or the other will be eliminated.

TABLE 12.7 Selection Favoring the Heterozygous Genotype at an Autosomal Locus

Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency before selection	p^2	$2pq$	q^2
Relative fitness	$1 - s_1$	1	$1 - s_2$
Frequency after selection	$p^2(1 - s_1)$	$2pq$	$q^2(1 - s_2)$
After one generation of selection			
Frequency $A_1 = (p - s_1p^2)/(1 - s_1p^2 - s_2q^2)$			
Frequency $A_2 = (q - s_2q^2)/(1 - s_1p^2 - s_2q^2)$			
The change in the frequency of A_2 from one generation to the next is $pq(s_1p - s_2q)/(1 - s_1p^2 - s_2q^2)$. Equating this quantity to zero gives the equilibrium allele frequencies, which are			
Frequency $A_1 = s_2/(s_1 + s_2)$			
Frequency $A_2 = s_1/(s_1 + s_2)$			

Let us now consider a balance between mutation and selection. Suppose the mutation rate from A_1 to A_2 is μ , and the relative fitnesses of the genotypes A_1A_1 , A_1A_2 , A_2A_2 , are 1, 1, $1 - s$, respectively. As shown in Table 12.6, the frequency of A_1 after selection is $p/(1 - sq^2)$. Thus, the increase in frequency of A_2 due to mutation from A_1 to A_2 is $\mu p/(1 - sq^2)$, while the decrease due to selection is $sq^2(1 - q)/(1 - sq^2)$. At equilibrium, $\mu p/(1 - sq^2) = sq^2(1 - q)/(1 - sq^2)$, which simplifies to $q = \sqrt{(\mu/s)}$. This equilibrium is stable and $q = \sqrt{\mu}$ when $s = 1$. Thus, for a lethal recessive disease and a mutation rate of 10^{-6} , the equilibrium frequency of the deleterious allele is 1/1000.

In the case of a deleterious dominant phenotype, the fitness of both the homozygote and the heterozygote is reduced. With selection coefficients of $1 - s$, $1 - s$, 1, the increase in the frequency of A_1 due to mutation is equal to the decrease due to selection when $q = \mu/s$, which reduces to $q = \mu$ for $s = 1$. If individuals with a dominant disease do not reproduce, the frequency of the deleterious allele in the next generation is equal to the mutation rate. Examples of such disorders are atelosteogenesis and thanatophoric dysplasia, which are both lethal forms of short-limbed dwarfism. In the case of achondroplasia, fitness is not zero, but it is considerably lower than one, and is estimated to be about 0.2. Thus, the equilibrium frequency of the deleterious allele is $10^{-5}/0.8 = 1.25 \times 10^{-5}$, or slightly higher than the mutation rate.

TABLE 12.8 X-Linked Locus: Selection Against the A_2A_2 Genotype in Females and the A_2 Genotype in Males

Females			
Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency before selection	p^2	$2pq$	q^2
Relative fitness	1	1	$1 - s$
Frequency after selection	p^2	$2pq$	$q^2(1 - s)$
Males			
Genotype	A_1	A_2	
Frequency before selection	p	q	
Relative fitness	1	$1 - s$	
Frequency after selection	p	$q(1 - s)$	
After one generation of selection in males			
Frequency $A_1 = p/(1 - sq)$			
Frequency $A_2 = (q - sq)/(1 - sq)$			

Selection against genotypes at loci on the X chromosome needs to be tabulated separately for males and females because males have only one allele at an X-linked locus. Table 12.8 shows the case in which the A_2A_2 genotype and the A_2 genotype are selected against in females and males, respectively. The decrease in frequency of A_2 due to reduced fitness in females is extremely small compared with the decrease due to reduced fitness in males. Thus, we will only consider males (Table 12.8). The loss of A_2 alleles is equal to $sq(1 - q)/(1 - sq)$, which is q if $s = 1$. In other words, if selection is complete, all male A_2 alleles are lost in each generation. Because males have only one allele and females have two, this loss represents one third of the A_2 alleles in the population. If the mutation rate from A_1 to A_2 is μ , the increase in frequency of A_2 due to mutation in males is $\mu p/(1 - sq)$. But mutation in males represents only one third of the mutations that are occurring in the population. Thus, an increase in frequency due to mutation balances a decrease due to selection when $3\mu p/(1 - sq) = sq(1 - q)/(1 - sq)$, which reduces to $q = 3\mu/s$. For an X-linked recessive lethal, $s = 1$, and $\mu = q/3$. In other words, one third of the deleterious alleles in the population, and, thus, one third of cases of diseases, such as Duchenne muscular dystrophy, are new mutations. In less severe X-linked disorders, the proportion of cases that are new mutations is not as high; for example, the relative fitness of individuals with hemophilia A is about 70%. Therefore, the proportion of new mutations is $0.3q/3$, meaning

that about 10% of cases are new mutations. Of course, during the initial years of the AIDS epidemic when blood was not being screened for HIV, the relative fitness of hemophiliacs was considerably lower than 70%. The effect of this transient reduction on the frequency of hemophilia A will be seen in future generations.

Example

Suppose the relative fitnesses of the genotypes A_1A_1 , A_1A_2 , A_2A_2 are 0.8, 1, 0.1, respectively. What are the equilibrium frequencies of the two alleles?

The heterozygote has a selective advantage in this population. At equilibrium, the frequencies of A_1 and A_2 are $0.9/1.1 = 0.82$ and $0.2/1.1 = 0.18$, respectively. Even though selection against the A_2A_2 homozygote is quite extreme, the equilibrium frequency of A_2 is relatively high because of overdominance. If the relative fitnesses were 0.9, 1, and 0.7, the frequencies of A_1 and A_2 would be 0.75 and 0.25, respectively.

12.3.2.4 Migration (Gene Flow)

Migration introduces alleles into the population and, like mutation, increases heterozygosity. In general, migration rates are higher than mutation rates, so migration is more effective than mutation in counterbalancing the effects of genetic drift.

Comparison of alleles in different ethnic groups demonstrates the contribution of gene flow to the current population gene pools. For example, the most common mutations in phenylalanine hydroxylase that cause PKU are likely to be of Celtic origin. These same mutations have been found in many different populations throughout the world, reflecting the migrations of the Celts.

12.4 APPLICATIONS IN POPULATION GENETICS

The evolutionary forces that govern the frequencies of genotypes and alleles provide us with the tools to understand the genetic structure of populations. Such tools can help deduce why some disease mutations are relatively common in some ethnic groups, but not in others. Ethnic variation in allele frequencies is found throughout the genome, and by examining this genetic diversity, evolutionary patterns can be inferred, and variants contributing to the cause of both common and rare diseases can be identified.

12.4.1 Ethnic Diversity of Rare Disease Alleles

The existence of different disease alleles among ethnic groups is significant both for understanding the origins of the disease in a population and for estimating recurrence risks that will depend on ethnicity. Several examples show the benefit of applying population history to medical genetics.

The thalassemias have relatively high frequencies in many different populations, presumably due to selective advantage (increased relative fitness) of the heterozygotes over the homozygotes against malaria. Mutations in the genes encoding both the α -chain (chromosome 16) and the β -chain (chromosome 11) of hemoglobin cause thalassemia. Most of the β -thalassemia mutations are single base-pair substitutions, as opposed to the α -thalassemias, in which complete genes are deleted. More than 80 β -chain point mutations that cause β -thalassemia have been described. These mutations have a wide ethnic distribution, with several different common alleles found in Mediterranean, African, and Southeast Asian populations.

Tay-Sachs disease provides another excellent example of ethnic-specific mutations. The most common mutation in the hexosaminidase-A α -subunit gene (chromosome 15) causing Tay-Sachs disease in the Ashkenazi Jewish population is a 4-bp insertion in exon 11. It is found in 80% of mutant alleles in this population, but in less than 20% in other populations. Three alleles account for 99% of mutations in the Ashkenazi Jewish population. The frequency of Tay-Sachs alleles is also relatively high in the French Canadian population, in which two different mutant alleles have been described. Members of several Acadian families in southwestern Louisiana were found with Tay-Sachs disease; in 11 of 12 disease alleles, the mutation was the 4-bp insertion that is the most frequent mutant allele in the Ashkenazi Jewish population.

Other diseases, such as gyrate atrophy and familial hypercholesterolemia, show similar ethnic diversity in the distribution of mutant alleles. Founder effect (random genetic drift), selection, and gene flow determine the frequencies of mutations in different populations.

12.4.2 Evolutionary Patterns

Various types of DNA marker alleles have been analyzed in studies of population structure and evolution, and

before the introduction of DNA markers, similar studies were done using blood group and protein polymorphisms. The evolutionary tree derived from these studies suggested four major groupings, consisting of Africa, Europe/Asia, Americas, and Australia/New Guinea, with the most likely origin being the African branch. Detailed analyses of mitochondrial DNA (mtDNA) and Y-chromosome haplogroups have extended these findings and confirmed that contemporary populations are largely the descendants of people who migrated out of Africa about 50,000 years ago. For example, the mtDNA and Y haplogroups found in southeastern Asia and Australia are distinct from those in the rest of Asia and Europe. This variation is most likely the result of random genetic drift and northern and southern migrations at different times. The Americas were the last continents to be colonized, and as would be anticipated, most native American Y chromosomes belong to a single haplogroup. Interestingly, the results of voyages by Europeans to the Americas and Oceania in the past 500 years are clearly revealed through mtDNA and Y-chromosome analyses. In these populations, European Y-chromosome haplogroups are relatively common, while the mtDNA haplogroups are those of the indigenous population.

Because of sex-specific gene flow and the small amount of the genome represented, studies of historical migration based only on mtDNA and Y chromosome are biased [1]. With the availability of next-generation sequencing, entire human genomes can be compared. In fact, a draft sequence of the Neanderthal genome has now been completed [2], revealing gene flow from Neanderthals to modern humans after moving out of Africa. There is also an extensive amount of genetic variation among African populations [3], with most African-Americans having a mixed ancestry that is not specific to any one African group [4]. This high level of admixture can be quite helpful for identifying genetic variants associated with disease or response to drugs.

12.4.3 Genome Variation

The advances of the human genome project have renewed appreciation and interest in the study of naturally occurring variation in the human genome. About 90% of human DNA variation is due to single nucleotide base changes. On average, a single bp difference between two human genomes is observed for every 1000 bp. But the odds of finding a difference may be as much as 100-fold greater in some regions of the genome than in others. Single nucleotide polymorphisms (SNPs) are

defined as loci with alleles that differ at a single base, with the rarer allele having a frequency of at least 1% in a random set of individuals in a population. In general, the likelihood of finding an SNP is much higher in noncoding regions than in coding regions. (An SNP in a coding region is sometimes called a cSNP.) Most SNPs found in the human genome are thought to have originated long after speciation, but before the separation into different human populations. This explains the observation that human SNPs are usually not shared with primates, but are common to all populations; only about 15% are thought to be “private,” or found only in one population. Also, only a few of the SNP alleles that were present when humans moved out of Africa have become fixed (either 0% or 100%) at this point in time.

As a result of major international initiatives, approximately 10 million SNPs, both common and rare, have been identified in the human genome, and more than 1.5 million have been genotyped in over 1000 individuals from 11 global populations in the original HapMap study [5]. In addition, the detection and characterization of copy number variants (CNVs), which tend to map to duplicated segments, have provided access to important variation, particularly in highly duplicated gene families, which is likely to contribute to some common diseases [6]. And most recently, the 1000 Genomes Project has generated and cataloged many millions more SNPs and CNVs, bringing the total to nearly 88 million variants across 26 populations. The goal of this project is to discover, genotype, and provide accurate haplotype information on DNA polymorphisms in multiple human populations, and the belief is that 99% of variants >1% MAF have been discovered in these populations [7,8]. These extensive data sets provide the population geneticist with a huge set of densely mapped polymorphisms for reconciling genome variation with population histories of bottlenecks, admixture, and migration, and for revealing evidence of natural selection. They also enable informative studies such as pinpointing functional elements affecting gene expression in noncoding DNA, which tend to be in regions of reduced variation [9]. Moreover, the knowledge gained from these studies is applicable to many disciplines including forensics, pharmacogenomics, and complex disease research.

12.4.4 Identification of Causal Variants for Common Diseases

Earlier in this chapter, the concept of LD was described, and it was mentioned that studies exploiting this

phenomenon have been helpful in defining the precise location of disease genes. The complexity of patterns of LD and the extent and explanation of variability among populations are critical factors that are only now becoming understood [10].

A large number of genome-wide association studies have been performed using the millions of SNPs that are now available. These studies have identified genomic regions containing genetic risk factors for many complex diseases, for example, Parkinson disease [11]. However, for the majority of such diseases that have been studied so far, much of the genetic influence on them remains to be explained [12].

Current work in the field of population genetics involves studying the differences in disease-causing variants between major population groups. As many GWAS studies have only been conducted in European populations, there are open questions as to the utility of results from these studies in other populations. The more we learn, the more it appears that studying diseases in multiple populations is going to be critical to discerning the true nature of causal variants for each ancestral group [13,14].

The availability of massive databases of genetic variation and automated technology for genotyping, sequencing, and bioinformatic analysis [15,16] is significantly enhancing collaborative efforts between population geneticists and molecular geneticists and advancing understanding of many diseases. As we would anticipate, thousands of interesting new questions are being raised and the tools to answer many of them are now at our fingertips.

Improved sequencing has led to significantly lower costs per genome as well as improved sequencing quality, allowing researchers to choose whole-genome sequencing as a study design more often. There is now a massively ballooning set of fully sequenced genomes for study. The proliferation of these modern, large-scale genome-wide sequencing studies is generating variant information on an incredible scale, enabling population geneticists to ask and answer questions that have previously only been theoretical in nature.

REFERENCES

- [1] Cox MP, Hammer MF. A question of scale: human migrations writ large and small. *BMC Biol* 2010;8(98).
- [2] Green RE, Krause J, Briggs AW, et al. A draft sequence of the Neandertal genome. *Science* 2010;328:710–22.
- [3] Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. *Science* 2009;324:1035–44.
- [4] Zakharia F, Basu A, Absher D, et al. Characterizing the admixed African ancestry of African Americans. *Genome Biol* 2009;10:R141.
- [5] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52–8.
- [6] Sudman PH, Kitzman JO, Antonacci F, et al. Diversity of human copy number variation and multicopy genes. *Science* 2010;330:641–6.
- [7] The 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
- [8] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;526:68–74.
- [9] Lomelin D, Jorgenson E, Risch N. Human genetic variation recognizes functional elements in noncoding sequence. *Genome* 2009;20:311–9.
- [10] Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;322:881–8.
- [11] The UK Parkinson's Disease Consortium, The Wellcome Trust Case Control Consortium 2. Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum Mol Genet* 2011;20:345–53.
- [12] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53.
- [13] Mathias RA, Taub MA, Gignoux CR, et al. A continuum or admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat Commun* 2016;7:12522.
- [14] Martin AR, Gignoux CR, Walters RK, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet* 2017;03:635–49.
- [15] Johnston HR, Chopra P, Wingo TS, et al. PEMapper and PECOler provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci Unit States Am* 2017;114(10):E1923–32.
- [16] McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit; a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.

Relevant Websites

International HapMap project. <http://hapmap.ncbi.nlm.nih.gov/>.
 A catalog of published genome-wide association studies. <http://www.ebi.ac.uk/gwas/>.

FURTHER READING

Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* May 27, 2009.

Maniolo T, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590–605.

Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatr* 2009;166:540–56.

Pathogenetics of Disease

Reed E. Pyeritz

Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, United States

To wrest from nature the secrets which have perplexed philosophers in all ages, to track to their sources the causes of disease, to correlate the vast stores of knowledge, that they may be quickly available for the prevention and cure of disease—these are our ambitions.

Osler W. Chauvinism in Medicine. *Montreal Med J* 1902;31:684–99.

The great ease with which molecular information can be collected on the genomes of higher organisms will tempt many. We can inevitably expect vast compendia of sequences but, without functional reference, these compendia will be uninterpretable, like an undeciphered ancient language. Many people and many computers will play games with these sequences, but we will have to find out by experiment what the sequences do and how the products they make participate in the physiology and development of the organism. Thus, although the analysis of the genotype has been taken care of, we still need better ways of analyzing phenotypes. Many of us are ultimately interested in the causal analysis of development and the reduction of the complex phenotypes of higher organisms to the level of gene products. This is still the major problem of biology. We must understand what cells can do because all of what we are is generated by cells growing, moving and differentiating.

Brenner S. Human Genetics: Possibilities and Relatives. *Ciba Found Symp Excerpta Medica* 1973;66:1–3.

The effort spent on the identification of genes is likely to prove only a small fraction of that required to work out their normal function in the tissues in which they are expressed. Yet that is where clues to the treatment and prophylaxis of disease are most likely to arise.

Maddox J. Genes and Patent Law. *Nature* 1994;37:1270.

13.1 INTRODUCTION

The foregoing quotations emphasize that the general theme of this chapter—all that occurs between the gene and the bedside—is not a new one. The true promise of the Human Genome Project only began to be realized when our genome was sequenced, and the really hard work persists [1,2].

At the outset, the definitions of four terms are fundamental to everything that follows. All deal with the causation of phenotypes, and they are distinguished as to method and scope of enquiry.

Etiology is the study of the causes of a phenomenon and, in the medical context, of disease. Its method is to discover the association between factors that are thought to be causes and certain features that we wish to explain. The goal and method of the discipline are strictly empirical, with at best minor interest in discerning the actual mechanisms involved.

Genetic etiology is a more specialized topic that deals with the properties of the genetic causal factors of disease and how they behave. Mendel's laws, which were

formulated before anything was known of the genes and their mechanisms, are arguably the high point of genetic etiology. A positive family history for early coronary artery disease is widely recognized as an important risk factor in the cause of myocardial infarction, with no appeal to explicit mechanisms, even genetic ones. Yet there are undoubted genetic versions, such as the Werner syndrome (OMIM*277700), an autosomal recessive disorder with premature aging and accelerated atherosclerosis.

Pathogenesis is the study of the mechanisms by which the etiologic factors are converted into disease states. For instance, the etiologic role of cholesterol in atherosclerosis has been ascribed to the infiltration of oxidized low-density lipoprotein into the arterial wall (the insudation theory); to the stimulation of organization and repair of small arterial thrombi (the encrustation theory); to the promotion of cellular and humoral immune processes within the arterial wall; and to the secondary accumulation of cholesterol in areas of minor initial damage. Study of these rival, or perhaps complementary, processes falls within the domain of pathogenesis.

Pathogenetics, a condensation of “genetic pathogenesis,” is the study of how anomalies in the genome are converted into the phenotypes of disorders. Numerous factors play roles, including epigenetics, the microbiome, mosaicism, chimerism for maternal and cells of other origins, and variations of nuclear and mitochondrial genes. The first appearance of the term in the literature was in 1949 in reference to the action of polio virus [3]. Subsequently, in 1951, the notion of pathogenetics was used to speculate about the treatment of peptic ulcer [4].

In this chapter, the emphasis is on disease, while stressing the relationship of genetic processes to ordinary developmental mechanisms and maintenance of the healthy body, known as *orthogenetics*. One such concern is preserving an intact vascular system. Minor leaks in the vascular tree are mended by the hemostatic plug, comprising platelets and fibrin. At times, this mechanism may become exuberant, resulting in thrombosis, which may occlude a strategic blood vessel. In turn, the thrombus may be covered by endothelium, with its remains, sealed in the arterial wall, perhaps converted into an atheromatous lesion. This gradation points out the perils of drawing sharp boundaries between the so-called normal and abnormal states, boundaries that are neither necessary nor illuminating. For this reason, orthogenetics and pathogenetics are best seen as parts of a single continuous field of enquiry [5]. Understanding

the phenotype comes from the interactions among the insights from many disciplines.

Galton–Fisher theory refers to the combination of quantitative genetic hypotheses originated by Sir Francis Galton and Sir Ronald Fisher early in the 20th century [6]. Their precepts were to a certain extent a response to the inability of Mendelian theory to explain the inheritance of common traits and diseases, such as intelligence and hypertension.

13.2 THE SCOPE OF ABNORMAL PHENOTYPES: SUSCEPTIBILITIES, DISEASES, AND MALFORMATIONS

It is a truism among pathologists that disease occurs only in the living. On the one hand, a cadaver cannot undergo poisoning or neoplasia. On the other hand, instantaneous death by catastrophe of a healthy person may occur without any disease even being started. Disease comprises both a disruption and a reaction to it: for example, homeostasis; inflammation and its sequelae; and curtailment of activities, whether voluntary or by invalidism.

13.2.1 Homeostasis

Homeostasis, a mainstay of physiology, is a concept attributed to Bernard [7]; studied and named by Cannon [8]; and cast in formal terms by Wiener [9]. It involves cybernetic devices that maintain the inner environment of the organism in a state that favors normal functioning. Despite its cardinal role in both normal and disturbed physiology, homeostasis has often been neglected in medical genetics. Its characteristic pattern is to offset departures of a measurable characteristic, such as body temperature, glucose concentration, or blood pressure, from a steady state that is in some sense optimal [10].

Much of the detailed workings of homeostatic systems depend on enzymes, receptors, and ligands, all of which exhibit genetic variation. Perturbations are a fact of life, and physiologic homeostasis to correct them is ubiquitous. No reasonable system can correct them fully and instantaneously. Furthermore, as we have seen in the control of periodic functions, other competing benefits of homeostasis are to be considered. Wherever there is an inescapable lag, there will be added difficulties in achieving a prompt response without undesirable overshoot and perhaps even total loss of control.

For many bodily characteristics, health is the avoidance of extremes. Diseases may then result from excess

or deficiency, and the dynamic reactive component will aim to restore the optimal. In Galton–Fisher theory, any variable genetic component becomes fixed in magnitude at conception. By contrast, the responses of health and disease are perpetually exercised in a fashion variable in both degree and direction, in accordance with the size and sign of each perturbation. Galton–Fisher theory, so useful in static traits, is irrelevant to most diseases. From time to time, Galton–Fisher theory has furnished more refined descriptors such as Pearson’s threshold model in congenital heart disease. But it generally does not lead to deeper questions and understanding of disease, its cause, its pathogenesis, or its genotypic fate at large (genetic population dynamics).

For example, many details suggest that most cases of type II diabetes mellitus at their outset are due to neither defective nor deficient insulin, but to impaired insulin regulation or sensitivity. Arguably, hyperglycemia and hypoglycemia are the same condition in different phases, as illustrated by potentially rapid oscillations between the two states in “brittle” diabetes. The interpretation of two apparently opposite disorders as a manifestation of instability of a single trait is not so startling as it may first appear. Besides the obvious precedent of bipolar mood disorder, there are many analogs, such as postural hypotension, dysautonomia, and anorexia–bulimia [11], but all such conditions would be inaccessible to Galton–Fisher theory, centered as it is on the first two moments of the Gaussian distribution as a gauge of the variation of means. Galton–Fisher theory is not sensitive to variation in tolerances as a segregating trait. Purely technical use of Galton–Fisher theory in almost purely genetic bipolar disease may yield heritability at or close to zero and lead to the mistaken conclusion that the disease is nongenetic. Diabetes occurring early in life is usually type I and insulin-dependent, and its etiology has long been recognized to have both environmental and multiple genetic components.

13.3 DEVELOPMENT OF ANATOMIC STRUCTURES: ANGULAR HOMEOSTASIS

The pathogenetic challenge for large and complex structures is translating a linear (single-dimensional) code in the genome into the fetus. The objective involves three dimensions in space and one in time. That the whole must be carefully orchestrated is self-evident. But perfect assembly of a spatial structure is rarely enough: it

must commonly be adequately matured by a specified ontogenic timeline. Development of the lips and palate calls for exquisite timing. Even in genetic etiology, the ontogenic phenomenon of “time windows” indicates some state of minimal precision. This complex feat of organizing form and timing may perhaps be achieved by “dead reckoning,” by rigorous specifications and a relentless timetable. However, there is evidence that the system is more robust than that, because of what Waddington [12] called homeorhesis. This process in ontogeny, which we call angular homeostasis, is akin to Bernard–Cannon homeostasis [13,14]. In it, discrepancies between the current and the ideal states of ontogeny are discerned and corrected. There is plenty of biologic evidence from age-old observations of tropisms and taxis and from modern studies of target tracking and predation to make the whole strategy highly plausible. What exactly the cybernetic details are and how they operate are emerging in expressly and explicitly quantitative chemical terms [15–17]. But the greater efficiency that continuous sensing and correction enjoys over unbending protocol is evident to anyone who has experience of the minutely directed process of firing a gun at a moving target, where the path of the bullet, once fired, is beyond correction.

If usual development is difficult to reconcile with a linear code of instruction, so too is maldevelopment. In this era of ever more refined definition of the human genome, a variety of approaches will be required to address both the normal and the abnormal.

For example, phenotypes associated with additions or deletions of many genes, such as in aneuploidy, or deletions or duplications of contiguous genes (“genomic disorders”) have stimulated considerable debate about pathogenesis. On the one hand are those who hold to the reductionist position that the phenotype reflects, in essence, mass action of too much or too little of the products of the affected genes. One result is the attempt to identify, for example, the gene or genes on chromosome 21 responsible for the cardiovascular malformations in Down syndrome. On the other hand are those who see the phenotypes as the result of complex interplays among multiple genes, which are being expressed in a local environment that is asynchronous with normal development.

Another example is coding the laterality of the brain, a more perplexing problem than ever. For the choice now does not seem to be an anatomic one at all, but a

question of which information and which kinds of mental process shall be assigned to the left side and which to the right. Wholly indifferent, random assignment is theoretically possible [18]. However, although the facts have been distorted in the past by social prejudice, there still seems to be a higher rate of “left brain dominance,” and although the patterns are not altogether clear, there is evidence that genetic factors are at work. The genetics of handedness, as a surrogate for cerebral dominance, has been studied for decades. According to one hypothesis, mothers tend to support infants using their left arm, perhaps so to sooth them with the sound of the maternal heart. Thus, mothers who were by nature right-handed (dextral) would find some evolutionary favor. Another hypothesis suggests that left-handed warriors were more successful. What is clear, however, is that left-handedness is less common in all human populations, despite some geographic variation. Thus, whatever selective pressures favor left-handedness must be balanced by some negative ones for the trait to remain less common. The prevalence of left-handedness decreases with age [19]. Whether sinistrality itself reduces life span, or serves as a marker for underlying neuropathology (perhaps related to cerebral dominance), remains unclear [20].

13.3.1 Elaborateness of Repair

Ontogenic robustness and recuperative power are reciprocally related. The brain is the most highly organized structure but, while having some functional capacity to recuperate from damage by the fluidity in allocating space to functions, only recently has it been shown to have any regenerative power through stem cells. The structure of the kidney may be less critical than that of the heart but, like the heart, it has little capacity to repair damage to its architecture. Anatomically at least, the liver is both less critically structured and more robust, but has much greater forces of recuperation. Tissues such as skin, bone marrow, spermatogonia, and intestinal endothelium are still less elaborate and are therefore so ready for regeneration as to be notorious sites of sensitivity to mutagens.

13.3.2 Life History

A natural feature of the impact of a disorder is how it affects well-being, fertility, and length of life. These three, although distinct, are obviously connected; yet the traditional methods of analysis pay little attention to

this fact. Discussions on fitness make much of the fact that clinical, genetic, evolutionary, athletic, and moral fitness are so different that they must be discussed separately. Indeed, excellence in one may go with mediocrity, even total incompetence in another. The super athlete may be sterile and morally bankrupt. The puny may live a long life free of disability. The fertile may be negligent in the care of their progeny. But it is absurd to suppose the several types of fitness to be totally unconnected. The issues are too large to deal with here, and we consider only the relationship between clinical fitness and length of life.

It is useful to distinguish between the unfolding of a disease and the impact of its complications. The wholly static disorders are mostly trivial: for example, red-green color blindness, tone deafness, pentosuria, synophrys, and the like. On the other hand, the usual patterns of deterioration may vary greatly. In severe cardiac malformations, the disability is evident at birth or soon after. Even so, major complications such as pulmonary hypertension and reversals of flow through a patent ductus arteriosus or an atrial septal defect may occur late. In hereditary polyposis coli, it is the course of the disease itself—the long latent phase, progressive polyposis, and malignancy—that is the chief concern, while other complications (except those due to therapy) are minor. By contrast, Alzheimer disease (OMIM*104300) often appears late enough that it is frequently obscured by intercurrent and competing diseases. Indeed, for centuries “senile dementia” was viewed by some as a concomitant of aging, and not as a disease. But while the duration of Alzheimer disease is typically less than a decade depending on diagnostic finesse and management, it is surprisingly difficult to find evidence that the patients ever have neurologic deaths. They seem to die of the complications (e.g., trauma, intercurrent infections, malnutrition) that occur at much increased risk [21].

The notion that well-being is eroded by overt catastrophes (e.g., strokes) or those imperceptible insults (e.g., chronic pyelonephritis) that we identify as “wear and tear” is so appealing that we are led, almost unconsciously, to accept multistage models. Where the genetic disorder becomes manifest early (as in Duchenne muscular dystrophy), the competing risks are small, and the pattern of deterioration is dominated by a single class of insults, with the typical survivorship curve positively skewed (i.e., with a long tail to the right), to an important extent because of interindividual genetic heterogeneity.

When the onset of disease is late, the patient shows the characteristic multiplex pathology so familiar to the geriatrician: damage in several body systems, so that it is often difficult to say which is the final cause of death. The survivorship is then often negatively skewed. That class of survivorship in which death is due to whichever of several partially damaged systems fails first is a “bingo” model, which is common but particularly difficult to examine either practically or quantitatively.

13.4 PATHOGENETICS OF REFINED TRAITS

A crucial process in molecular biology is generating the primary sequence of the polypeptide in the proper cell at the right time. This is attained by the elaborate apparatus of genetic coding, transcription and its regulation, and translation, which is highly conserved in evolutionary time, but the high organization largely ends at that stage. Once formed, the polypeptide assumes its secondary and higher structure by processes that are little understood; aside from posttranslational modifications catalyzed by enzymes, there seems to be little need to direct these processes. The polypeptide quickly assumes a stable low-energy state. Whether it ever becomes completely fixed is not readily established. But in or near that state, it functions most efficiently. The subsequent fate of the polypeptide may be largely random. For example, the theory of red cell survival suggests that the cell is eventually destroyed by random wear and tear, and the hemoglobin with it. However, survival of the whole is still shortened by some mutant forms of the primary structure of hemoglobin or of components of the erythrocyte wall.

The speed at which the polypeptide is made is certainly important. For instance, sickle hemoglobin is manufactured more slowly than the wild-type, such as to lead to a representation in the heterozygote in a ratio of 2:1 to 3:1. Furthermore, in heterozygotes, A and S hemoglobins tend to be concentrated in particular cells. However, one does not ordinarily regard translation (as opposed to transcription) primarily as a timed or quantitative process. Posttranslational modification is also sensitive to time. A mutation that results in substitution of a glycine in the triple-helical domain of type I procollagen results in slower winding of the helix. This in turn exposes for a longer time critical amino acids to the enzymes that catalyze modifications, such as glycosylation. The net result is a much more “damaged” molecule

than a simple amino acid substitution might predict. This, in turn, is reflected in the degrees of severity of osteogenesis imperfecta.

But there is even a higher-order effect possible when a protein is misfolded or otherwise damaged as it traverses the cellular machinery. When the endoplasmic reticulum encounters a misfolded protein, processing slows; if severe, a situation of “ER stress” ensues, which can lead to marked cellular dysfunction, even cell death [22–24]. Interestingly, the cellular phenotype may be the same for different mutations that affect entirely separate proteins. Understanding the importance of ER stress to the overall phenotype may afford a generic approach to therapy, whereby refolding of the mutant protein is facilitated.

Where the components are interchangeable (e.g., β^A - and β^S -globins), systems are appropriately described by their corporate properties. Where the numbers are large (e.g., numbers of erythrocytes), the usual device is the probabilistic model; and where the numbers are even larger (e.g., molecules), deterministic methods greatly simplify the analysis with negligible loss of accuracy. However, whatever the value of deterministic models in microbial populations, they have little place in studies of human beings; even in molecular studies, they must be handled with circumspection. This is a major difference between classic population genetics and the highly individualized character of medical genetics.

13.5 PATHWAYS AND MULTIPLE-STAGE PROCESSES

Two highly refined approaches have much in common and may, in certain circumstances, be united by a single theory.

13.5.1 Simple Pathways

The simplest possible process involves synthesis of B from A by enzyme *ab* (Fig. 13.1). There are three potential deleterious consequences:

- Precursor toxicity: Because *ab* fails, A accumulates and proves harmful. Alkaptonuria (OMIM*203500)

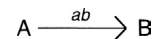


Figure 13.1 An enzyme, *ab*, catalyzes conversion of substrate A to product B.

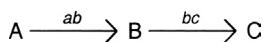


Figure 13.2 An enzyme, *ab*, catalyzes conversion of substrate A to intermediate B, which is converted to final product C by enzyme *bc*. A defect in *ab* impairs production of C. The gene specifying *ab* is epistatic to that encoding *bc*.

is such a disorder, as are most enzymopathies in catabolic pathways, such as lysosomal storage disorders.

- **Product deficit:** Because *ab* fails, B is reduced or absent. Examples include the various forms of albinism due to failure to produce pigment (e.g., OMIM*203100) and most enzymopathies involving posttranslational processing of proteins. In a few mammalian species, the inability due to deficiency of one enzyme, L-gulonolactone oxidase, to synthesize ascorbic acid is another example. Does deficiency of this enzyme in all humans exclude it from the category of “disease”? At a minimum, deficiency of this enzyme creates a risk factor for scurvy in all of us.
- **Combined product deficit and precursor excess:** The glycogen storage disorders are examples (e.g., OMIM*232200). The glycogen that accumulates disrupts cellular and tissue processes, while failure to release glucose from glycogen leads to hypoglycemia. Phenylketonuria (OMIM*261600) is another such example; phenylalanine is toxic in excess, and synthesis of tyrosine is impaired, resulting in the pleiotropic manifestations of phenylalanine hydroxylase deficiency.

This elemental pattern extends to the three-step process: $A \rightarrow B \rightarrow C$ (Fig. 13.2). If A is absent, then B is lacking, and C cannot be synthesized. This suppression is epistasis; the gene governing the first step is epistatic to that governing the second. The classic example in humans is the rather trivial Bombay blood group phenomenon (OMIM*211100) in which the failure to generate H substance destroys all means of expressing the ABO blood group phenotype.

Consider a typical multistage metabolic pathway, such as the synthesis of cholesterol or thyroid hormone. Each step is under the control of an enzyme. It will be at once evident that total failure at one or more steps means total blockade of later stages and that substrate accumulates before the first failed step. The gene for the enzyme at any step is therefore epistatic to all subsequent steps. The combined effect of defects in all genes will be the same as that of any subset of defective genes.

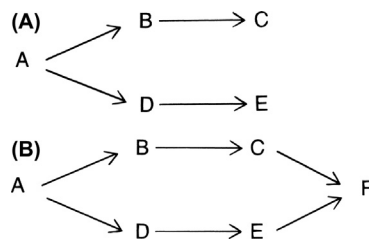


Figure 13.3 Metabolic pathways with branches. (A) Open branched pathway. (B) Closed branched pathway.

In this it is quite different from the usual additivity of traits in Galton–Fisher theory.

13.5.2 Branching Pathways

Two kinds of branching pathways, the open and the closed, can be distinguished. In the open type (Fig. 13.3A), the branches do not rejoin and pool their products; thus, they compete for substrate, and the flow through each is correspondingly decremented. In the closed type (Fig. 13.3B), the paths rejoin, and the result is a parastasis; two or more pathways run in parallel, which accelerates the entire process and acts as a fail-safe device should any of them fail. This scheme can be used as an advantage in treatment, such as by promoting remethylation of homocysteine to methionine through an alternative pathway dependent on the cofactor betaine. Those classic inborn errors of metabolism that lack adequate alternate pathways are the most severe clinically. Their rarity argues that metabolic processes without auxiliary paths are the exception, and the selective disadvantages may explain why.

On the other hand, a defect in one branch of a pathway may generate all or most of its pathology by leading to overflow through the alternative branch. For example, a defect in the enzyme hypoxanthine-guanine phosphoribosyltransferase (OMIM*308000) leads to overproduction of phosphoribosylpyrophosphate. This in turn drives overproduction of purines, which leads to hyperuricemia, hyperuricosuria, and gout.

13.5.3 Pathways with Feedback

Metabolic pathways may be actively regulated in some cases by demands downstream. Negative feedback, positive feedback, or both can achieve a desired rate of processing or level of synthesis. This represents a form of physiologic homeostasis. Production of most hormones

involves feedback at multiple levels. For example, estrogen is secreted by ovarian follicular cells in response to the anterior pituitary hormone and follicle-stimulating hormone (FSH); estrogen in turn feeds back on both the hypothalamus, to inhibit production of gonadotropin-releasing hormone, and on the anterior pituitary, to inhibit release of FSH, thereby modulating estrogen production and preparing the endometrium for implantation. Once an embryo implants in the endometrium, synthesis of chorionic gonadotropin signals the ovary to continue production of progesterone, which maintains the endometrium as a nurturing environment for continuing the pregnancy.

13.5.4 Multiple-Hit Processes

A metabolic process involving several steps may, where necessary, be viewed in more quantitative terms. In the synthesis of insulin in response to a carbohydrate load, multiple steps are involved, but physiologic impact occurs only in the final step that results in a physiologically active form. The lag time for the response, then, is that for the transit through the entire system, and the characteristics of inert precursors make little difference. In this sense, what matters is only the process as a whole; permutations of the components do not matter.

13.5.5 Multiple-Compartment Models

To deal with chemical processes in the foregoing fashion has certain uses, notably, in understanding steady-state processes. A more quantitative approach is necessary where changes need to be more rapid. Many of the processes of converting one class of compound into another are of so-called zero-order kinetics, that is, other things being equal, the rate of transfer equals the concentration of the substrate multiplied by the Michaelis constant, m , of the enzyme. In a system without replenishment of substrate, these conditions define the negative exponential process, and the mean time for conversion, the waiting time, is precisely $1/m$. This pattern may be viewed from two perspectives: first, as a chemist sees it, deterministically conforming to the law of mass action; and second, as a probabilist sees it, a random exponential (one-hit) process in which every eligible molecule is at a fixed instantaneous hazard of change. As long as the number of molecules remains large, the distinction hardly matters. But the probabilistic model is more appealing because of its wider relevance in biology (e.g., when the number of decaying items—molecules, body cells, recurrent bodily

insults—is small, and random uncertainty may no longer be safely ignored). Assuming that each step operates independently, the transit time through the metabolic chain is then the sum of the waiting times of each step. The whole is termed a multiple-hit process. Its mean value is the sum of the waiting times, that is, the sum of the reciprocals of the Michaelis constants.

The smaller any particular Michaelis constant compared to constants for steps elsewhere in the chain, the larger is its reciprocal and the greater its impact on the whole transit time. Moreover, other things being equal, for any given variation in m , the smaller the mean, the larger the variation in its reciprocal, and the more sensitively the impact of variations in it will be detected. At some, quite arbitrarily small, value of m and large value for its reciprocal, this dominating step is termed the rate-limiting step; viewed genetically, if this step is itself Mendelian, the whole will be termed Mendelian. It will be evident that the conventional distinction between Mendelian and Galtonian (“multifactorial”) traits is both vague and arbitrary.

13.6 MOLECULAR PATHOGENETICS

Describing the defect in a genetic disorder at the level of the mutation specifies its etiology. Anything more remote from the mutation represents phenotype and at least the first layer of complexity in the pathogenetics. At the most remote layer is the ultimate phenotype. Any intermediate phenotype reflects the pathogenetics. Thus, the resolution or sensitivity of the methods being brought to bear on the investigation of how the disorder arises determines how closely the mutation can be approached. The clinician has long had to deal with crude tools—stethoscope, tape measure, electrocardiogram, radiograph, urinalysis—to define the phenotype; what generally results is a perception of pathogenesis that is shallow, often complex, even confused. At the bedside, and even in the clinical laboratory, one usually sees only the leaves on the pathogenetic tree. Advances in clinical chemistry, biochemical genetics, cytogenetics, immunology, noninvasive and invasive imaging of many types, and pathology have all led to a more radical, hence sensitive, discernment of what is wrong with the patient and elucidation of pathogenesis. All these advances have facilitated in brushing aside the leaves and clambering part way down branches toward the trunk of the pathogenetic tree (Fig. 13.4).

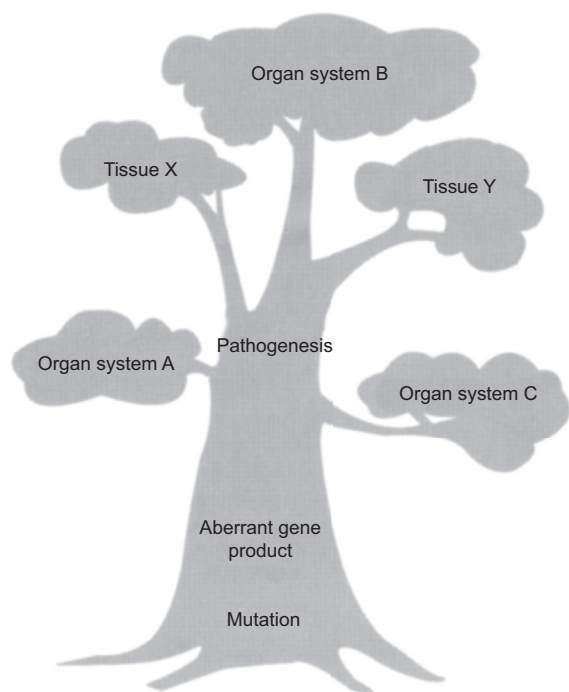


Figure 13.4 Pathogenetic tree for a Mendelian condition. Leaves correspond to the phenotypic features, detectable by bedside investigation. Branches represent the pathogenetic pathways leading to organ- and tissue-specific pathology. The trunk corresponds to the gene product. Roots indicate the cause, in this case, the mutation.

Fig. 13.4 also illustrates a fundamental characteristic of many human phenotypes termed *pleiotropy*. This word encompasses several concepts in biology; here it refers to multiple, even seemingly unrelated, aspects of the same syndrome. Indeed, *syndrome* embodies this notion of several clinical properties “running together.” Each of the leaves on the pathogenetic tree represents an aspect of the phenotype, connected through the limbs of pathogenesis. The analogy breaks down in that, while all leaves appear similar, the clinical details of the phenotype may be quite diverse. For example, dislocated lens, elongated digits, dural ectasia, and aortic root aneurysm are cardinal manifestations of Marfan syndrome (OMIM*154700), but outwardly bear no connection to each other. These features are all rooted in mutations in the gene (*FBN1*) encoding a large structural protein, fibrillin-1, and at the first level of pathogenesis, share defects in an extracellular structure, the microfibril

(corresponding to the trunk in the figure). However, microfibrils have distinct functions, including regulating TGF- β signaling pathways and providing structural integrity to certain structures, so the trunk branches in two. Further, since these two distinct functions vary in different organs and tissues, each of the limbs of the tree heads in its own direction. The molecular bases of pleiotropy are as diverse as the features of some syndromes [25]. For example, the complexity of some Mendelian inborn errors of metabolism is due to the mutant enzyme having functional roles beyond that in the specific metabolic pathway at issue. The extent of such “moonlighting enzymes” in human biology is uncertain [26].

The phenotype can be explored systematically, beginning with the first product of the mutant gene, mRNA. Various defects in the structure and amount of a given mRNA can be described, albeit only with considerable effort and sophistication in techniques. The thalassemias constitute one group of diseases that beautifully illustrate the molecular pathology of mRNA.

It is more feasible and instructive to focus on the stable product of most genes, the protein, and describe the types of molecular pathology that arise from variation. In the most fundamental terms, a mutation can affect the quantity of a protein, the quality of a protein, and occasionally both aspects. The quantity of a protein synthesized by a gene is regulated at the level of transcription, by promoters, enhancers, and other locus control elements, and at the level of translation. DNA variants in any of a number of sites, *cis* and *trans*, to the gene of interest can affect the amount of protein produced. Usually, but not always, production from the variant alleles associated with a disease (often termed mutations) is decreased. One class of variation that has garnered considerable attention is the expansion of a trinucleotide repeat within or, more commonly, outside the actual coding sequence. The number of repeats may be inversely proportional to transcription of the gene. Furthermore, the more repeats, often the more severe the phenotypic change.

A change in the primary structure, the amino acid sequence, of a protein may alter its function (i.e., the quality of the protein). The study of diverse variants of the same protein has greatly advanced the understanding of molecular pathogenetics and investigating authentic relationships between genotype and phenotype. This inquiry calls for a new commitment to meaning and the authenticity of the descriptors, matters that are much more sophisticated than the correlations and coefficients that

have dominated classic Galton–Fisher theory. The latter may lead to such paradoxical results as that, a nearly perfect correspondence between genotype and phenotype, may nevertheless yield a zero correlation and hence zero heritability. How the quality of a protein can be affected depends, in the first instance, on its normal function.

A DNA variant can change both the quality and quantity of a protein. For example, a change in primary sequence might affect the stability of the protein and lead to enhanced (or retarded) degradation. In some situations, the amount of the altered protein is crucial to the severity of the phenotype, especially in a dominant-negative scenario.

Proteins can be divided into three classes based on function: those whose essential functions involve interactions with small molecules, such as enzymes, receptors, and transporters; those that perform regulatory roles, such as transcription factors and hormones; and those that function in complex systems, often in a structural role, and often in association with other proteins.

Most proteins have one or more domains associated with specific functions. Not surprisingly, proteins in the same class often have domains in common, and there is remarkable conservation of sequences among domains. For example, transcription factors all have one or more amino acid sequence motifs (such as leucine zipper or zinc finger) that facilitate binding of the protein to DNA sequences. Cellular receptor molecules have domains that enable interaction with the lipid bilayer of the cell membrane, an extracellular domain that binds a ligand, and often a domain that resides in the cytoplasm, perhaps exhibiting kinase activity. Some molecules have many domains, some of which are composed of dozens of repeated motifs—witness fibrillin (OMIM*134797), lipoprotein(a) (OMIM*152200), and plasminogen (OMIM*173350). The conservation of domains has facilitated discovery of the cause of numerous diseases through positional cloning. Thus, when a newly identified open reading frame is sequenced and found to be a strong candidate for the cause of a disorder, generally through identification of a variant, the logical question is what the function of the protein encoded by the gene might be. If the coding sequence specifies an amino acid motif typical of a zinc finger, the protein is likely to be a transcription factor. This process is aided considerably by large databases that incorporate knowledge of genetic sequences and protein structure and function from both humans and all other organisms.

Proteins that interact with themselves (to form multimers) or with other proteins are subject to enhancement of a pathologic effect when one copy of their gene is mutant. Even though the patient is heterozygous for the mutation, the defective protein, by interacting with the product of its normal allele, or the products of other nonmutant genes, consumes these normal proteins; the result is a much more severe phenotype than would be expected from having half-normal levels of the normal protein. This is termed the dominant-negative effect, and is rather common. The irony is that a “more severe” variant, such as one that eliminates transcription from the variant allele altogether (i.e., a null allele), has less effect on phenotype than does a missense variant that leads to normal transcription and translation of a mutant protein.

Within each of the three classes of proteins, a mutation can have one of four consequences: quantitative increase or decrease in function and qualitative gain or loss of function. Each of these consequences can have a number of molecular explanations.

A quantitative increase in function can be due to a regulatory mutation. An example is loss of sensitivity to inhibition, such as by a repressor molecule. A variant could also affect the active site of an enzyme, such that its V_{\max} was increased, or the binding site of a hormone, such that the K_M was lowered.

A quantitative decrease in function could operate by the converse of any of these mechanisms. The extreme of the spectrum of decreased function is loss of function, perhaps the easiest to conceptualize, and certainly the most prevalent consequence. For example, most inborn errors of metabolic pathways result from an enzymatic failure. The enzymopathy can be due to a variant in or around the locus encoding that enzyme, resulting in a qualitative or quantitative defect as described earlier; to abnormal posttranslational processing of the nascent enzyme; to abnormal subcellular localization or extracellular trafficking; to altered affinities for substrates or cofactors; or to altered responsiveness to allosteric regulators of activity. Other examples of loss-of-function phenotypes include familial hypercholesterolemia due to many of the defects in the low-density lipoprotein receptor, and neoplasia due to defects in tumor suppressor genes such as the retinoblastoma or neurofibromatosis type 1 genes. Strains of mice bearing gene “knockouts” represent specified loss-of-function mutations; these are especially popular tools for studying development and neoplasia.

Quantitative and qualitative loss of function clearly overlap. A variant that reduces the ability of an enzyme to bind substrate also might lead to enhanced degradation and a reduced steady-state amount of the protein molecule.

Variants that cause a gain in function, that is, a function not intrinsic to the wild-type protein, are less common. The diverse familial amyloidoses are examples, in which a change in amino acid sequence of one or another protein (e.g., transthyretin) results in enhanced stability of the protein and abnormal tissue deposition (OMIM*176300).

The least commonly recognized molecular phenotype, also qualitative, is a change in function. One example is the product of the fusion of *BCR* and *ABL* in chronic myelogenous leukemia. Another example is the *p53* protein, which when mutated in some ways, assumes regulatory capabilities foreign to the normal product.

As useful as these protein phenotypes are for classification (and education), there are limitations in making the intellectual leap to the next level of pathogenetic complexity. For example, gene knockout mutations are relatively easy to generate in mice, and increasingly in other species. Many investigators see this technique as a facile way to isolate the physiologic role of a particular gene product, to generate an animal model of a given disease, or to serve as the background strain into which a defined mutation is introduced. There is no question that the approach has been brilliantly successful in a number of instances; however, the pitfalls have been underemphasized. For example, some mice homozygous for the absence of transforming growth factor- β_1 (TGF- β_1) are born normal in appearance and survived, both unexpected results given the prominent role of this cytokine in many aspects of development. The reason is “rescue” of the embryo by maternal TGF- β_1 that presumably crosses the placenta in sufficient quantity to replace the deficient fetal sources. But at a more fundamental level, the “null” mutant animal cannot be viewed as an artificial isolated system focused on that deficient gene product. Rather, the mutant strain is a complex homeostatic system capable of responding to loss of a specific protein, even compensating for it. Thus, if the null strain shows no phenotype, it would be inappropriate to conclude that the missing protein is not important to the physiology of a given system (physiologic or developmental).

The actual effect of the variant may be loss of function at the protein level, but gain of function at the cellular

level. For example, Rett syndrome (OMIM*312750) is a pleiotropic, severe neurologic disorder that primarily affects girls. The cause is mutation of the *MECP2* locus, which encodes a protein that represses transcription of other genes. Mutations in *MECP2* (OMIM*300005) that inactivate the protein result in enhanced or inappropriate production of proteins in various tissues, most obviously the brain.

13.7 CONCLUSIONS

All of the steps that occur between causation and the bedside constitute pathogenesis. When considering the genetic causal factors and variations in pathogenesis, the term “pathogenetics” applies. The phenotype can be studied at a coarse level, such as the clinical features, and this is the object of evolution. Alternatively, the phenotype can be studied at more and more refined levels, for example, biochemical pathways, to specific enzymatic defects, and to aberrations of messenger RNA processing. These intermediate steps, which require traditional disciplines of pathology, physiology, and biochemistry, help to elucidate pathogenetics. In the vast majority of instances, only through understanding pathogenetics will novel and effective therapies emerge.

The prognosis of a disease is largely a matter of pathogenesis. For instance, its age of onset, the rapidity of its course, and the vulnerable points at which disease and complications may occur all depend on details that in principle, as much as in fact, may be difficult to infer even from the most detailed knowledge of the basic defect. Some knowledge of the prognosis may come from “black box” empirical inquiries—the natural history of myotonic dystrophy, for instance—but this course calls for extensive data, and there may be disturbing discrepancies between one study and another that are not readily reconciled. If the pathogenesis is understood, even partially, more incisive methods may be available, including direct measurements of the progress of components of the disease. For instance, the pathogenesis of familial polyposis coli is not clearly established, but currently the course of this disease and its response to treatment are easier to study than Alzheimer dementia. Refined studies at the molecular level make for very precise statements about etiology. It is tempting, but rather treacherous, to view pathogenesis in the same way. But where the concern lies in either the assessment of morbidity or the study of the population and eugenic behavior of the mutant,

to attach too much weight to refined biochemistry may push the precision of the statement at the expense of its significance. For the overt clinical pattern and the target of selection are very coarse matters; the many modifying factors, which to the basic scientist are largely a nuisance, may have important attenuating effects on the course of the disorder.

Many advances in therapeutics have resulted from largely empirical reasoning as to choosing an approach and from an understanding of natural history in judging whether the therapy was successful. A more rational approach to targeting therapy is based on an understanding of pathogenesis. Some fondly held the hope of circumventing “indirect” therapies for genetic disorders by simply replacing the defective gene. But considerable experience has amply shown the general fallacy in this approach. Until the molecular pathogenesis of a disorder is elucidated, the effects of simply adding back, or even replacing, a gene that should have been functioning perhaps from conception will be as empirical as any-thing physicians had available in the 18th century.

ACKNOWLEDGMENTS

The original chapter on which this revision is based owes much to the research and insights of the late Edmond A. Murphy, MD, ScD.

REFERENCES

- [1] Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470:204–13.
- [2] Lander ES. Initial impact of the sequencing of the human genome. *Nature* 2011;470:187–97.
- [3] Barondes RD. Extrahuman sources of polio virus; new concept on the pathogenesis of the viruses. *Mil Surg* 1949;105:400–8.
- [4] Barondes RD. Duodenal ulcer; pathogenetics, and the re-evaluation of therapeutics. *Mil Surg* 1951;109:720–31.
- [5] Wolf U. The genetic contribution to the phenotype. *Hum Genet* 1995;595:127.
- [6] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 1918;52:399–433.
- [7] Bernard C. *De la Physiologie Gèneérale*. Paris: Hachette; 1872.
- [8] Cannon WB. *The wisdom of the body*. New York: Norton; 1932.
- [9] Wiener N. *Cybernetics or control and communication in the animal and the machine*. New York: John Wiley & Sons; 1948.
- [10] Murphy EA, Pyeritz RE. Homeostasis. VII. A conspectus. *Am J Med Genet* 1986;24:735–51.
- [11] Hebebrand J, Remschmidt H. Anorexia nervosa viewed as an extreme weight condition: genetic implications. *Hum Genet* 1995;595:1–11.
- [12] Waddington CH. *The strategy of the genes*. London: Allen & Unwin; 1957.
- [13] Murphy EA, Berger KR, Trojak JE, Sagawa Y. Angular homeostasis. V. Some issues in genetics, ontogeny and evolution. *Am J Med Genet* 1988;31:963–79.
- [14] Murphy EA, Berger KR, Pyeritz RE, Sagawa Y. Angular homeostasis. VI. Threshold processes with bivariate liabilities. *Am J Med Genet* 1990;36:115–21.
- [15] Mallo M, Wellik DM, Deschamps J. Hoxgenes and regional patterning of the vertebrate body plan. *Dev Biol* 2010;344:7–15.
- [16] Pauli A, Rinn JL, Schier AF. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* 2011;12:136–49.
- [17] Sampath K, Ephrussi A. *Development* 2016;143:1234–41.
- [18] Laland KN, Kumm J, Van Horn JD, Feldman MW. A gene-culture model of human handedness. *Behav Genet* 1995;25:433–45.
- [19] Llaurens V, Raymond M, Faurie C. Why are some people left-handed? An evolutionary perspective. *Philos Trans R Soc B* 2009;364:881–94.
- [20] Halpern DF, Coren S. Handedness and life span. *N Engl J Med* 1991;324:998.
- [21] Spalletta G, Long JD, Robinson RG, et al. Longitudinal neuropsychiatric predictors of death in Alzheimer’s disease. *J Alzheimers Dis* 2015;48:627–36.
- [22] Macario AJL, de Macario EC. Sick chaperones, cellular stress, and disease. *N Engl J Med* 2005;353:1489–501.
- [23] Tabas I, Ron D. Integrating the mechanisms of apoptosis induced by endoplasmic reticulum stress. *Nat Cell Biol* 2011;13:184–90.
- [24] Xu C, Bailly-Maitre B, Reed JC. Endoplasmic reticulum stress: cell life and death decisions. *J Clin Invest* 2005;115:2656–64.
- [25] Pyeritz RE. Pleiotropy revisited. *J Med Genet* 1989;34:124–34.
- [26] Siriam G, Martinez JA, McCabe ER, et al. Single-gene disorders: what role could moonlighting enzymes play? *Am J Hum Genet* 2005;76:911–94.
- [27] Brenner S. Human genetics: possibilities and relatives. *Ciba Found Symp Excerpta Medica* 1973;66:1–3.
- [28] Maddox J. Genes and patent law. *Nature* 1994;37:1270.
- [29] Osler W. Chauvinism in medicine. *Montreal Med J* 1902;31:684–99.

Twins and Twinning

*Mark P. Umstad^{1,2}, Lucas Calais-Ferreira^{3,4},
Katrina J. Scurrah³, Judith G. Hall⁵,
Jeffrey M. Craig^{6,7}*

¹Department of Maternal-Fetal Medicine, The Royal Women's Hospital, Melbourne, VIC, Australia

²University Department of Obstetrics and Gynaecology, University of Melbourne, Melbourne, VIC, Australia

³Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, VIC, Australia

⁴CAPES Foundation, Ministry of Education, Brasilia, Brazil

⁵Departments of Medical Genetics and Pediatrics, British Columbia's Children's Hospital, Vancouver, BC, Canada

⁶Deakin University School of Medicine, Geelong, VIC, Australia

⁷Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC, Australia

14.1 INTRODUCTION

Twins have sparked curiosity in our society since ancient history, such as the Biblical accounts of Esau and Jacob, a pair of twins who were physically similar in spite of striking differences in their personalities. In science, Galton [1] is recognized as the pioneer in involving twins as participants in what we know today as twin studies. However, it is arguable whether he had a correct understanding of zygosity and its implications in relation to Mendel's laws of inheritance which were, back then, still largely undiscovered [2].

The existence of two types of twins was first noted in the 19th century [3], but the realization that monozygotic (MZ or "identical") twin pairs share close to 100% of their genetic material, while dizygotic (DZ or "fraternal") pairs share on average 50% of genetic variation, as any other sibling, was only conceptualized in the early 1900s [4]. Fisher [5] then developed quantitative genetic theory, which allowed for the quantification of similarity in human traits within groups of MZ and DZ twin pairs, which can be seen as an important increment in the "collective" creation

of the classical twin design and its early applications in the 1920s. The first twin registries began to be established about 30 years later and facilitated the collection of valuable prospective and retrospective longitudinal data.

Since then, the fast-paced development of technology and analytical techniques has substantially advanced the core understanding of the twinning process and consequently the applicability of twin designs in genetic epidemiology. The further understanding of atypical twinning and the importance of differences in placentation and chorionicity have played a role in shaping the perception that traditional models of twinning may be no longer sufficient to explain the phenomena [6] (see Section 14.2). This perception has an obvious effect not only on medical practice related to twins, but also on some of the assumptions under which the twin study designs operate. In spite of such challenges, twin studies continue to be recognized as an important tool in any health researcher's repertoire, especially in the current era of omics and molecular studies [7].

This chapter summarizes what is known about the different types of twins, including frequency within the population; the mechanisms of twinning, including genetic contributions; and practical advice for working with twins as patients. It also discusses the value of twin research, including landmark studies and modern applications, and showcases how global collaborative networks of twin registries and researchers are reshaping the way in which twin studies can be relevant to the understanding of human traits and diseases toward prevention, prognosis, and treatment.

14.2 TYPICAL TWINNING IN HUMANS

The traditional model of twinning proposes that DZ twins result from fertilization of two distinct ova by two separate spermatozoa and MZ twins are the

product of a single ovum and sperm fertilization that subsequently divides to form two embryos. The most widely accepted model of MZ twinning is based on the unproven hypothesis of postzygotic division of the conceptus [6]. In this model the numbers of fetuses, chorions, and amnions are determined by the timing of embryo splitting (Table 14.1, Figs. 14.1 and 14.2).

Herranz [8] argued that the postzygotic splitting model lacked scientific evidence and that factors initiating cleavage have not been specified. He also noted that the rate of postzygotic splitting becomes more unlikely with the passage of time and that splitting has never been observed in vitro. He proposed an alternative theory of MZ twinning based on two principles:

- 1. MZ twinning occurs at the first cleavage division of the zygote; and

TABLE 14.1 Chorionicity and Amnionicity by Time of Zygote Splitting					
Zygosity	Twins	Time of Split	Chorions	Amnions	Fetal Mass
Dizygotic	DC DA	No split	2	2	2
Monozygotic	DC DA	Days 1–3	2	2	2
Monozygotic	MC DA	Days 3–8	1	2	2
Monozygotic	MC MA	Days 8–13	1	1	2
Monozygotic	Conjoined	>Day 13	1	1	1

DA, diamniotic; DC, dichorionic; MA, monoamniotic; MC, monochorionic.

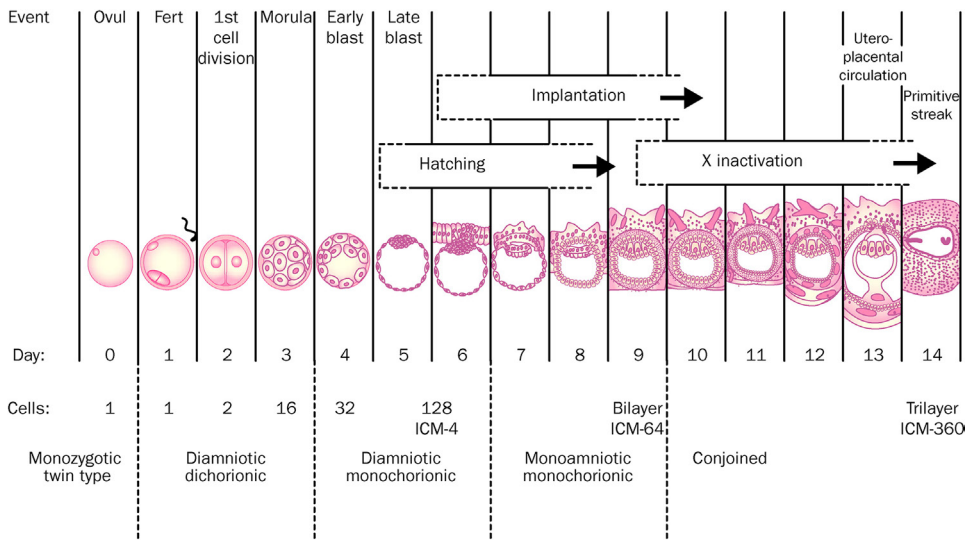


Figure 14.1 Process of Monozygotic Twinning During Postfertilization. Ovul, ovulation; fert, fertilization; divis, division; blast, blastulation. Reproduced from Hall, J. (2003). Twinning. *The Lancet*, 362(9385), 735-743. doi: 10.1016/s0140-6736(03)14237-7, with permission from Elsevier.

2. subsequent chorionicity and amnionicity is determined by the degree of fusion of embryonic membranes within the zona pellucida [8].

Both the traditional “fission model” and the “fusion model” of Herranz are unsubstantiated [9].

14.3 ATYPICAL TWINNING

Insights and challenges to the traditional models of twinning are seen in the variety of atypical or unusual twins. These variations from the usual ideas of twinning are discussed in this section.

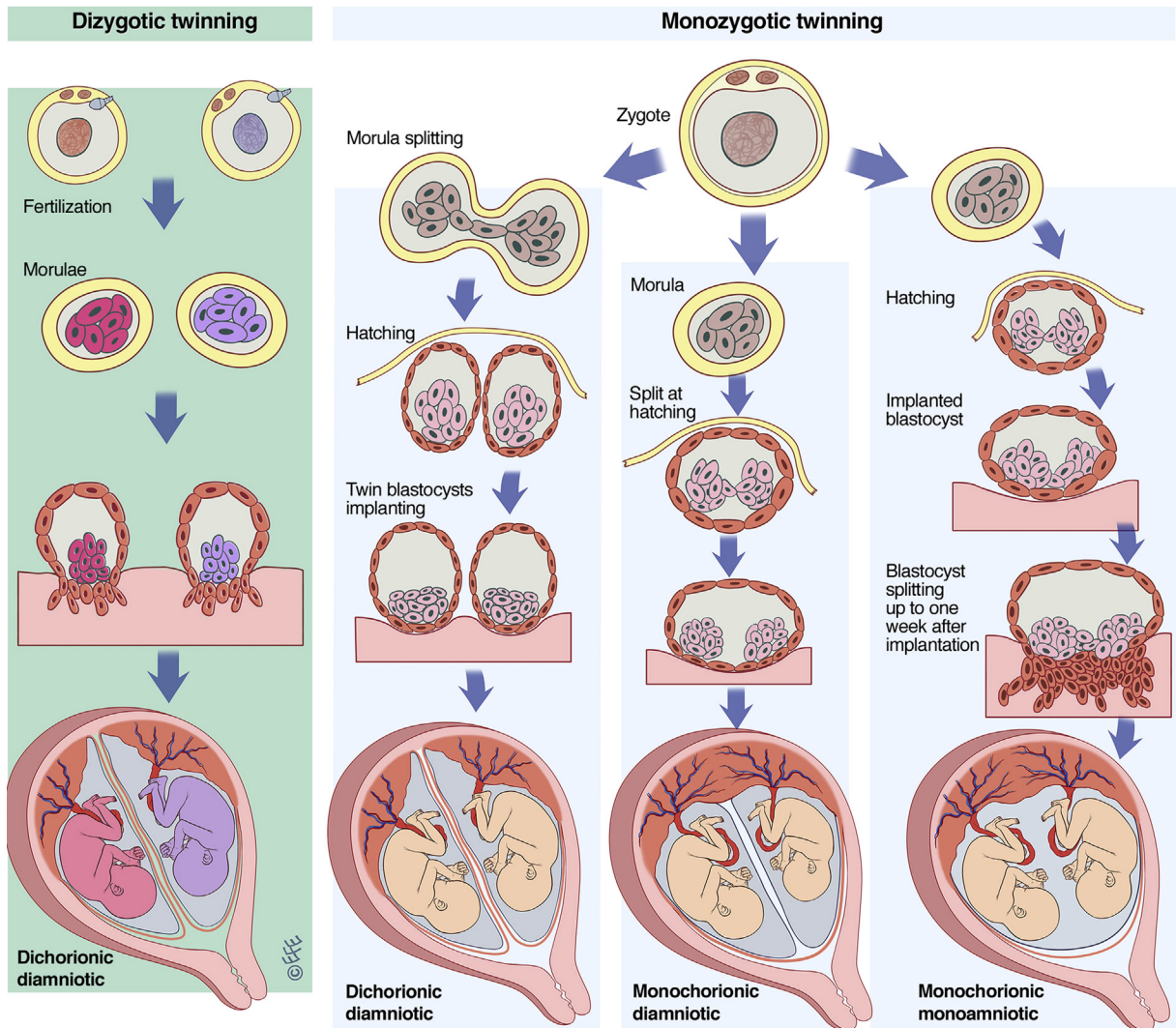


Figure 14.2 The Formation of the Main Types of Twins. DZ twins result from two separate fertilization events and are dichorionic and diamniotic. MZ twins result from the splitting of a single embryo early in gestation. Approximately one-third of MZ twins split early and are dichorionic and diamniotic. Approximately two-thirds of MZ twins split later and are monochorionic and diamniotic. Approximately 1% of twins split even later and share a single chorion and amnion. *Reproduced from McNamara, H.C., Kane, S.C., Craig, J.M., Short, R.V., Umstad, M.P., 2016. A review of the mechanisms and evidence for typical and atypical twinning. Am J Obstet Gynecol 214, 172–191.*

14.3.1 Chimeric Twins

A chimera is a single organism containing two populations of genetically distinct cells originating from two different zygotes. Chimerism has now been extensively described in monochorionic DZ twins. The etiology of human twin chimerism in monochorionic zygotic twin pregnancies is incompletely defined [10].

Theories proposed to explain chimeric twinning include the presence of placental anastomoses that allow early transfer of genetic material [11–14], fusion of elements of two genetically distinct zygotes [15–18], and fertilization of a binovular follicle [19–21]. The evidence for the last of these theories is poor. The importance of recognizing chimerism is to dismiss the longstanding dogma that monochorionicity is indicative of monozygosity.

Clinically, individuals with blood or tissue chimerism may be at increased risk in the context of transfusion or transplantation [22]. Chimeric twins may also exhibit reproductive dysfunction analogous to that of the bovine freemartin [23,24].

14.3.2 Mirror-Image Twins

Mirror-image twinning is observed in as many as 25% of MZ twin pairs. These twins have phenotypic features that are asymmetrical, including the direction of hair whorls, cleft lip and palate, unilateral eye and ear defects, and tumor patterns [25–36].

Mirror-image central nervous system abnormalities, including arachnoid cysts, colpocephaly and optic glioma have been described [37–40]. Although higher-order cerebral functions, including dominant handedness [41], eye dominance [32], and cerebral lateralization [42] exhibit mirror asymmetries, Derom et al. found there is no evidence to suggest that discordant handedness represents mirror-imaging [41]. The traditional model of twinning explains mirror imaging as a consequence of late zygotic splitting between days 9 and 12 [3,36].

14.3.3 Polar Body Twins

A polar body is a small cellular byproduct of the meiotic division of an oocyte. Fragments remain within the zona pellucida following apoptosis within 17–24 h of formation [43]. It has been proposed that fertilization of an ovum and its first or second polar body by two distinct sperm may result in polar body twinning [6].

Bieber et al. [44] described a monochorionic twin pregnancy complicated by twin-reversed arterial

perfusion sequence. The normal fetus had an XY karyotype and the acardiac female was triploid XXX. Genetic studies and human leukocyte antigen (HLA) typing suggested independent fertilization of a haploid ovum and its diploid first polar body. The authors proposed that the proximity of the ovum and its first polar body allowed development of distinct inner cell masses with a common trophoblast.

Fisk and colleagues [45] performed cytogenetic analyses on nine twin reversed arterial perfusion sequence-affected pregnancies. They calculated the likelihood of fertilization of an ovum and its first polar body at less than 3.6% and for its second polar body at less than 0.0003%. They argued against the existence of polar body twinning, suggesting previous reported cases were due to chimerism.

14.3.4 Vanishing Twin Syndrome

Vanishing twin syndrome (VTS) refers to the spontaneous loss of an embryo or fetus in the first trimester of a multiple pregnancy. The rate of VTS is underreported in spontaneous twin pregnancies as most occur prior to 9 weeks' gestation and early ultrasound may not have been performed. VTS is well recognized in assisted reproductive technology (ART) pregnancies, with rates ranging from 10% to 18% [46–48]. There is conflicting evidence regarding the impact of a vanishing twin on the remaining pregnancies. Generally, when compared with singletons, survivors are smaller, have an increase in congenital anomalies of the vascular type, and have higher rates of preterm birth, with the risk being inversely proportional to the timing of loss [47,49].

Initial reports suggested that VTS might increase the risk of cerebral palsy for survivors [50], but subsequent studies have revealed no statistically significant increase in cerebral palsy or adverse neurological sequelae [47,51].

VTS has a significant potential impact on prenatal screening and diagnosis. A vanishing twin increases pregnancy-associated plasma protein A on average by 21%, alpha-fetoprotein by 10%, and dimeric inhibin A by 13% [52]. These changes influence biochemical-based screening programs.

Fetal haplotypes remain detectable via noninvasive prenatal testing using cell-free fetal DNA for up to 8 weeks after fetal demise. Careful counseling and result interpretation are required.

14.3.5 Superfetation

Superfetation refers to fertilization and implantation of a second conception during pregnancy.

Upregulation of the hypothalamic–pituitary–ovarian axis by luteal and then placental progesterone in the first trimester of pregnancy suppresses ovulation and makes the possibility of spontaneous superfetation very unlikely [53,54]. Previously described cases of superfetation are likely to represent either fetal growth discordance, interval delivery, or sonographic error [53]. The advent of ART means that natural barriers to superfetation can be overcome and cases of superfetation as a consequence of ART, with or without additional spontaneous conception, have been reported [55,56].

14.3.6 Superfecundation

Superfecundation refers to the fertilization of oocytes via separate instances of coital or artificial insemination during a polyovulatory period. Heteropaternal superfecundation occurs after coitus with multiple partners. Monopaternal superfecundation occurs after coitus with one partner on multiple occasions. Both spontaneous superfecundation and fecundation associated with ART have been widely reported [55,57–59].

It has been proposed that 1 in 400 twin pairs born to married white women in the United States are the result of heteropaternal superfecundation [60]. Monopaternal superfecundation has an estimated prevalence of 1 in 12 DZ twins born to mothers in the United Kingdom [61]. The reported incidence is increasing because of increased paternity testing [57].

14.3.7 Complete Hydatidiform Mole with Coexistent Twin

A multiple pregnancy with a complete hydatidiform mole (CHM) and a coexisting live fetus (CLF) is characterized by the presence of a fetus with a normal karyotype, anatomy, and placentation, alongside a molar component with no identifiable fetal parts and a placenta with diploid paternal chromosomes. The incidence varies from 1 in 22,000 to 1 in 100,000 pregnancies [62,63]. A multiple pregnancy with CHM-CLF must be distinguished from a singleton pregnancy with a partial hydatidiform mole in which the fetus has triploidy resulting from dispermic fertilization of a haploid normal oocyte. It must also be differentiated from a twin pregnancy with a normal twin in one sac and a partial mole in the other sac and also from

mesenchymal dysplasia, which is associated with an enlarged cystic placenta and fetal growth restriction.

The live birth rate for CHM-CLF varies from 21% to 40% [64,65], but the pregnancies can be complicated by hyperemesis gravidarum, thyrotoxicosis, vaginal bleeding, severe preeclampsia, and fetal death.

Rates of persisting gestational trophoblast disease after CHM-CLF are independent of gestation and whether the pregnancy is terminated or allowed to continue [66]. Persistence rates for gestational trophoblastic disease range from 14% to 50% [64,65].

14.3.8 Fetus-in-Fetu

Fetus-in-fetu, or a parasitic twin, refers to one or more partially formed fetuses situated entirely within the body of another normally formed fetus [6]. This occurs in approximately 1 in 500,000 births. A fetus-in-fetu is an MZ monochorionic diamniotic twin that occurs as a consequence of persistent anastomoses of the vitelline circulation which leads to the absorption of one twin inside the other during the ventral folding of the trilaminar embryonic disc [67]. A fetus-in-fetu should be contrasted to a teratoma, although both can be located within the retroperitoneum and are histopathologically similar. A fetus-in-fetu is characterized by the presence of vertebrae with appropriately organized limbs and organs. Fetus-in-fetu is part of the parasitic continuum of conjoined twins, acardiac twins, and teratomata [68]. They have been identified in the mediastinum, scrotum, mouth, and skull. Usually there is one fetal mass but up to 11 fetuses-in-fetu have been reported [69–72].

Fetuses-in-fetu do not demonstrate malignant potential, unlike teratomata, but may cause significant mass effects necessitating surgical removal [73,74].

14.4 PLACENTATION

DZ twins, with the exception of chimeric twins, have dichorionic diamniotic placentation. That is, they have two placentas with two chorions and two amnions. The placentas in DZ twins can fuse to appear as one placenta but they are functionally independent and there are usually no communications between the two.

The placentation of MZ twins is determined by the assumed timing of postzygotic splitting (Fig. 14.1). If the split occurs on days 1–3, up to the morula stage, dichorionic diamniotic twins are formed. A split between

days 4 and 6, during which blastocyst hatching starts, results in monochorionic diamniotic twins. If the split occurs between days 7 and 9 the result is monochorionic monoamniotic twins. If no split occurs by day 10, conjoined twins are formed [3,6].

Approximately 30% of MZ twins are dichorionic diamniotic, around 70% are monochorionic diamniotic, and 1% are monochorionic monoamniotic. In clinical practice, MZ dichorionic diamniotic twins have no more obstetric or perinatal complications than DZ dichorionic diamniotic twins, as they do not share a circulation.

Monochorionic diamniotic twins, with their shared circulation, are complex pregnancies with a range of significant clinical risks. If there is unequal sharing of the circulation, then twin–twin transfusion syndrome or twin reversed arterial perfusion sequence may ensue. These conditions result from one twin having an excess of blood and the other a deficiency. In their untreated state these conditions have very high rates of perinatal mortality and morbidity, particularly neurological injury to surviving twins. Rarer complications of monochorionic diamniotic twins include twin anemia polycythemia sequence (acardiac twins), and selective growth restriction in one of the pair of monochorionic twins.

Monochorionic monoamniotic twins, that is twins who share both a single placenta and a single amniotic sac, universally have entangled umbilical cords. This can lead to sudden fetal demise in either or both twins, and is regarded as a very high-risk multiple pregnancy.

14.5 INCIDENCE OF TWINS

14.5.1 Incidence of Monozygotic Twinning

MZ twins occur at a constant rate throughout the world, at approximately 1 in 250 pregnancies. The spontaneous rate of MZ twinning is unaffected by maternal age, height, weight, or parity [75]. The rate of MZ twinning is increased with ART. MZ twinning is sporadic, usually with nonrecurrence in families, but there are very rare exceptions (see Section 14.9.2).

14.5.2 Incidence of DZ Twinning

The incidence of spontaneous DZ twins is highly dependent on many factors: age, nationality, parity, height, weight, and family history all influence the incidence of spontaneous DZ twinning.

The highest rate of spontaneous twinning is seen in Africa, particularly East Africa, with rates as high as 1 in 20 [76], and the lowest in Japan with rates as low as 1 in 500 births [77]. Intermediate rates are seen in Europe, North America, and Australia.

DZ twinning increases with maternal age, with the peak at age 35–39 years, the next highest at 40–44 years, and then the 30–34-year range. DZ twinning also increases with higher parity, maternal height, and maternal weight [78–80].

In contrast to MZ twinning, DZ twinning does run in families (see Section 14.9.3).

DZ twinning rates following ART pregnancies peaked in the mid 2000s, and in most developed countries with careful regulatory monitoring, those levels have dropped substantially. Twinning rates after ART peaked at around 25% of all multiple pregnancies in Australia in 2005 but are now less than 10% [81].

The major influence on the DZ twinning rate has been ART. Multiple pregnancy rates with oral agents such as clomiphene citrate are in the order of 5–8%, and with injectable gonadotrophins are in the order of 20–25%. These rates are dependent on the level of ultrasound monitoring of follicular development.

14.6 SEX RATIO IN MZ TWINNING

The sex ratio (the proportion of males compared to the combination of males and females) is lower in MZ twins than in either DZ twins or singletons [82,83]. In DZ twins and singletons the sex ratio is 0.514 and for all MZ twins 0.496 [84]. For monoamniotic twins, including conjoined twins, the sex ratio is 0.2; that is, around 80% of all monoamniotic twins, including conjoined twins, are female.

It has been shown that the number of female embryos is less than male embryos at specific stages of early development in mice, suggesting that female conceptions are at higher risk of late splitting of the embryo due to their delayed embryonic development [85].

14.7 STRUCTURAL DEFECTS IN TWINS

Both MZ and DZ twins are known to have an increased risk for structural defects compared to singletons [86,87]. Deformations are increased because of the external pressure due to two growing fetuses sharing the space usually meant for one. All anatomical sites

TABLE 14.2 Structural Defects in Monozygotic Twins

Associated with the Twinning Process due to Incomplete Splitting of the Embryo	Due to Shared Vascular Connections or to in utero Death of Second Twin	Due to Fetal Constraint or Crowding in utero
Conjoined twin, fetus in fetu	Acardia (TRAP sequence), asplenia, microcephaly, hydrocephaly, intestinal atresia, aplasia cutis, terminal limb defects, gastroschisis, fetus papyraceus	Craniosynostosis, positional defects of the foot, bowing of the limbs, some contracture

appear to be involved in the overall increase rather than just those expected to be increased by external compression. Structural defects in MZ twins are three times more frequent than among DZ twins and two to three times more frequent than in singletons [88–90]. The structural defects seen in MZ twins can be divided into three groups (Table 14.2) based on the type of process producing them. Because of the high rate of structural anomalies in twins, and the high rate of conversion of a twin pregnancy to a singleton (see below), many singletons born with congenital anomalies may have started as part of a twinning process.

14.8 ZYGOSITY DETERMINATION

In the past, the only way of differentiating between MZ and DZ twins at birth was their sex and appearance. If twins were of unlike sex, they were said to be DZ, whereas if they were like-sexed and looked identical, they were said to be MZ [91]. Although not always definitive, placentation was another means of establishing zygosity [92]. Other methods that have been used to differentiate between MZ and DZ twins include blood types, blood protein polymorphisms, human leukocyte antigen (HLA) typing, and dermatoglyphic studies that document fingerprints, palm prints, and creases. More recently, DNA analyses have been used to establish zygosity. Notably, zygosity assessment based on physical similarity alone has been shown to be approximately 95% accurate [93].

A statistical approach to establish zygosity for population studies was proposed by Weinberg [94], who used what was later termed the “Weinberg differential method” to estimate the number of MZ and DZ twin pairs. This method assumes that all MZ twins are like-sexed. It also assumes that in DZ pairs the sex of twin pairs occurs at random, so that one half is like-sexed and one half is unlike-sexed (i.e., in one-fourth of DZ twin

pairs both are female, in one half there will be male/female pairs, and in one-fourth both twins are male). Thus, if A is the number of like-sexed twins and B is the number of unlike-sexed twin pairs observed, the estimated proportion of MZ twin pairs would be $(A - B) / (A + B)$ and the estimated proportion of DZ twin pairs would be $2B / (A + B)$. The reliability of the Weinberg method has been questioned by new evidence. For instance, unlike-sexed twins are not always DZ, as when one is 46XY and the other is 45X [95–97].

Today, sex, placentation, DNA fingerprinting [98–100], DNA microsatellites [101] and, recently, single nucleotide polymorphisms [102], are all used to differentiate between MZ and DZ twin pairs, with DNA analysis now being the most accurate method for determining zygosity. The analysis of DNA from cells such as skin fibroblasts or buccal cells may be more accurate than analyzing DNA from blood cells, as chimerism of blood supply is known to occur in both DZ and MZ twins [103,104].

The determination of zygosity at birth of all same-sex twins has been recommended by geneticists and epidemiologists since the early 1990s [105–108]. This recommendation was made because the information is useful in determining organ transplant compatibility, for research concerning the biology and pathology of MZ twinning, and for the investigation of discordance and concordance for many genetic diseases and disorders with multifactorial inheritance. Knowledge of zygosity also provides twins and their families with an understanding of their identity and relationships. Additionally, increased education for health professionals and the community will help ensure that twins have the same opportunity to participate in research as nontwins; to do so requires clarity about their zygosity.

Three parameters are used together to differentiate between MZ and DZ twins: chorionicity, sex, and

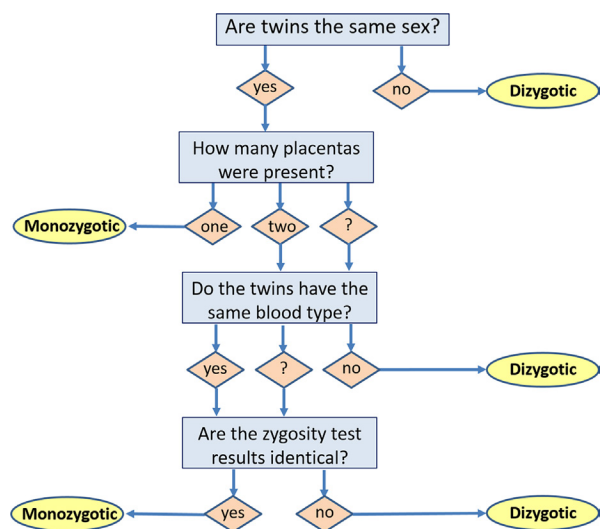


Figure 14.3 Decision Tree for Determining Zygosity in Twins. *Number of placentas should be determined by ultrasound prior to 12 weeks' gestation and confirmed at birth by physical examination of membranes.

genetic zygosity testing (Fig. 14.3) [3,109,110]. Blood type is also sometimes useful. Chorionicity is most accurately determined by the thickness of the membrane between the twins as determined by ultrasound examination, most accurately between weeks 6 and 9 of gestation, but it may be determined up to week 14 [111]. In dichorionic twins a thick membrane forms a lambda shape and separates the twins. In monochorionic twins this membrane is much thinner and joins the placenta to form a "T" shape. Ultrasounds taken later in gestation are less reliable due to the increased crowding of twins in the uterus. Sonography may provide additional clarification if needed. Physical examination of the intertwin membranes at birth should also be used to determine chorionicity. This will provide confirmation of early ultrasound data and determination of chorionicity in twins without early ultrasound information. Again, dichorionic membranes are thick, opaque, and can be pulled apart, whereas monochorionic membranes are thin, semitransparent, and inseparable. When sex and chorionicity are known, we can assign different-sex twins as DZ and same-sex monochorionic twins as MZ (with rare exceptions in cases of chromosomal abnormalities). The steps in Fig. 14.3 can be used to accurately determine twin zygosity in almost all pairs irrespective

of age. When blood type is known, we can assign twins with different blood types as DZ. In same-sex twins, if chorionicity and blood type are unknown then a zygosity test is required to accurately determine zygosity. Same-sex dichorionic twin pairs with the same blood type also require a zygosity test to determine their zygosity.

14.8.1 Incorrect Assumptions about Zygosity

Our research [112] and that of others [113,114] have found that a substantial proportion of parents and twins are misinformed about their zygosity status, and that this misinformation may come from parents or medical professionals.

Incorrect assumptions lead to misclassification of twins by healthcare professionals and the twins and their families. The main two incorrect assumptions being that (1) all dichorionic twin pairs are DZ and (2) genetically identical MZ twins must have identical phenotypes. The false belief that all dichorionic twins are DZ can lead to one-third of MZ twins being incorrectly categorized as DZ. In addition to this, the number of chorions can be difficult to determine from ultrasounds. Dichorionic twin pairs may have fused placentas, which can be mistaken for a single placenta without careful examination at birth. This can lead to misclassification of DZ pairs as MZ. MZ twins usually look and behave more similarly than DZ twins due to their greater genetic similarity. However, MZ twins are often not physically and behaviorally identical due to differences in the environments they encounter from conception onwards (see Section 14.11.2). This can lead to the misclassification of phenotypically different MZ pairs as DZ.

14.8.2 The Importance of Zygosity Knowledge

Accurate knowledge of zygosity is important to twins and people with whom they come into contact. Below we summarize the range of medical, personal, financial, scientific, legal, and ethical reasons supporting the testing and reporting of twin zygosity [106,115,116].

MZ pairs are perfectly compatible organ donors for one another, requiring much less post-transplant immunosuppression than DZ twins and have better chances of long-term survival. As nearly all diseases have at least some genetic component to their origin, the diagnosis of a disease in one twin means the co-twin is at increased risk of that disease, more so for MZ pairs than DZ pairs. Genetic sequence data will almost always be the same for MZ co-twins and the implications of this for the

co-twin should be considered if testing is undertaken. In the event that one or both twins die before or soon after birth, it is vital that parents and/or the surviving twin have this fundamental information about twin zygosity as it bears upon the immediate bereavement response, the long-term identity of the surviving twin, and future family planning. Accurate determination of zygosity is also important postnatally for estimation of the likelihood of the mother or close female relatives giving birth to further sets of twins, because only DZ twins can run in families. Zygosity knowledge is also important for understanding the physical and behavioral differences and similarities between twins.

Increased understanding of zygosity helps define social relationships and helps define twins as individuals. It helps avoid embarrassment over uncertainty when asked about by zygosity by family, friends, teachers, and strangers. It can provide peace of mind and positive emotional responses for twins and their families. Some twins experience significant emotional stress if they discover later in life that their belief about their zygosity is incorrect. Knowledge of zygosity is also a prerequisite for twin research, and many twins feel they are unable to participate because of this [112].

Knowledge of a genetic disorder manifested in only one of a pair of identical twins is likely to lead to early detection in the second twin, thus leading to improved health outcomes and potentially savings in costs for treatment and management compared to detection at a later stage of disease. Costs incurred for zygosity testing would be outweighed, in the cost of suspected genetic disorder, by the savings from genetic testing of the second twin after a genetic diagnosis in the first (this will only happen with MZ twins).

Accurate knowledge of zygosity will affect the results and findings of medical research involving twins and saves both time and expense for researchers and participants as additional testing would not be required.

The International Council for Multiple Birth Organisations (ICOMBO) and the International Society of Twin Studies (ISTS), in their *Declaration of Rights* state that “Parents have a right to expect accurate recording of placentation, determination of chorionicity and amniocinity via ultrasound, and the diagnosis of zygosity of same sex multiples at birth” and that “older, same sex multiples of undetermined zygosity have a right to testing to ascertain their zygosity” and “Zygosity should be respected as any other human trait and deserves

the same privacy rules.” (<http://icombo.org/wp-content/uploads/2010/11/Declaration-of-Rights-2014.pdf>). Respect for the individual recognizes the importance of the concept of identity for a person, which is important for wellbeing, and for avoiding the harm of misinformation. Accurate knowledge of zygosity at birth also avoids any erroneous assumptions from the outset, including those that would result in damaging psychological or emotional impact if zygosity assumptions are proven incorrect, for example if twins had assumed incorrectly that they were identical.

In summary, research has shown that accurate knowledge of twin zygosity can be very reassuring for both twins and their families [112,113].

14.9 THE ETIOLOGY OF TWINNING

14.9.1 Genetic Causes of MZ Twinning

There have been a number of reports of families in which MZ twinning occurs more frequently than expected [117–121], which has been termed “familial MZ twinning.” Interestingly, there does not seem to be an increase in congenital anomalies among the MZ twins in these families. Because familial MZ twinning has been reported on both the maternal and paternal sides of the family, it has also been suggested that it may be caused by a single gene effect that is unaffected by the sex of the parent transmitting the gene [120,122]. However, data from Lichtenstein and colleagues [123] have suggested that there is no paternal effect on familial MZ twinning.

Recently, a gene has been characterized that is likely to play a role in MZ twinning. *PITX2*, a transcription factor, was shown to be involved in the formation of embryonic axis formation in a model of cleavage-induced “experimental” twinning in chickens [124]. The authors concluded that this meant that the pathway associated with this gene “guarantees” that the two products of an early embryonic split are each “guaranteed” to form separate embryos, and that the opposing axes may even explain a proportion of mirror twins. Genetic studies of rare familial MZ twinning may shed light on whether variants of this and other genes influence the likelihood of MZ twinning in such cases.

14.9.2 Genetic Causes of DZ Twinning

There are many reports of familial DZ twinning [125]. The female members of these families are thought to have an inherited predisposition to multiple ovulation

and in turn have a higher number of DZ twin pairs when compared to the general population [126]. The risk of having twins is up to 2.5 times higher for a woman with a sister with DZ twins than it is for the general population [127,128]. An established association between higher gonadotrophin levels and higher incidence of DZ twins in certain families is thought to be the basis for familial DZ twinning. While there appears to be some controversy whether this is an autosomal maternal or paternal effect [123,127], in reality a twinning gene could be inherited through either the maternal or paternal side, although it will only be expressed in females. It is possible that some genetic disorders may also predispose to DZ twinning [129].

Studies of species other than humans have revealed a number of genes that contribute to DZ twinning. The study of sheep has proved particularly informative, as sheep typically give birth to a single offspring at a time, but some strains have high incidences of multiple births [130]. To date, three genes have been confirmed to influence DZ twinning rates in sheep: the growth differentiation factor 9 (*GDF9*) and bone morphogenetic protein 15 (*BMP15*) genes, both of which are expressed in the oocyte and are essential for follicle development, and the bone morphogenetic protein receptor 1B (*BMPR1B*), the receptor for *BMP15* and expressed in multiple cell types in the ovary. Interestingly, while mutations in *GDF9* and *BMP15* can increase twinning rates when one copy is present (i.e., when the mother is heterozygous for the mutation), they can also cause female infertility when two copies are present in one individual (i.e., in homozygous form) [131,132].

At present, only mutations in *GDF9* appear to influence DZ twinning in humans, although such mutations are rare. Screening these genes in large numbers of DZ twinning families has revealed a loss-of-function mutation and a two-base deletion in *GDF9* in heterozygous form in three families [133,134], and it appears that overall genetic variation in *GDF9* is more common in mothers of DZ twins than it is in controls [134]. No such effect has been found for *BMP15* [135] or *BMPR1B* [136]. Interestingly, both *GDF9* and *BMP15* have been implicated in premature ovarian failure [137].

There are a number of additional genes known to have roles in ovulation and DZ twinning in humans [79]. Genetic variants that result in changes in amino acids in the follicle-stimulating hormone receptor (FSHR) protein were suggested to contribute to DZ

twinning [138] and a variant with a known functional effect located in the promoter of the *FSHR* gene was found to segregate with the DZ twinning phenotype in one large family [139], but further studies found no evidence for the involvement of this gene [140]. A variant of the *FSHB* gene, which codes for the beta subunit of FSH, and *SMAD3*, the product of which is involved in gonadal responsiveness of FSH, were recently found to be associated with a higher rate of spontaneous DZ twinning [141]. Both variants are also associated with other aspects of female fertility, such as earlier age at first and last child, confirming the link between fertility and DZ twinning.

Evidence for the involvement of serine proteinase inhibitor clade A member 1 (*SERPINA1*, commonly known as alpha-1-antitrypsin [142,143]), peroxisome proliferator-activated receptor gamma (*PPARG* [144]) and the fragile X (FRAXA) “premutation” [145,146]: has not been borne out in later studies [79]. Family-based “linkage” studies have found no evidence for increased levels of genetic sharing among family members over chromosomal regions in which such candidate genes are located [139,147,148]. Linkage studies have, however, indicated chromosomal locations that may harbor new candidate genes for DZ twinning [139,147,149]. A study including 525 DZ twinning families suggests the presence of such genes on a number of chromosomes, most notably, chromosomes 6, 12, and 20, in Australian and Dutch DZ twinning families and confirmed that DZ twinning is a complex trait likely to be influenced by multiple genes [139]. Much work remains to be done to find the genes underlying the tendency to human DZ twinning.

14.9.3 Other Causes of Twinning

Several nongenetic mechanisms of MZ twinning have been proposed. The finding that mammalian female embryos are somewhat behind male embryos in the number of cells present at a certain stage during early stages of embryonic development [85] and the fact that there is a slightly higher incidence of female MZ twins, particularly among conjoined twins, in which the twinning process is assumed to occur relatively late in the very early embryonic developmental process, support the suggestion that MZ twinning is somehow related to delayed implantation and that the timing of different developmental clocks plays a critical role in MZ twinning.

Several authors have suggested that skewed X-chromosomal inactivation may play a role in female MZ twinning if, during embryogenesis, two different foci were to arise: one expressing the maternal X and the other expressing the paternal X. A number of female MZ twins have been discordant for a variety of X-linked recessive diseases [150], suggesting, and often demonstrating, nonrandom X-inactivation. Goodship and colleagues [151] have tested the hypothesis that skewed X-inactivation can trigger MZ twinning in females by studying umbilical cord tissue in female MZ twins. They observed random X-inactivation in most pairs of female MZ twins, but some showed marked skewing. Thus, it would appear that skewed X-inactivation does not explain all female MZ twinning but could be responsible for the excess of MZ female twins. Interestingly, Tan and colleagues [152] have shown that X-inactivation occurs at different times in different tissues postimplantation in the mouse embryo. These findings suggest that since X-inactivation occurs at the time of tissue differentiation, X-inactivation in blood and skin may not be representative of the rest of the tissue in an organism. To properly determine the exact role of X-inactivation in female MZ twinning, it would be necessary to study many different tissues. Bamforth and colleagues [153] have studied the parent-of-origin of X-inactivation in placental membranes and umbilical cords in twins and triplets. The chorion did show asymmetric X-inactivation in MZ dichorionic twins. Of course, the chorion is not representative of the whole embryo, but may represent processes that occurred early in development. The study also suggested that monochorionic MZ twins may react differently from dichorionic ones, reflecting the importance of timing in the MZ twinning process.

Observations of discordance in the expression of genetic material in MZ twins have suggested the intriguing possibility that some cases of MZ twinning may occur because of epigenetic events [154,155] (explained in [Section 14.10.2](#)). Weksberg and colleagues [156] found differential imprinting in female MZ twins discordant for Beckwith–Wiedemann syndrome and proposed that in such cases either unequal splitting of the inner cell mass or, alternatively, a lack of maintenance of DNA methylation, leads to a loss of imprinting that predisposes to MZ twinning. Such discordance would be expected to arise early in development among cells from a single zygote. A discordance of expression of genetic information could then lead to division of the zygote

into two separate embryos during a specific period, early in development, perhaps from the stage of eight cells to approximately 360 cells in the inner cell mass, when differentiation and primitive streak formation begin. After birth, this discordance of genetic information could be mosaic in each twin, but it could be present to different degrees in the two different twins, sufficient to cause observable phenotypic discordance. In other words, once measurement error and environmental influences are accounted for, genetic discordance or differences in the expression of genetic information should be suspected in cases of discordant MZ twins. This topic is covered in more detail in [Section 14.11.2](#).

Although temperature, delay from the time of ovulation until fertilization or implantation, oxygen supply, and various teratogenic agents have been shown to affect MZ twinning rates in other animals [157], no such factor has been associated with MZ twinning rates in humans. A recent association between an increase of twin births (both MZ and DZ) and periconceptual vitamin supplementation, specifically folic acid, has been reported by Czeizel and colleagues [158], suggesting that adequate maternal nutrition is important for survival to birth for human twins (e.g., loss or conversion to a singleton may occur with inadequate nutrition). The population rate of spontaneous MZ twins seems to be increasing [75], and a 3–5 times increase in MZ twin births has been seen with ART [159,160], perhaps related to ovarian stimulation, disturbance of the zona pellucida, or culturing conditions or handling (e.g., blastocyst transfer) during ART procedures [159]. MZ twins and MZ triplets are frequent among spontaneous triplets [161].

Stockard [162] suggested that MZ twinning may be due to a lack of oxygen prior to implantation, which causes developmental arrest and splitting in the zygote. His work was supported by the finding that the implantation of the ovum is delayed in the armadillo, which results in MZ quadruplets or octuplets [163], and by studies in rabbit and roe deer showing that twinning in these animals is also associated with delayed implantation [164]. These findings suggested that MZ twinning is associated with disturbance of development clocks or thresholds and that delayed fertilization or delayed implantation may play a role in MZ twinning.

On the basis of observations of a higher-than-expected incidence of MZ twins after ART [99,165], Edwards and colleagues [166] suggested that abnormalities or rupture of the zona pellucida may lead to

herniation of the blastocyst and predispose to MZ twinning. Boklage [167–169] estimated that differentiation of the chorion occurs at approximately the fourth day after fertilization and that, in monochorionic MZ twins, the physical separation of two embryos is unlikely if the zona is still intact when the chorion begins to develop. Boklage suggested that if the zona is intact, rather than a physical separation of the MZ twins, there may be “developmental” separation, rendering two groups of cells within a morula that organize themselves separately and continue with embryogenesis separately. A recent study investigating the elevated frequency of MZ twinning resulting from ART [170] found that extended culture (or embryo stage of transfer) was a major risk factor. In one report [171], extended culture of spare 2–10-cell ART embryos resulted in two cases of ectopic adhesion of cells from a blastocyst’s inner cell mass to the opposing inner trophoctoderm wall, which was followed in one by blastocyst splitting and both products hatching. It was proposed that blastocyst collapse and re-expansion observed in mouse blastocysts [172] could occasionally result in adhesion of the inner cell mass to the opposite wall, to which it would transfer a portion of the inner cell mass, triggering MZ splitting.

Other investigators have suggested that twinning itself may be a type of congenital anomaly or an abnormality of development, with the “twinning” fertilized egg (i.e., a fertilized egg resulting in twins) developing at a different rate and in a different way, as compared to a “normal” fertilized egg (i.e., an egg resulting in a singleton). There must be a relatively narrow window during which MZ twinning can occur (normally only up until 11–13 days postfertilization, when the primitive streak forms), and there are a number of different events taking place during post fertilization, including hatching, implantation, genomic imprinting, and X-inactivation (see Fig. 14.1). If the twinning zygote is maturing at a different rate than the normal zygote, the timing for all these events may be shifted and may even occur in an order different from the predicted normal timing for singletons.

14.10 GENETIC AND EPIGENETIC DIFFERENCES WITHIN PAIRS OF MZ TWINS

MZ twins have been described as natural clones; however, this is incorrect. In the course of the development of every large multicellular organism, somatic

mutations as well as epigenetic and stochastic processes lead to distinct genetic differences between MZ twins [109,173–175].

14.10.1 Genetic Differences Within Pairs of MZ Twins

Genetic differences within MZ twin pairs include differences in chromosomal aneuploidy [176–178], uniparental disomy [179], chromosomal rearrangement [180], triplet expansion [181], or nuclear [182,183] or mitochondrial [184,185] point mutations that have occurred postzygotically [175]. Within-pair differences in telomere crossover [186] and length [187] have also been observed.

A small number of case studies using a candidate gene approach have been particularly informative clinically. For example, Vadlamudi and colleagues identified a *de novo* mutation in the sodium channel alpha 1 subunit gene *SCN1A* as the likely cause of the epileptic disorder Dravet syndrome in the affected twin of a discordant MZ pair [188]. The extent to which DNA sequence has been shown to differ within any given MZ twin pair, whether phenotypically similar or discordant, has depended on the genomic platform used. Medium-to-large-scale studies of single nucleotide polymorphism (SNP) arrays, typically measuring 500,000–1,000,000 SNPs, have shown that fewer than 1% of phenotypically normal MZ twins show genetic discordance at single nucleotides and copy number variants [189–191]. The rate of discordance in disease-discordant pairs detected using SNP arrays may be higher [192–194], although numbers of twin pairs in such studies have been low (typically less than 10 pairs per study) and validation of sequence differences using locus-specific techniques has not been performed for all putative variants. Studies of disease-discordant MZ twin pairs on a similar scale have been performed on coding regions using exome sequencing; however, these have either failed to validate putative within-pair differences [195–197] or detected only mosaic variants [198]. More recently, whole-genome sequencing has enabled sequence comparisons across coding and noncoding regions within pairs of MZ twins. Although such studies are still finding null results with small numbers of twin pairs [199,200], two studies are worth highlighting. Francioli and colleagues derived whole-genome sequences from 11 pairs of MZ twins and their parents, eight pairs of DZ twins and their parents, and 231 child–parents trios [201].

Validated variants in MZ twins enabled them to estimate that germline and postzygotic somatic mutations occur at a ratio of approximately 33:1. They also found that only 7% of such variants detected in single twins were specific to that twin only and were more likely than expected to be located in coding regions. Weber-Lehman and colleagues performed whole-genome sequencing on a pair of father and son and the father's twin brother [202]. The authors found five validated SNPs in sperm DNA from the father and in blood DNA from his son, all of which were not present in the father's co-twin. Although none were located in coding regions or in SNP databases, three of the five were located in putative regulatory regions. Only one of the five variant SNPs were found in the father's blood and four were found in his buccal cells.

In summary, we know that there are likely to be genetic differences within many pairs of MZ twins, but until larger studies that compare multiple tissues at different sequencing read depths, we will not know the true frequency of such events or their contribution, if any, to phenotype.

14.10.2 Epigenetic Differences Within MZ Twin Pairs

Epigenetics is the study of changes to gene activity, perpetuated through cell division, that are not accompanied by changes to the DNA sequence. Epigenetic change, which is associated with both early development and aging, involves the addition and removal of small molecules onto DNA and histones—the DNA-packaging proteins. The most understood epigenetic change is DNA methylation—the addition of a methyl group (CH_3) to the cytosine nucleotide of a cytosine–guanine (CpG) dinucleotide. Studies involving genome-wide analysis of DNA methylation have started to show promise in identifying causal mechanisms, and diagnostic biomarkers for complex human diseases [203].

Twin studies have shown that throughout mammalian genomes, epigenetic state is influenced mostly by variation in nonshared environment (75–85% of the variance explained), to a lesser extent by genetic variation (10–20% of the variance explained), with variation in shared environment having the smallest influence [204–208]. Like other traits, the epigenetic state of individual locations throughout the epigenome (the sum total of epigenetic marks in any given tissue) can vary greatly in their components of variance.

Twin studies have demonstrated within-pair epigenetic differences are present at birth, whether at individual loci [209] or throughout the epigenome [210,211]. These results also showed that some MZ twin pairs can be more epigenetically discordant than DZ twins and some unrelated individuals, emphasizing the formative nature of the nonshared intrauterine environment. Also of note, the same studies found that genes with the highest levels of within-pair epigenetic discordance were enriched in those associated with development and morphogenesis. In agreement with this, in an MZ twin pair discordant for the developmental anomaly of caudal duplication, the twins differed in DNA methylation state in a gene, *AXIN1*, whose genetic state had previously been associated with the disorder [212].

As tissues age, their epigenetic state changes as a function of age and environment, and this has been termed epigenetic drift [213]. A cross-sectional study of MZ twins of different ages found that epigenetic discordance on a genome-wide scale increases with age [214]; however, these findings were not replicated in another cross-sectional study [210] or longitudinal studies of infants [215] and adults [216]. Although we do know that the epigenetic state of a small subset of the epigenome does correlate highly with age [217,218], we also know that factors such as tissue and genomic location also influence epigenetic change over time [219,220]. More longitudinal epigenome-wide studies are needed to resolve this issue.

It is also worth noting that nonshared environment includes twin-specific environmental factors, measurement error, and stochastic factors, the latter being an intrinsic part of eukaryotic development [221–223]. This implies that we need to improve the accuracy of measuring phenotypes in twin studies and need to test hypotheses that specific (intrauterine) nonshared environmental factors are associated with the epigenetic state. To date, MZ twin discordance for birth weight, an easily measurable proxy for intrauterine growth and viability, has been associated with epigenetic discordance in genes involved in metabolism when measured at birth [210] or in adulthood [224], and in the growth-associated gene *IGF1R* when measured in adulthood [225]; however, two other epigenome-wide studies found no evidence for differential DNA methylation levels in adult twins discordant for birth weight. Again, more studies are needed to resolve this issue.

14.10.3 Nonshared Environment and Chronic Disease

Twins studies have made a tremendous contribution to the understanding of the causes of chronic disorders, from cancers to neurodevelopmental conditions such as schizophrenia, allergies, and cardiometabolic disorders such as cardiovascular disease [7,173,204,226,227]. Such disorders originate in very early life [228–230] and like the epigenetic state itself, are mainly influenced by the nonshared environment [231]. This finding was initially surprising because maternal diet and lifestyle had been assumed to be influential on the developing fetus; however, multiple nonshared factors have now been shown to influence risk for chronic disease [116,173,232–234]. Most of these factors are associated with the “fetoplacental unit” [235]—the placenta, cords, and fetus. Such factors include uterine implantation site, the placental location at which the umbilical cord is inserted, and the physical characteristics of the umbilical cord such as length, width, and torsion. All these factors have the potential to affect the growth rate of individual twins via the transplacental transport of oxygen, nutrients, and teratogens.

MZ twins can be discordant for inflammatory state [236–240], which is a major contributor to chronic disease risk [241,242]. Twin studies have shown that the nonshared environment dominates as the largest component of variance of immune factors in adults [243]. Prenatal inflammation can occur in the umbilical cord or placenta of a single MZ twin [237,244] and soluble inflammatory factors can pass to the associated fetus [245]. More longitudinal twin birth cohorts are required to improve our understanding of the way the nonshared environment influences health and disease and its epigenetic mediators.

14.11 TWIN RESEARCH: DESIGNS AND ANALYTIC APPROACHES

Studying twins can provide insights about the health of both twins and nontwin individuals—and whole populations. A number of different study designs involving twins are possible and these have different aims, statistical analysis techniques, model assumptions, advantages, and limitations.

14.11.1 The “Classic Twin Design”

One of the most common twin designs, and one that researchers first think of when considering a twin study, is the “classic twin design.” Statistically, familial

resemblance can be quantified by estimating the correlation, either for all types of twins together or separately for MZ and DZ twins. Correlations are always between -1 and 1 , and correlations greater than 0 for pairs of twins suggest that familial resemblance exists for the particular trait. The “classic twin model” initially estimates correlations separately for MZ and DZ pairs and compares these. If the correlation is greater for MZ twins than for DZ twins, this is consistent with (but does not prove) the existence of genetic effects influencing variation. Even when results are consistent with genetic effects, the specific genes responsible usually cannot be identified simply by studying disease traits. In addition, this conclusion relies on several strong assumptions: (1) MZ and DZ twins share their environment to the same extent (the “equal environments” assumption [246]) and (2) the only difference between MZ and DZ twins is the proportion of genes they share. The first assumption does not mean that all pairs of twins are assumed to share their environments equally, just that the extent of sharing does not depend on zygosity, and that on average MZ pairs do not share their environment more (or less) than DZ pairs.

The trait often depends on measured variables such as age and sex, and in such cases should be adjusted for these before correlations are estimated. This usually requires fitting statistical models which estimate the effects of the measured variables on the trait of interest and the correlation of the residuals in MZ and DZ pairs simultaneously.

If correlations for MZ and DZ pairs are statistically significantly different, additional models which divide the residual variation into components due to shared genetic effects, shared environmental effects, and unshared effects can be fitted. These “variance components models” can be fitted using maximum likelihood estimation [247] or via a structural equations approach [248], and should also incorporate adjustments for measured variables. The residual variance (which includes variation due to measurement error and stochastic factors as well as factors unique to individual twins), and the amounts of variation attributable to shared genes and shared environmental effects can also provide information, as can opposite-sex fraternal pairs, for example, about whether different genes or environmental effects are influencing variation of a trait in males and females. Often, the components of variance are reported as proportions of the total variance, and in this case the

proportion of shared genetic effects is referred to as the *heritability*. Although it is commonly reported, focusing on the heritability loses a lot of information and is not recommended [249,250].

Correlations are usually estimated for continuous traits such as height, but it is also often of interest to estimate how similar twins are for a categorical outcome, such as a diagnosis (or not) of schizophrenia or of metastatic prostate cancer, or severity of baldness (none, mild, moderate, or severe). This similarity is usually quantified using the *concordance* (either casewise or pairwise [251]). In the case of schizophrenia, the pairwise concordance for MZ twins is 0.48, meaning that if one member of a pair receives a diagnosis there is close to a 50% chance that the other one will too [252]. For DZ twins the concordance is 0.17. If we assume again that the two twin types share their environments equally and only differ in the proportion of genes they share, this is again consistent with genetic influences on liability of schizophrenia, but factors in the environment matter too; otherwise the concordance for MZ twins would be 1.00. Variance components models can also be fitted for binary traits, and these estimate the proportions of variation on the liability scale, which makes interpretation of results more challenging. These models for binary outcomes also have lower power. Alternatives exist, including estimation of “intrinsic correlation” after adjustment for measured variables such as age, which can affect concordances [253,254], and generalized linear mixed models (GLMMs) fitted using Markov chain Monte Carlo (MCMC) methods [255]. Methods for outcomes which are best represented as ordinal, categorical [256,257], or as censored survival times [258,259] also exist.

Advantages of the classic twin design include the ability to include all types of twins regardless of their measured outcomes or exposures, the ability to estimate variation and covariation as well as correlations and heritability, and the capacity to assess effects of measured covariates, such as age and sex, on both the trait and the variances/covariances. This approach also has limitations, perhaps most importantly, regarding the equal environments assumption, which is crucial to the model and yet difficult to test. The model also has low power to detect shared environmental effects (see, e.g., [246]) and assumes that any differences in correlations are due solely to differences in genetics.

14.11.2 Other Types of Twin Designs

Studies with twins extend beyond the classic twin design and are very useful in epidemiology generally, not just in genetic epidemiology. Some examples of other designs and of the types of questions that twin research can address are included in Table 14.3. Most designs require statistical models which allow for correlated observations to be fitted, such as GLMMs, also known as mixed effects models, or models fitted using generalized estimating equations (GEEs). These models also have advantages and disadvantages. For example, the GEE approach is appropriate when interest focuses on the association of measured covariates with the outcome, while GLMMs are required if estimation of variation and covariation is also of interest, but both these models are appropriate for continuous and binary outcomes and both can be fitted in standard statistical software packages such as Stata [274]. The most appropriate approach and model will depend on the specific research question and available data.

Each of these study types also has advantages and disadvantages. In outcome-discordant twin studies, the case-control pairs are matched for both measured and unmeasured factors—for age, genetic factors (perfectly for MZ pairs; 50% for DZ), nongenetic familial factors (not necessarily to the same degree for MZ and DZ pairs), mother, father, uterus and, perhaps, placenta, sex (if only same-sex pairs are included), calendar year, and season of birth. Although this type of study can be less costly and time-consuming compared with cohort studies, it has similar limitations to standard case-control studies (potential recall bias, inefficient for rare exposures). In contrast, exposure-discordant twin studies also match for both unmeasured and measured factors (other than the exposure of interest), and have the potential to allow causal inferences, but may not be representative of the general twin population. This study type has similar advantages and limitations to standard matched cohort studies, namely that it is advantageous for rare exposures but that it can be difficult to find and recruit exposure-discordant twin pairs. For example, over 2000 pairs of female twins were screened to find 20 pairs discordant by 20 or more pack-years of smoking [266]. If twins are included in an intervention study, they will be matched for genes, their participation may be enhanced by the pairing, and they may be a highly motivated group, but the close bond between twins may mean a potential failure to adhere to study protocol (e.g., by discussing or swapping treatments).

TABLE 14.3 Designs and Statistical Approaches in Twin Research

Design	Examples/Applications	Statistical Model/Analytic Approach
Co-twin control study—Disease discordant: matched case-control design, generally with identical twins discordant for a trait	Earlier onset of puberty is not associated with breast cancer risk in disease-discordant twins [260]	Conditional logistic regression or binary GLMMs/GEEs
Co-twin control study—Exposure-discordant: identical twins discordant for a health-related characteristic	Military service in Vietnam is associated with higher risk of post-traumatic stress symptoms than military service elsewhere [261]	Linear or logistic regression with adjustment for within-pair correlation (e.g., GEEs or mixed effects models)
Randomized controlled trial: identical twins, naturally matched for age, sex, and genes, given the same or different interventions (or possibly both, if a crossover design is used)	Calcium supplementation in adolescence has little effect on bone density [262]	As for exposure-discordant studies
Epigenetics: MZ twins share the same DNA but the way in which this DNA operates can differ	Birth weight is associated with epigenetic differences in growth and metabolism genes [210]	As for exposure-discordant studies
Causal models in twins	The observed association between BMI and mortality is unlikely to be causal and appears largely due to shared confounding [263,264]	As for exposure-discordant studies
Within-between study (related to causal models)	The association between birth weight and cord blood erythropoietin appears due to individual rather than pair-specific factors [265]	Linear or logistic regression with adjustment for within-pair correlation (e.g., GEEs or mixed effects models)
Differences study: Both exposure and outcome are continuous	Differences in smoking consumption predict differences in bone density [266]	Linear regression
Issues specific to twins: Raising twins can be challenging as well as rewarding, and research can help parents with day-to-day decisions	For most twin pairs, staying together in the first year of school is a critical social support [267]	Regression-based approach; depends on research question and outcome type
Longitudinal designs: Following twins over growth and development can identify changing patterns of genetic and environmental influence	Higher cumulative exposure to solvents over the lifetime is associated with increased risk of Parkinson's disease [268]. Many twin registries have longitudinal data on twins [269–271]	Regression-based approach accounting for correlation (e.g., mixed effects models)
Multivariate designs: Studying two or more traits simultaneously can identify shared and separate genetic and environmental influences	Genes that influence children's reading abilities are in part shared with genes that influence mathematics abilities [272]	Multivariate variance components models
Extended twin family study: Inclusion of other family members, such as parents and siblings, extends the range of genetic and environmental factors that can be understood	Different patterns of covariation are apparent for height, BMI, and blood pressure [273]	Variance components models

14.11.3 General Statistical Issues

General statistical principles apply when analyzing data from twins and families. Good statistical practice in this context includes thorough exploration of data prior to model fitting, being aware of and testing model assumptions, reporting estimates, 95% confidence intervals and *P*-values, starting with simple analyses and models, and building on these, adjusting for measured variables before considering unmeasured effects, and remembering that analyses of continuous outcomes are usually more powerful than those of binary outcomes.

14.11.4 Summary

While the classic twin design has been applied for around 100 years as a powerful tool for disentangling the etiology of complex traits and diseases, researchers now increasingly recognize the value of including twins in many other types of studies, particularly in those related to health and medical research. Technological advances such as epigenetic, microbiome, and other “omics” platforms have resulted in new opportunities and possibilities for studies and study designs [7], and new statistical models and approaches (such as causal inference utilizing data from twins) being developed, suggesting that twin studies will continue to remain highly relevant in medical research.

14.12 TWIN REGISTRIES AND INTERNATIONAL COLLABORATION

The first twin registries were set up in Scandinavia around the middle of the 20th century, aimed at collecting and maintaining contact information from twins who were identified through national birth registers, and accessing their longitudinal health data through links with medical data sources to conduct twin studies. The value of such resources has been since recognized based on a number of other twin registries established in developed and developing countries throughout the years [275]. Their relevance expands beyond the collection and analysis of twin data and biospecimens toward further development of the twin methodology and novel applications in health and social sciences.

While some twin registries have relied upon population-based recruiting strategies such as identifying twins from national birth registers [276,277], drivers' licensing departments [278], regional health systems [279],

and war veterans' databases [269], others have opted for a volunteer-based approach, using traditional and social media channels [280] and twin festivals to improve recruitment and engagement with members [270]. In all twin registries, zygosity determination strategies have also varied from questionnaire self-reports to DNA testing in twin registries where there is a large amount of genotypes collected [281].

Global collaboration in twin research has also been a task facilitated by twin registries, which have formed the International Network of Twin Registries [282] with that purpose. The network aims to build a worldwide catalogue of available twin data and biospecimens to be used in ethically approved future studies, and it is formally linked to the International Society of Twin Studies (ISTS), the main professional society in twin research, which is also responsible for the *Twin Research and Human Genetics* journal. A good example of what can be achieved in multicenter twin studies is the CODATwins project, a consortium with mostly anthropometrical data on 434,723 twins in 22 different countries [283]. Such collaborations will be especially relevant for achieving the necessary sample sizes to find discordant twin pairs for specific traits and conditions and to study rare diseases, as they also provide means for expert knowledge generation and exchange.

14.13 CONCLUSIONS

In the field of twins and twinning, there is still a great deal to be learned. The development of new DNA molecular and cytogenetic techniques, the use of prenatal diagnosis such as chorionic villus sampling, amniocentesis, and ultrasound examination in humans, as well as embryo pathology, histology, and genetic advances all give clues to the increased understanding of MZ and DZ twins and the twinning process itself.

REFERENCES

- [1] Galton F. The history of twins, as a criterion of the relative powers of nature and nurture. Trubner and Co.; 1876. p. 391.
- [2] Fisher RA. Has Mendel's work been rediscovered? *Ann Sci* 1936;1:115.
- [3] Hall J. Twinning. *Lancet* 2003a;362:735–43.
- [4] Mayo O. Early research on human genetics using the twin method: who really invented the method? *Twin Res Hum Genet* 2009;12:237–45.

- [5] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 1918;52:399–433.
- [6] McNamara HC, Kane SC, Craig JM, Short RV, Umstad MP. A review of the mechanisms and evidence for typical and atypical twinning. *Am J Obstet Gynecol* 2016;214:172–91.
- [7] van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. *Nat Rev Genet* 2012;13:640–53.
- [8] Herranz G. The timing of monozygotic twinning: a criticism of the common model. *Zygote* 2013;1–14.
- [9] Denker HW. Comment on G. Herranz: the timing of monozygotic twinning: a criticism of the common model. *Zygote* (2013). *Zygote* 2013;1–3.
- [10] Ginsberg NA, Ginsberg S, Rechitsky S, Verlinsky Y. Fusion as the etiology of chimerism in monochorionic dizygotic twins. *Fetal Diagn Ther* 2005;20:20–2.
- [11] Assaf SA, Randolph LM, Benirschke K, Wu S, Samadi R, Chmait RH. Discordant blood chimerism in dizygotic monochorionic laser-treated twin–twin transfusion syndrome. *Obstet Gynecol* 2010;116:483–5.
- [12] Chen K, Chmait RH, Vanderbilt D, Wu S, Randolph L. Chimerism in monochorionic dizygotic twins: case study and review. *Am J Med Genet* 2013;161A:1817–24.
- [13] Fumoto S, Hosoi K, Ohnishi H, Hoshina H, Yan K, Saji H, Oka A. Chimerism of buccal membrane cells in a monochorionic dizygotic twin. *Pediatrics* 2014;133:e1097–1100.
- [14] Umstad MP, Short RV, Wilson M, Craig JM. Chimaeric twins: why monochorionicity does not guarantee monozygosity. *Aust N Z J Obstet Gynaecol* 2012;52:305–7.
- [15] Miura K, Niikawa N. Do monochorionic dizygotic twins increase after pregnancy by assisted reproductive technology? *J Hum Genet* 2005;50:1–6.
- [16] Nylander PP, Osunkoya BO. Unusual monochorionic placentation with heterosexual twins. *Obstet Gynecol* 1970;36:621–5.
- [17] Tarkowski AK, Wojewodzka M. A method for obtaining chimaeric mouse blastocysts with two separate inner cell masses: a preliminary report. *J Embryol Exp Morphol* 1982;71:215–21.
- [18] Williams CA, Wallace MR, Drury KC, Kipersztok S, Edwards RK, Williams RS, Haller MJ, Schatz DA, Silverstein JH, Gray BA, Zori RT. Blood lymphocyte chimerism associated with IVF and monochorionic dizygous twinning: case report. *Hum Reprod* 2004;19:2816–21.
- [19] Safran A, Reubinoff BE, Porat Katz A, Werner M, Friedler S, Lewin A. Intracytoplasmic sperm injection allows fertilization and development of a chromosomally balanced embryo from a binovular zona pellucida. *Hum Reprod* 1998;13:2575–8.
- [20] Van de Leur SJ, Zeilmaker GH. Double fertilization in vitro and the origin of human chimerism. *Fertil Steril* 1990;54:539–40.
- [21] Vicdan K, Işık AZ, Dagli HG, Kaba A, Kişnişçi H. Fertilization and development of a blastocyst-stage embryo after selective intracytoplasmic sperm injection of a mature oocyte from a binovular zona pellucida: a case report. *J Assist Reprod Genet* 1999;16:355–7.
- [22] Walker SP, Meagher S, White SM. Confined blood chimerism in monochorionic dizygous (MCDZ) twins. *Prenat Diagn* 2007;27:369–72.
- [23] Choi DH, Kwon H, Lee SD, Moon MJ, Yoo EG, Lee KH, Hong YK, Kim G. Testicular hypoplasia in monochorionic dizygous twin with confined blood chimerism. *J Assist Reprod Genet* 2013;30:1487–91.
- [24] Short RV. The bovine freemartin: a new look at an old problem. *Philos Trans R Soc Lond B Biol Sci* 1970;259:141–7.
- [25] Dirani M, Chamberlain M, Garoufalos P, Chen CY, Guymer RH, Baird PN. Mirror-image congenital esotropia in monozygotic twins. *J Pediatr Ophthalmol Strabismus* 2006;43:170–1.
- [26] Goto T, Nemoto T, Okuma T, Kobayashi H, Funata N. Mirror-image solitary bone cyst of the humerus in a pair of mirror-image monozygotic twins. *Arch Orthop Trauma Surg* 2008;128:1403–6.
- [27] Hu JT, Liu T, Qian J, Zhang YB, Zhou X, Zhang QG. Occurrence of different external ear deformities in monozygotic twins: report of 2 cases. *Plast Reconstr Surg Glob Open* 2014;2:e206.
- [28] Karaca C, Yilmaz M, Karatas O, Menderes A, Karademir S. Mirror imaging cleft lip in monozygotic twins. *Eur J Plast Surg* 1995;18:260–1.
- [29] Morison D, Reyes CV, Skorodin MS. Mirror-image tumors in mirror-image twins. *Chest* 1994;106:608–10.
- [30] Novak RW. Laryngotracheoesophageal cleft and unilateral pulmonary hypoplasia in twins. *Pediatrics* 1981;67:732–4.
- [31] Riess A, Dufke A, Riess O, Beck Woedl S, Fode B, Skladny H, Klaes R, Tzschach A. Mirror-image asymmetry in monozygotic twins with kabuki syndrome. *Mol Syndromol* 2012;3:94–7.
- [32] Rife DC. Genetic studies of monozygotic twins: III. Mirror-imaging. *J Hered* 1933;24:443–6.
- [33] Satoh K, Shibata Y, Tokushige H, Onizuka T. A mirror image of the first and second branchial arch syndrome associated with cleft lip and palate in monozygotic twins. *Br J Plast Surg* 1995;48:601–5.

- [34] Sperber GH, Machin GA, Bamforth FJ. Mirror-image dental fusion and discordance in monozygotic twins. *Am J Med Genet* 1994;51:41–5.
- [35] Springer SP, Searleman A. Laterality in twins: the relationship between handedness and hemispheric asymmetry for speech. *Behav Genet* 1978;8:349–57.
- [36] Wang ED, Xu X, Dagum AB. Mirror-image trigger thumb in dichorionic identical twins. *Orthopedics* 2012;35:e981–3.
- [37] Helland CA, Wester K. Monozygotic twins with mirror image cysts: indication of a genetic mechanism in arachnoid cysts? *Neurology* 2007;69:110–1.
- [38] Nigro MA, Wishnow R, Maher L. Colpocephaly in identical twins. *Brain Dev* 1991;13:187–9.
- [39] Pascual-Castroviejo I, Verdú A, Román M, De la Cruz-Medina M, Villarejo F. Optic glioma with progressive occlusion of the aqueduct of Sylvius in monozygotic twins with neurofibromatosis. *Brain Dev* 1988;10:24–9.
- [40] Zhou JY, Pu JL, Chen S, Hong Y, Ling CH, Zhang JM. Mirror-image arachnoid cysts in a pair of monozygotic twins: a case report and review of the literature. *Int J Med Sci* 2011;8:402–5.
- [41] Derom C, Thiery E, Vlietinck R, Loos R, Derom R. Handedness in twins according to zygosity and chorion type: a preliminary report. *Behav Genet* 1996;26:407–8.
- [42] Sommer IE, Ramsey NF, Bouma A, Kahn RS. Cerebral mirror-imaging in a monozygotic twin. *Lancet* 1999;354:1445–6.
- [43] Schmerler S, Wessel GM. Polar bodies—more a lack of understanding than a lack of respect. *Mol Reprod Dev* 2011;78:3–8.
- [44] Bieber FR, Nance WE, Morton CC, Brown JA, Redwine FO, Jordan RL, Mohanakumar T. Genetic studies of an acardiac monster: evidence of polar body twinning in man. *Science* 1981;213:775–7.
- [45] Fisk NM, Ware M, Stanier P, Moore G, Bennett P. Molecular genetic etiology of twin reversed arterial perfusion sequence. *Am J Obstet Gynecol* 1996;174:891–4.
- [46] La Sala GB, Villani MT, Nicoli A, Gallinelli A, Nucera G, Blickstein I. Effect of the mode of assisted reproductive technology conception on obstetric outcomes for survivors of the vanishing twin syndrome. *Fertil Steril* 2006;86:247–9.
- [47] Pinborg A, Lidegaard O, la Cour Freiesleben NI, Andersen AN. Consequences of vanishing twins in IVF/ICSI pregnancies. *Hum Reprod* 2005;20:2821–9.
- [48] Rodríguez-González M, Serra V, García-Velasco JA, Pellicer A, Remohí J. The ‘vanishing embryo’ phenomenon in an oocyte donation programme. *Hum Reprod* 2002;17:798–802.
- [49] Pinborg A, Lidegaard O, la Cour Freiesleben NI, Andersen AN. Vanishing twins: a predictor of small-for-gestational age in IVF singletons. *Hum Reprod* 2007;22:2707–14.
- [50] Pharoah PO, Cooke RW. A hypothesis for the aetiology of spastic cerebral palsy - the vanishing twin. *Dev Med Child Neurol* 1997;39:292–6.
- [51] Newton R, Casabonne D, Johnson A, Pharoah P. A case-control study of vanishing twin as a risk factor for cerebral palsy. *Twin Res* 2003;6:83–4.
- [52] Huang T, Boucher K, Aul R, Rashid S, Meschino WS. First and second trimester maternal serum markers in pregnancies with a vanishing twin. *Prenat Diagn* 2015;35:90–6.
- [53] Blickstein I. Superfecundation and superfetation: lessons from the past on early human development. *J Matern Fetal Neonatal Med* 2003;14:217–9.
- [54] Czyz W, Morahan JM, Ebers GC, Ramagopalan SV. Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. *BMC Med* 2012;10:93.
- [55] Amsalem H, Tsvieli R, Zentner BS, Yagel S, Mitrani-Rosenbaum S, Hurwitz A. Monopaternal superfecundation of quintuplets after transfer of two embryos in an in vitro fertilization cycle. *Fertil Steril* 2001;76:621–3.
- [56] Peigné M, Andrieux J, Deruelle P, Vuillaume I, Leroy M. Quintuplets after a transfer of two embryos following in vitro fertilization: a proved superfecundation. *Fertil Steril* 2011;95(2124):e2113–24. e2116.
- [57] Girela E, Lorente JA, Alvarez JC, Rodrigo MD, Lorente M, Villanueva E. Indisputable double paternity in dizygous twins. *Fertil Steril* 1997;67:1159–61.
- [58] Harris DW. Superfecundation. *J Reprod Med* 1982;27:39.
- [59] Terasaki PI, Gjertson D, Bernoco D, Perdue S, Mickey MR, Bond J. Twins with two different fathers identified by HLA. *N Engl J Med* 1978;299:590–2.
- [60] Wenk RE, Houtz T, Chiafari F, Brooks M. Superfecundation identified by HLA, protein, and VNTR DNA polymorphisms. *Transfus Med* 1991;1:253–5.
- [61] James WH. The incidence of superfecundation and of double paternity in the general population. *Acta Genet Med Gemellol* 1993;42:257–62.
- [62] Bristow RE, Shumway JB, Khouzami AN, Witter FR. Complete hydatidiform mole and surviving coexistent twin. *Obstet Gynecol Surv* 1996;51:705–9.
- [63] Fishman DA, Padilla LA, Keh P, Cohen L, Frederiksen M, Lurain JR. Management of twin pregnancies consisting of a complete hydatidiform mole and normal fetus. *Obstet Gynecol* 1998;91:546–50.

- [64] Massardier J, Golfier F, Journet D, Frappart L, Zalaquett M, Schott A, Lenoir V, Dupuis O, Hajri T, Raudrant D. Twin pregnancy with complete hydatidiform mole and coexistent fetus: obstetrical and oncological outcomes in a series of 14 cases. *Eur J Obstet Gynecol Reprod Biol* 2009;143:84–7.
- [65] Sebire NJ, Foskett M, Paradinas FJ, Fisher RA, Francis RJ, Short D, Newlands ES, Seckl MJ. Outcome of twin pregnancies with complete hydatidiform mole and healthy co-twin. *Lancet* 2002;359:2165–6.
- [66] Niemann I, Sunde L, Petersen LK. Evaluation of the risk of persistent trophoblastic disease after twin pregnancy with diploid hydatidiform mole and coexisting normal fetus. *Am J Obstet Gynecol* 2007;197:45.e41–5.
- [67] George V, Khanna M, Dutta T. Fetus in fetu. *J Pediatr Surg* 1983;18:288–9.
- [68] Spencer R. Parasitic conjoined twins: external, internal (fetuses in fetu and teratomas), and detached (acardi-acs). *Clin Anat* 2001;14:428–44.
- [69] Brand A, Alves MC, Saraiva C, Loio P, Goulão J, Malta J, Palminha JM, Martins M. Fetus in fetu—diagnostic criteria and differential diagnosis—a case report and literature review. *J Pediatr Surg* 2004;39:616–8.
- [70] Gerber RE, Kamaya A, Miller SS, Madan A, Cronin DM, Dwyer B, Chueh J, Conner KE, Barth RA. Fetus in fetu: 11 fetoid forms in a single fetus: review of the literature and imaging. *J Ultrasound Med* 2008;27:1381–7.
- [71] Hoeffel CC, Nguyen KQ, Phan HT, Truong NH, Nguyen TS, Tran TT, Fornes P. Fetus in fetu: a case report and literature review. *Pediatrics* 2000;105:1335–44.
- [72] Huddle LN, Fuller C, Powell T, Hiemenga JA, Yan J, Deuell B, Lyders EM, Bodurtha JN, Papenhausen PR, Jackson-Cook CK, Pandya A, Jaworski M, Tye GW, Ritter AM. Intraventricular twin fetuses in fetu. *J Neurol Surg Pediatr* 2012;9:17–23.
- [73] Escobar MA, Rossman JE, Caty MG. Fetus-in-fetu: report of a case and a review of the literature. *J Pediatr Surg* 2008;43:943–6.
- [74] Hopkins KL, Dickson PK, Ball TI, Ricketts RR, O'Shea PA, Abramowsky CR. Fetus-in-fetu with malignant recurrence. *J Pediatr Surg* 1997;32:1476–9.
- [75] Bressers WM, Eriksson AW, Kostense PJ, Parisi P. Increasing trend in the monozygotic twinning rate. *Acta Genet Med Gemellol* 1987;36:397–408.
- [76] Nylander PP. The twinning incidence of Nigeria. *Acta Genet Med Gemellol* 1979;28:261–3.
- [77] Imaizumi Y. Triplets and higher order multiple births in Japan. *Acta Genet Med Gemellol* 1990;39:295–306.
- [78] Campbell DM, Campbell AJ, MacGillivray I. Maternal characteristics of women having twin pregnancies. *J Biosoc Sci* 1974;6:463–70.
- [79] Hoekstra C, Zhao ZZ, Lambalk CB, Willemsen G, Martin NG, Boomsma DI, Montgomery GW. Dizygotic twinning. *Hum Reprod Update* 2008;14:37–47.
- [80] Nylander PP. Biosocial aspects of multiple births. *J Biosoc Sci Suppl* 1971:29–38.
- [81] Umstad MP, Hale L, Wang YA, Sullivan EA. Multiple deliveries: the reduced impact of in vitro fertilisation in Australia. *Aust N Z J Obstet Gynaecol* 2013;53:158–64.
- [82] James WH. Sex ratio in twin births. *Ann Hum Biol* 1975;2:365–78.
- [83] James WH. Sex ratio and placentation in twins. *Ann Hum Biol* 1980b;7:273–6.
- [84] James WH. Gestational age in twins. *Arch Dis Child* 1980a;55:281–4.
- [85] Tsunoda Y, Tokunaga T, Sugie T, Katsumata M. Production of monozygotic twins following the transfer of bisected embryos in the goats. *Theriogenology* 1985;24:337–43.
- [86] Bryan E, Little J, Burn J. Congenital anomalies in twins. *Baillieres Clin Obstet Gynaecol* 1987;1:697–721.
- [87] Mastroiacovo P, Castilla EE, Arpino C, Botting B, Cocchi G, Goujard J, Marinacci C, Merlob P, Metneki J, Mutchinick O, Ritvanen A, Rosano A. Congenital malformations in twins: an international study. *Am J Med Genet* 1999;83:117–24.
- [88] Doyle PE, Beral V, Botting B, Wale CJ. Congenital malformations in twins in England and Wales. *J Epidemiol Community Health* 1991;45:43–8.
- [89] Kallen B. Congenital malformations in twins: a population study. *Acta Genet Med Gemellol* 1986;35:167–78.
- [90] Schinzel AA, Smith DW, Miller JR. Monozygotic twinning and structural defects. *J Pediatr* 1979;95:921–30.
- [91] Derom C, Derom R, Loos RJ, Jacobs N, Vlietinck R. Retrospective determination of chorion type in twins using a simple questionnaire. *Twin Res* 2003;6:19–21.
- [92] Burn J, Corney G. Zygosity determination and the types of twinning. In: MacGillivray I, Campbell DM, Thompson B, editors. *Twinning and twins*. Chichester: John Wiley & Sons; 1988.
- [93] Forget-Dubois N, Perusse D, Turecki G, Girard A, Billette JM, Rouleau G, Boivin M, Malo J, Tremblay RE. Diagnosing zygosity in infant twins: physical similarity, genotyping, and chorionicity. *Twin Res* 2003;6:479–85.
- [94] Weinberg W. Contribution on the physiology and pathology of multiple birth in man. *Pflügers Arch Physiol* 1901;88:346.
- [95] Dallapiccola B, Stomeo C, Ferranti G, Di Lecce A, Purpura M. Discordant sex in one of three monozygotic triplets. *J Med Genet* 1985;22:6–11.
- [96] Edwards JH, Dent T, Kahn J. Monozygotic twins of different sex. *J Med Genet* 1966;3:117–23.

- [97] Kurosawa K, Kuromaru R, Imaizumi K, Nakamura Y, Ishikawa F, Ueda K, Kuroki Y. Monozygotic twins with discordant sex. *Acta Genet Med Gemellol* 1992;41:301–10.
- [98] Akane A, Matsubara K, Shiono H, Yamada M, Nakagome Y. Diagnosis of twin zygosity by hypervariable RFLP markers. *Am J Med Genet* 1991;41:96–8.
- [99] Derom C, Bakker E, Vlietinck R, Derom R, Van den Berghe H, Thiery M, Pearson P. Zygosity determination in newborn twins using DNA variants. *J Med Genet* 1985;22:279–82.
- [100] Hill AV, Jeffreys AJ. Use of minisatellite DNA probes for determination of twin zygosity at birth. *Lancet* 1985;2:1394–5.
- [101] Becker A, Busjahn A, Faulhaber HD, Bahring S, Robertson J, Schuster H, Luft FC. Twin zygosity. Automated determination with microsatellites. *J Reprod Med* 1997;42:260–6.
- [102] Hannelius U, Gherman L, Makela VV, Lindstedt A, Zucchelli M, Lagerberg C, Tybring G, Kere J, Lindgren CM. Large-scale zygosity testing using single nucleotide polymorphisms. *Twin Res Hum Genet* 2007;10:604–25.
- [103] Bianchi DW, Fisk NM. Fetomaternal cell trafficking and the stem cell debate: gender matters. *J Am Med Assoc* 2007;297:1489–91.
- [104] Erlich Y. Blood ties: chimerism can mask twin discordance in high-throughput sequencing. *Twin Res Hum Genet* 2011;14:137–43.
- [105] Bajoria R, Kingdom J. The case for routine determination of chorionicity and zygosity in multiple pregnancy. *Prenat Diagn* 1997;17:1207–25.
- [106] Craig JM, Segal NL, Umstad MP, Cutler TL, Keogh LA, Hopper JL, Rankin M, Denton J, Derom CA, Sumathipala A, Harris JR, International Society for Twin S, International Council of Multiple Birth O. Zygosity testing should be encouraged for all same-sex twins: FOR: a genetic test is essential to determine zygosity. *BJOG* 2015;122:1641.
- [107] Derom R, Vlietinck RF, Derom C, Keith LG, Van Den Berghe H. Zygosity determination at birth: a plea to the obstetrician. *J Perinat Med* 1991;19(Suppl 1):234–40.
- [108] Machin GA. Why is it important to diagnose chorionicity and how do we do it? *Best Pract Res Clin Obstet Gynaecol* 2004;18:515–30.
- [109] Machin G. Non-identical monozygotic twins, intermediate twin types, zygosity testing, and the non-random nature of monozygotic twinning: a review. *Am J Med Genet C Semin Med Genet* 2009b;151C:110–27.
- [110] Segal NL. Zygosity testing: laboratory and the investigator's judgment. *Acta Genet Med Gemellol* 1984;33:515–21.
- [111] Maruotti GM, Saccone G, Morlando M, Martinelli P. First-trimester ultrasound determination of chorionicity in twin gestations using the lambda sign: a systematic review and meta-analysis. *Eur J Obstet Gynecol Reprod Biol* 2016;202:66–70.
- [112] Cutler TL, Murphy K, Hopper JL, Keogh LA, Dai Y, Craig JM. Why accurate knowledge of zygosity is important to twins. *Twin Res Hum Genet* 2015;18:298–305.
- [113] Bamforth F, Machin G. Why zygosity of multiple births is not always obvious: an examination of zygosity testing requests from twins or their parents. *Twin Res* 2004;7:406–11.
- [114] van Jaarsveld CH, Llewellyn CH, Fildes A, Fisher A, Wardle J. Are my twins identical: parents may be misinformed by prenatal scan observations. *BJOG* 2012;119:517–8.
- [115] Craig JM, All A. Re: zygosity testing should be encouraged for all same-sex twins. AGAINST: the benefit of this knowledge should be weighed against the potential pitfalls. *BJOG* 2016;123:1560–1.
- [116] Keith L, Machin G. Zygosity testing. Current status and evolving issues. *J Reprod Med* 1997;42:699–707.
- [117] Cyranoski D. Developmental biology: two by two. *Nature* 2009;458:826–9.
- [118] Harvey MA, Huntley RM, Smith DW. Familial monozygotic twinning. *J Pediatr* 1977;90:246–7.
- [119] Machin G. Familial monozygotic twinning: a report of seven pedigrees. *Am J Med Genet C Semin Med Genet* 2009a;151C:152–4.
- [120] Shapiro LR, Zemek L, Shulman MJ. Genetic etiology for monozygotic twinning. *Birth Defects Orig Artic Ser* 1978;14:219–22.
- [121] St Clair JB, Golubovsky MD. Paternally derived twinning: a two century examination of records of one Scottish name. *Twin Res* 2002;5:294–307.
- [122] Michels VV, Riccardi VM. Twin recurrence and amniocentesis: male and MZ heritability factors. *Birth Defects Orig Artic Ser* 1978;14:201–11.
- [123] Lichtenstein P, Kallen B, Koster M. No paternal effect on monozygotic twinning in the Swedish Twin Registry. *Twin Res* 1998;1:212–5.
- [124] Torlopp A, Khan MA, Oliveira NM, Leek I, Soto-Jimenez LM, Sosinsky A, Stern CD. The transcription factor Pitx2 positions the embryonic axis and regulates twinning. *Elife* 2014;3:e03743.
- [125] Parisi P, Gatti M, Prinzi G, Caperna G. Familial incidence of twinning. *Nature* 1983;304:626–8.
- [126] Meulemans WJ, Lewis CM, Boomsma DI, Derom CA, Van den Berghe H, Orlebeke JF, Vlietinck RF, Derom RM. Genetic modelling of dizygotic twinning in pedigrees of spontaneous dizygotic twins. *Am J Med Genet* 1996;61:258–63.

- [127] Lewis CM, Healey SC, Martin NG. Genetic contribution to DZ twinning. *Am J Med Genet* 1996;61:237–46.
- [128] Samra JS, Hampton N, Fitzgibbon MN, Obhrai MS. The second twin. *Lancet* 1990;336:883.
- [129] Healey SC, Duffy DL, Martin NG, Turner G. Is fragile X syndrome a risk factor for dizygotic twinning? *Am J Med Genet* 1997;72:245–6.
- [130] Montgomery GW, McNatty KP, Davis GH. Physiology and molecular genetics of mutations that increase ovulation rate in sheep. *Endocr Rev* 1992;13:309–28.
- [131] Galloway SM, McNatty KP, Cambridge LM, Laitinen MP, Juengel JL, Jokiranta TS, McLaren RJ, Luiro K, Dodds KG, Montgomery GW, Beattie AE, Davis GH, Ritvos O. Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nat Genet* 2000;25:279–83.
- [132] Hanrahan JP, Gregan SM, Mulsant P, Mullen M, Davis GH, Powell R, Galloway SM. Mutations in the genes for oocyte-derived growth factors GDF9 and BMP15 are associated with both increased ovulation rate and sterility in Cambridge and Belclare sheep (*Ovis aries*). *Biol Reprod* 2004;70:900–9.
- [133] Montgomery GW, Zhao ZZ, Marsh AJ, Mayne R, Treloar SA, James M, Martin NG, Boomsma DI, Duffy DL. A deletion mutation in GDF9 in sisters with spontaneous DZ twins. *Twin Res* 2004;7:548–55.
- [134] Palmer JS, Zhao ZZ, Hoekstra C, Hayward NK, Webb PM, Whiteman DC, Martin NG, Boomsma DI, Duffy DL, Montgomery GW. Novel variants in growth differentiation factor 9 in mothers of dizygotic twins. *J Clin Endocrinol Metab* 2006;91:4713–6.
- [135] Zhao ZZ, Painter JN, Palmer JS, Webb PM, Hayward NK, Whiteman DC, Boomsma DI, Martin NG, Duffy DL, Montgomery GW. Variation in bone morphogenetic protein 15 is not associated with spontaneous human dizygotic twinning. *Hum Reprod* 2008;23:2372–9.
- [136] Luong HT, Chaplin J, McRae AF, Medland SE, Willemsen G, Nyholt DR, Henders AK, Hoekstra C, Duffy DL, Martin NG, Boomsma DI, Montgomery GW, Painter JN. Variation in BMP1B, TGFRB1 and BMP2 and control of dizygotic twinning. *Twin Res Hum Genet* 2011;14:408–16.
- [137] Dixit H, Rao L, Padmalatha V, Raseswari T, Kapu AK, Panda B, Murthy K, Tosh D, Nallari P, Deenadayal M, Gupta N, Chakrabarty B, Singh L. Genes governing premature ovarian failure. *Reprod Biomed Online* 2010;20:724–40.
- [138] Al-Hendy A, Moshynska O, Saxena A, Feyles V. Association between mutations of the follicle-stimulating-hormone receptor and repeated twinning. *Lancet* 2000;356:914.
- [139] Painter JN, Willemsen G, Nyholt D, Hoekstra C, Duffy DL, Henders AK, Wallace L, Healey S, Cannon-Albright LA, Skolnick M, Martin NG, Boomsma DI, Montgomery GW. A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins. *Hum Reprod* 2010;25:1569–80.
- [140] Montgomery GW, Duffy DL, Hall J, Kudo M, Martin NG, Hsueh AJ. Mutations in the follicle-stimulating hormone receptor and familial dizygotic twinning. *Lancet* 2001;357:773–4.
- [141] Mbarek H, Steinberg S, Nyholt DR, Gordon SD, Miller MB, McRae AF, Hottenga JJ, Day FR, Willemsen G, de Geus EJ, Davies GE, Martin HC, Penninx BW, Jansen R, McAloney K, Vink JM, Kaprio J, Plomin R, Spector TD, Magnusson PK, Reversade B, Harris RA, Aagaard K, Kristjansson RP, Olafsson I, Eyjolfsson GI, Sigurdardottir O, Iacono WG, Lambalk CB, Montgomery GW, McGue M, Ong KK, Perry JR, Martin NG, Stefansson H, Stefansson K, Boomsma DI. Identification of common genetic variants influencing spontaneous dizygotic twinning and female fertility. *Am J Hum Genet* 2016;98:898–908.
- [142] Boomsma DI, Frants RR, Bank RA, Martin NG. Protease inhibitor (Pi) locus, fertility and twinning. *Hum Genet* 1992;89:329–32.
- [143] Lieberman J, Borhani NO, Feinleib M. Twinning as a heterozygous advantage for alpha1-antitrypsin deficiency. *Prog Clin Biol Res* 1978;24 Pt B:45–54.
- [144] Jauniaux E, Elkazen N, Leroy F, Wilkin P, Rodesch F, Hustin J. Clinical and morphologic aspects of the vanishing twin phenomenon. *Obstet Gynecol* 1988;72:577–81.
- [145] Fryns JP. The female and the fragile X. A study of 144 obligate female carriers. *Am J Med Genet* 1986;23:157–69.
- [146] Kenneson A, Warren ST. The female and the fragile X reviewed. *Semin Reprod Med* 2001;19:159–65.
- [147] Derom C, Jawaheer D, Chen WV, McBride KL, Xiao X, Amos C, Gregersen PK, Vlietinck R. Genome-wide linkage scan for spontaneous DZ twinning. *Eur J Hum Genet* 2006;14:117–22.
- [148] Duffy D, Montgomery G, Treloar S, Birley A, Kirk K, Boomsma D, Beem L, de Geus E, Slagboom E, Knighton J, Reed P, Martin N. IBD sharing around the PPARG locus is not increased in dizygotic twins or their mothers. *Nat Genet* 2001;28:315.
- [149] Busjahn A, Knoblauch H, Faulhaber HD, Aydin A, Uhlmann R, Tuomilehto J, Kaprio J, Jedrusik P, Januszewicz A, Strelau J, Schuster H, Luft FC, Muller-Myhsok B. A region on chromosome 3 is linked to dizygotic twinning. *Nat Genet* 2000;26:398–9.

- [150] Valleix S, Vinciguerra C, Lavergne JM, Leuer M, Delpech M, Negrier C. Skewed X-chromosome inactivation in monochorionic diamniotic twin sisters results in severe and mild hemophilia A. *Blood* 2002;100:3034–6.
- [151] Goodship J, Carter J, Burn J. X-inactivation patterns in monozygotic and dizygotic female twins. *Am J Med Genet* 1996;61:205–8.
- [152] Tan SS, Williams EA, Tam PP. X-chromosome inactivation occurs at different times in different tissues of the post-implantation mouse embryo. *Nat Genet* 1993;3:170–4.
- [153] Bamforth F, Machin G, Innes M. X-chromosome inactivation is mostly random in placental tissues of female monozygotic twins and triplets. *Am J Med Genet* 1996;61:209–15.
- [154] Hall JG. Twinning: mechanisms and genetic implications. *Curr Opin Genet Dev* 1996;6:343–7.
- [155] Shur N. The genetics of twinning: from splitting eggs to breaking paradigms. *Am J Med Genet C Semin Med Genet* 2009;151C:105–9.
- [156] Weksberg R, Shuman C, Caluseriu O, Smith AC, Fei YL, Nishikawa J, Stockley TL, Best L, Chitayat D, Olney A, Ives E, Schneider A, Bestor TH, Li M, Sadowski P, Squire J. Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum Mol Genet* 2002;11:1317–25.
- [157] Kaufman MH, O'Shea KS. Induction of monozygotic twinning in the mouse. *Nature* 1978;276:707–8.
- [158] Czeizel AE, Metneki J, Dudas I. Higher rate of multiple births after periconceptional vitamin supplementation. *N Engl J Med* 1994;330:1687–8.
- [159] Aston KI, Peterson CM, Carrell DT. Monozygotic twinning associated with assisted reproductive technologies: a review. *Reproduction* 2008;136:377–86.
- [160] Sills ES, Tucker MJ, Palermo GD. Assisted reproductive technologies and monozygous twins: implications for future study and clinical practice. *Twin Res* 2000;3:217–23.
- [161] Elizur SE, Levron J, Shrim A, Sivan E, Dor J, Shulman A. Monozygotic twinning is not associated with zona manipulation procedures but increases with high-order multiple pregnancies. *Fertil Steril* 2004;82:500–1.
- [162] Stockard CR. Developmental rate and structural expression: an experimental study of twins, 'double monsters' and single deformities, and the interaction among embryonic organs during their origin and development. *Am J Anat* 1921;28:115–277.
- [163] Storrs EE, Williams RJ. A study of monozygous quadruplet armadillos in relation to mammalian inheritance. *Proc Natl Acad Sci U S A* 1968;60:910–4.
- [164] Bulmer M. The biology of twinning in man. Oxford: Clarendon Press; 1970.
- [165] Yovich JL, Stanger JD, Grauaug A, Barter RA, Lunay G, Dawkins RL, Mulcahy MT. Monozygotic twins from in vitro fertilization. *Fertil Steril* 1984;41:833–7.
- [166] Edwards RG, Mettler L, Walters DE. Identical twins and in vitro fertilization. *J In Vitro Fert Embryo Transf* 1986;3:114–7.
- [167] Boklage CE. On the timing of monozygotic twinning events. In: Gedda L, Parisi P, Nance WE, editors. *Twin research*. New York: Alan R. Liss; 1981. p. 155–65.
- [168] Boklage CE. The organization of the oocyte and embryogenesis in twinning and fusion malformations. *Acta Genet Med Gemellol* 1987a;36:421–31.
- [169] Boklage CE. Twinning, nonrighthandedness, and fusion malformations: evidence for heritable causal elements held in common. *Am J Med Genet* 1987b;28:67–84.
- [170] Knopman JM, Krey LC, Oh C, Lee J, McCaffrey C, Noyes N. What makes them split? Identifying risk factors that lead to monozygotic twins after in vitro fertilization. *Fertil Steril* 2014;102:82–9.
- [171] Payne D, Okuda A, Wakatsuki Y, Takeshita C, Iwata K, Shimura T, Yumoto K, Ueno Y, Flaherty S, Mio Y. Time-lapse recording identifies human blastocysts at risk of producing monozygotic twins. *Hum Reprod* 2007;22:i9–10.
- [172] Niimura S. Time-lapse videomicrographic analyses of contractions in mouse blastocysts. *J Reprod Dev* 2003;49:413–23.
- [173] Martin N, Boomsma D, Machin G. A twin-pronged attack on complex traits. *Nat Genet* 1997;17:387–92.
- [174] Silva S, Martins Y, Matias A, Blickstein I. Why are monozygotic twins different? *J Perinat Med* 2011;39:195–202.
- [175] Zwijnenburg PJ, Meijers-Heijboer H, Boomsma DI. Identical but not the same: the value of discordant monozygotic twins in genetic research. *Am J Med Genet B Neuropsychiatr Genet* 2010;153B:1134–49.
- [176] Gilgenkrantz S, Janot C. Monozygotic twins discordant for trisomy 21 or chimeric dizygotic twins? *Am J Med Genet* 1983;15:159–60.
- [177] Marcus-Soekarman D, Hamers G, Velzeboer S, Nijhuis J, Loneus WH, Herbergs J, de Die-Smulders C, Schrandt-Stumpel C, Engelen J. Mosaic trisomy 11p in monozygotic twins with discordant clinical phenotypes. *Am J Med Genet* 2004;124A:288–91.
- [178] Rogers JG, Voullaire L, Gold H. Monozygotic twins discordant for trisomy 21. *Am J Med Genet* 1982;11:143–6.
- [179] West PM, Love DR, Stapleton PM, Winship IM. Paternal uniparental disomy in monozygotic twins discordant for hemihypertrophy. *J Med Genet* 2003;40:223–6.

- [180] Wakita Y, Narahara K, Tsuji K, Yokoyama Y, Ninomiya S, Murakami R, Kikkawa K, Seino Y. De novo complex chromosome rearrangement in identical twins with multiple congenital anomalies. *Hum Genet* 1992;88:596–8.
- [181] Kruyer H, Mila M, Glover G, Carbonell P, Ballesta F, Estivill X. Fragile X syndrome and the (CGG)n mutation: two families with discordant MZ twins. *Am J Hum Genet* 1994;54:437–42.
- [182] Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, Howard E, de Lima RL, Daack-Hirsch S, Sander A, McDonald-McGinn DM, Zackai EH, Lammer EJ, Aylsworth AS, Ardinger HH, Lidral AC, Pober BR, Moreno L, Arcos-Burgos M, Valencia C, Houdayer C, Bahuau M, Moretti-Ferreira D, Richieri-Costa A, Dixon MJ, Murray JC. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 2002;32:285–9.
- [183] Robertson SP, Thompson S, Morgan T, Holder-Espinasse M, Martinot-Duquenoy V, Wilkie AO, Manouvrier-Hanu S. Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. *Eur J Hum Genet* 2006;14:549–54.
- [184] Biousse V, Brown MD, Newman NJ, Allen JC, Rosenfeld J, Meola G, Wallace DC. De novo 14484 mitochondrial DNA mutation in monozygotic twins discordant for Leber's hereditary optic neuropathy. *Neurology* 1997;49:1136–8.
- [185] Blakely EL, He L, Taylor RW, Chinnery PF, Lightowers RN, Schaefer AM, Turnbull DM. Mitochondrial DNA deletion in "identical" twin brothers. *J Med Genet* 2004;41:e19.
- [186] Shaffer LG, Kashork CD, Bacino CA, Benke PJ. Caution: telomere crossing. *Am J Med Genet* 1999;87:278–80.
- [187] Graakjaer J, Bischoff C, Korsholm L, Holstebro S, Vach W, Bohr VA, Christensen K, Kolvraa S. The pattern of chromosome-specific variations in telomere length in humans is determined by inherited, telomere-near factors and is maintained throughout life. *Mech Ageing Dev* 2003;124:629–40.
- [188] Vadlamudi L, Dibbens LM, Lawrence KM, Iona X, McMahon JM, Murrell W, Mackay-Sim A, Scheffer IE, Berkovic SF. Timing of de novo mutagenesis—a twin study of sodium-channel mutations. *N Engl J Med* 2010;363:1335–40.
- [189] Abdellaoui A, Ehli EA, Hottenga JJ, Weber Z, Mbarek H, Willemsen G, van Beijsterveldt T, Brooks A, Hudziak JJ, Sullivan PF, de Geus EJ, Davies GE, Boomsma DI. CNV concordance in 1,097 MZ twin pairs. *Twin Res Hum Genet* 2015;18:1–12.
- [190] Li R, Montpetit A, Rousseau M, Wu SY, Greenwood CM, Spector TD, Pollak M, Polychronakos C, Richards JB. Somatic point mutations occurring early in development: a monozygotic twin study. *J Med Genet* 2014;51:28–34.
- [191] McRae AF, Visscher PM, Montgomery GW, Martin NG. Large autosomal copy-number differences within unselected monozygotic twin pairs are rare. *Twin Res Hum Genet* 2015;18:13–8.
- [192] Breckpot J, Thienpont B, Gewillig M, Allegaert K, Vermeesch JR, Devriendt K. Differences in copy number variation between discordant monozygotic twins as a model for exploring chromosomal mosaicism in congenital heart defects. *Mol Syndromol* 2012;2:81–7.
- [193] Castellani CA, Awamleh Z, Melka MG, O'Reilly RL, Singh SM. Copy number variation distribution in six monozygotic twin pairs discordant for schizophrenia. *Twin Res Hum Genet* 2014;17:108–20.
- [194] Ehli EA, Abdellaoui A, Hu Y, Hottenga JJ, Kattenberg M, van Beijsterveldt T, Bartels M, Althoff RR, Xiao X, Scheet P, de Geus EJ, Hudziak JJ, Boomsma DI, Davies GE. De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems. *Eur J Hum Genet* 2012;20:1037–43.
- [195] Lu P, Wang P, Li L, Xu C, Liu JC, Guo X, He D, Huang H, Cheng Z. Exomic and epigenomic analyses in a pair of monozygotic twins discordant for cryptorchidism. *Twin Res Hum Genet* 2017;20:349–54.
- [196] Petersen BS, Spehlmann ME, Raedler A, Stade B, Thomsen I, Rabionet R, Rosenstiel P, Schreiber S, Franke A. Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genom* 2014;15:564.
- [197] Zhang R, Thiele H, Bartmann P, Hilger AC, Berg C, Herberg U, Klingmuller D, Nurnberg P, Ludwig M, Reutter H. Whole-exome sequencing in nine monozygotic discordant twins. *Twin Res Hum Genet* 2016;19:60–5.
- [198] Morimoto Y, Ono S, Imamura A, Okazaki Y, Kinoshita A, Mishima H, Nakane H, Ozawa H, Yoshiura KI, Kurotaki N. Deep sequencing reveals variations in somatic cell mosaic mutations between monozygotic twins with discordant psychiatric disease. *Hum Genome Var* 2017;4:17032.
- [199] Meltz Steinberg K, Nicholas TJ, Koboldt DC, Yu B, Mardis E, Pamphlett R. Whole genome analyses reveal no pathogenetic single nucleotide or structural differences between monozygotic twins discordant for amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener* 2015;16:385–92.

- [200] Tang J, Fan Y, Li H, Xiang Q, Zhang DF, Li Z, He Y, Liao Y, Wang Y, He F, Zhang F, Shugart YY, Liu C, Tang Y, Chan RCK, Wang CY, Yao YG, Chen X. Whole-genome sequencing of monozygotic twins discordant for schizophrenia indicates multiple genetic risk factors for schizophrenia. *J Genet Genomics* 2017;44:295–306.
- [201] Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands C, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, Slagboom PE, Boomsma DI, Ye K, Guryev V, Arndt PF, Kloosterman WP, de Bakker PI, Sunyaev SR. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* 2015;47:822–6.
- [202] Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J, Rolf B. Finding the needle in the haystack: differentiating “identical” twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Sci Int Genet* 2014;9:42–6.
- [203] Mikeska T, Craig JM. DNA methylation biomarkers: cancer and beyond. *Genes* 2014;5:821–64.
- [204] Bell JT, Saffery R. The value of twins in epigenetic epidemiology. *Int J Epidemiol* 2012;41:140–50.
- [205] Busche S, Shao X, Caron M, Kwan T, Allum F, Cheung WA, Ge B, Westfall S, Simon MM, Multiple Tissue Human Expression R, Barrett A, Bell JT, McCarthy MI, Deloukas P, Blanchette M, Bourque G, Spector TD, Lathrop M, Pastinen T, Grundberg E. Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol* 2015;16:290.
- [206] McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, Painter JN, Martin NG, Visscher PM, Montgomery GW. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol* 2014;15:R73.
- [207] van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, Ehli EA, Davies GE, van IJterson M, Breeze CE, Beck S, Consortium B, Suchiman HE, Jansen R, van Meurs JB, Heijmans BT, Slagboom PE, Boomsma DI. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun* 2016;7:11115.
- [208] Yet I, Tsai PC, Castillo-Fernandez JE, Carnero-Montoro E, Bell JT. Genetic and environmental impacts on DNA methylation levels in twins. *Epigenomics* 2016;8:105–17.
- [209] Ollikainen M, Smith KR, Joo EJ, Ng HK, Andronikos R, Novakovic B, Abdul Aziz NK, Carlin JB, Morley R, Saffery R, Craig JM. DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum Mol Genet* 2010;19:4176–88.
- [210] Gordon L, Joo JE, Powell JE, Ollikainen M, Novakovic B, Li X, Andronikos R, Cruickshank MN, Conneely KN, Smith AK, Alisch RS, Morley R, Visscher PM, Craig JM, Saffery R. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res* 2012;22:1395–406.
- [211] Gordon L, Joo JH, Andronikos R, Ollikainen M, Wallace EM, Umstad MP, Permezel M, Oshlack A, Morley R, Carlin JB, Saffery R, Smyth GK, Craig JM. Expression discordance of monozygotic twins at birth: effect of intrauterine environment and a possible mechanism for fetal programming. *Epigenetics* 2011;6:579–92.
- [212] Oates NA, van Vliet J, Duffy DL, Kroes HY, Martin NG, Boomsma DI, Campbell M, Coulthard MG, Whitelaw E, Chong S. Increased DNA methylation at the AXIN1 gene in a monozygotic twin from a pair discordant for a caudal duplication anomaly. *Am J Hum Genet* 2006;79:155–62.
- [213] Martin GM. Epigenetic drift in aging identical twins. *Proc Natl Acad Sci U S A* 2005;102:10413–4.
- [214] Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* 2005;102:10604–9.
- [215] Martino D, Loke YJ, Gordon L, Ollikainen M, Cruickshank MN, Saffery R, Craig JM. Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biol* 2013;14:R42.
- [216] Zhang N, Zhao S, Zhang SH, Chen J, Lu D, Shen M, Li C. Intra-monozygotic twin pair discordance and longitudinal variation of whole-genome scale DNA methylation in adults. *PLoS One* 2015;10:e0135022.
- [217] Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. *Aging Cell* 2015;14:924–32.
- [218] Jylhava J, Pedersen NL, Hagg S. Biological age predictors. *EBioMedicine* 2017;21:29–36.
- [219] Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, Myers RM, Absher D. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol* 2013;14:R102.

- [220] Tan Q, Heijmans BT, Hjelmborg JV, Soerensen M, Christensen K, Christiansen L. Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort. *Int J Epidemiol* 2016;45:1146–1158.
- [221] Gartner K. A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 year long effort to standardize laboratory animals? *Int J Epidemiol* 2012;41:335–41.
- [222] Pujadas E, Feinberg AP. Regulated noise in the epigenetic landscape of development and disease. *Cell* 2012;148:1123–31.
- [223] Whitelaw NC, Chong S, Whitelaw E. Tuning in to noise: epigenetics and intangible variation. *Dev Cell* 2010;19:649–50.
- [224] Chen M, Baumbach J, Vandin F, Rottger R, Barbosa E, Dong M, Frost M, Christiansen L, Tan Q. Differentially methylated genomic regions in birth-weight discordant twin pairs. *Ann Hum Genet* 2016;80:81–7.
- [225] Tsai PC, Van Dongen J, Tan Q, Willemsen G, Christiansen L, Boomsma DI, Spector TD, Valdes AM, Bell JT. DNA methylation changes in the IGF1R gene in birth weight discordant adult monozygotic twins. *Twin Res Hum Genet* 2015;18:635–46.
- [226] Chiarella J, Tremblay RE, Szyf M, Provencal N, Booij L. Impact of early environment on children's mental health: lessons from DNA methylation studies with monozygotic twins. *Twin Res Hum Genet* 2015;1–12.
- [227] Craig JM. Epigenetics in twin studies. *Med Epigenetics* 2013;1:70–7.
- [228] Barker DJ, Osmond C. Low birth weight and hypertension. *BMJ* 1988;297:134–5.
- [229] Gluckman PD, Hanson MA, Buklijas T. A conceptual framework for the developmental origins of health and disease. *J Dev Orig Health Dis* 2010;1:6–18.
- [230] Woo Baidal JA, Locks LM, Cheng ER, Blake-Lamb TL, Perkins ME, Taveras EM. Risk factors for childhood obesity in the first 1,000 Days: a systematic review. *Am J Prev Med* 2016;50:761–79.
- [231] Rappaport SM. Genetic factors are not the major causes of chronic diseases. *PLoS One* 2016;11:e0154387.
- [232] Machin GA. Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. *Am J Med Genet* 1996;61:216–28.
- [233] Plomin R. Commentary: why are children in the same family so different? Non-shared environment three decades later. *Int J Epidemiol* 2011;40:582–92.
- [234] Stromswold K. Why aren't identical twins linguistically identical? Genetic, prenatal and postnatal factors. *Cognition* 2006;101:333–84.
- [235] Dwyer T, Blizzard L, Morley R, Ponsonby AL. Within pair association between birth weight and blood pressure at age 8 in twins from a cohort study. *Br Med J* 1999;319:1325–9.
- [236] Bekhit MT, Greenwood PA, Warren R, Aarons E, Jauniaux E. In utero treatment of severe fetal anaemia due to parvovirus B19 in one fetus in a twin pregnancy—a case report and literature review. *Fetal Diagn Ther* 2009;25:153–7.
- [237] Dickinson JE, Keil AD, Charles AK. Discordant fetal infection for parvovirus B19 in a dichorionic twin pregnancy. *Twin Res Hum Genet* 2006;9:456–9.
- [238] Jamieson DJ, Read JS, Kourtis AP, Durant TM, Lampe MA, Dominguez KL. Cesarean delivery for HIV-infected women: recommendations and controversies. *Am J Obstet Gynecol* 2007;197:S96–100.
- [239] Pimentel JD, Szymanski LJ, Samuel LP, Meier FA. Discordant *Streptococcus agalactiae* (Group B streptococcus) gestational infection in monochorionic/diamniotic and dichorionic/diamniotic twins. *Fetal Pediatr Pathol* 2012;31:176–83.
- [240] Schiesser M, Sergi C, Enders M, Maul H, Schnitzler P. Discordant outcomes in a case of parvovirus b19 transmission into both dichorionic twins. *Twin Res Hum Genet* 2009;12:175–9.
- [241] Pawelec G, Goldeck D, Derhovanessian E. Inflammation, ageing and chronic disease. *Curr Opin Immunol* 2014;29:23–8.
- [242] Tabas I, Glass CK. Anti-inflammatory therapy in chronic disease: challenges and opportunities. *Science* 2013;339:166–72.
- [243] Brodin P, Jovic V, Gao T, Bhattacharya S, Angel CJ, Furman D, Shen-Orr S, Dekker CL, Swan GE, Butte AJ, Maecker HT, Davis MM. Variation in the human immune system is largely driven by non-heritable influences. *Cell* 2015;160:37–47.
- [244] Jacques SM, Qureshi F. Chronic villitis of unknown etiology in twin gestations. *Pediatr Pathol* 1994;14:575–84.
- [245] Phung DT, Blickstein I, Goldman RD, Machin GA, LoSasso RD, Keith LG. The Northwestern Twin Chorionicity Study: I. Discordant inflammatory findings that are related to chorionicity in presenting versus nonpresenting twins. *Am J Obstet Gynecol* 2002;186:1041–5.
- [246] Hopper JL. Why 'common environmental effects' are so uncommon in the literature. In: Spector TD, Snieder H, MacGregor AJ, editors. *Advances in twin and sib-pair analysis*. London: Oxford University Press; 2000. p. 151–65.
- [247] Lange K, Weeks D, Boehnke M. Programs for pedigree analysis: MENDEL, Fisher and dGene. *Genet Epidemiol* 1988;5:471–2.
- [248] Rijdsdijk FV, Sham PC. Analytic approaches to twin data using structural equation models. *Brief Bioinform* 2002;3:119–33.

- [249] Fisher RA. Limits to intensive production in animals. *Br Agric Bull* 1951;4:217–8.
- [250] Hopper JL. Heritability, *Encyclopedia of Biostatistics*. In: Amitage P, Colton T, editors. John Wiley & Sons, Ltd; 2005.
- [251] Witte JS, Carlin JB, Hopper JL. Likelihood-based approach to estimating twin concordance for dichotomous traits. *Genet Epidemiol* 1999;16:290–304.
- [252] Plomin R, DeFries JC, Knopik VS, Neiderhiser JM. *Behavioral Genetics (6th Edition)* Worth Publishers. New York, 2013, ISBN-10: 1-4292-4215-9.
- [253] Hannah MC, Hopper JL, Mathews JD. Twin concordance for a binary trait. I. Statistical models illustrated with data on drinking status. *Acta Genet Med Gemellol* 1983;32:127–37.
- [254] Munteanu SE, Menz HB, Wark JD, Christie JJ, Scurrah KJ, Bui M, Erbas B, Hopper JL, Wluka AE. Hallux valgus, by nature or nurture? A twin study. *Arthritis Care Res* 2016.
- [255] Burton P, Tiller K, Gurrin L, Cookson W, Musk A, Palmer L. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genet Epidemiol* 1999;17(118):140.
- [256] Nyholt DR, Gillespie NA, Heath AC, Martin NG. Genetic basis of male pattern baldness. *J Invest Dermatol* 2003;121:1561–4.
- [257] Zaloumis SG, Scurrah KJ, Harrap SB, Ellis JA, Gurrin LC. Non-proportional odds multivariate logistic regression of ordinal family data. *Biom J* 2015;57:286–303.
- [258] Scurrah KJ, Palmer LJ, Burton PR. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genet Epidemiol* 2000;19:127–48.
- [259] Yashin I, Vaupel J, Iachine I. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Math Popul Stud* 1995;5:145–59.
- [260] Hamilton AS, Mack TM. Puberty and genetic susceptibility to breast cancer in a case-control study in twins. *N Engl J Med* 2003;348:2313–22.
- [261] Goldberg J, Fischer M. Co-twin control methods, *encyclopedia of statistics in behavioral science*. John Wiley & Sons, Ltd; 2005.
- [262] Nowson CA, Green RM, Hopper JL, Sherwin AJ, Young D, Kaymakci B, Guest CS, Smid M, Larkins RG, Wark JD. A co-twin study of the effect of calcium supplementation on bone density during adolescence. *Osteoporos Int* 1997;7:219–25.
- [263] Sjölander A, Frisell T, Öberg S. Causal interpretation of between-within models for twin research, epidemiologic methods. 2012. p. 217.
- [264] Sjölander A, Lichtenstein P, Larsson H, Pawitan Y. Between-within models for survival analysis. *Stat Med* 2013;32:3067–76.
- [265] Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T. Regression models for twin studies: a critical review. *Int J Epidemiol* 2005;34:1089–99.
- [266] Hopper JL, Seeman E. The bone density of female twins discordant for tobacco use. *N Engl J Med* 1994;330:387–92.
- [267] Staton S, Thorpe K, Thompson C, Danby S. To separate or not to separate? Parental decision-making regarding the separation of twins in the early years of schooling. *J Early Child Res* 2012;10:196–208.
- [268] Goldman SM, Quinlan PJ, Ross GW, Marras C, Meng C, Bhudhikanok GS, Comyns K, Korell M, Chade AR, Kasten M, Priestley B, Chou KL, Fernandez HH, Cambi F, Langston JW, Tanner CM. Solvent exposures and Parkinson's disease risk in twins. *Ann Neurol* 2012;71:776–84.
- [269] Gatz M, Harris JR, Kaprio J, McGue M, Smith NL, Snieder H, Spiro 3rd A, Butler DA, Institute of Medicine Committee on Twins S. Cohort profile: the national Academy of sciences-national research Council twin registry (NAS-NRC twin registry). *Int J Epidemiol* 2015;44:819–25.
- [270] Hopper JL, Foley DL, White PA, Pollaers V. Australian twin registry: 30 years of progress. *Twin Res Hum Genet* 2013;16:34–42.
- [271] Moayyeri A, Hammond CJ, Hart DJ, Spector TD. The UK adult twin registry (TwinsUK resource). *Twin Res Hum Genet* 2013;16:144–9.
- [272] Davis OSP, Band G, Pirinen M, Haworth CMA, Meaburn EL, Kovas Y, Harlaar N, Docherty SJ, Hanscombe KB, Trzaskowski M, Curtis CJC, Strange A, Freeman C, Bellenguez C, Su Z, Pearson R, Vukcevic D, Langford C, Deloukas P, Hunt S, Gray E, Dronov S, Potter SC, Tashakkori-Ghanbaria A, Edkins S, Bumpstead SJ, Blackwell JM, Bramon E, Brown MA, Casas JP, Corvin A, Duncanson A, Jankowski JAZ, Markus HS, Mathew CG, Palmer CNA, Rautanen A, Sawcer SJ, Trembath RC, Viswanathan AC, Wood NW, Barroso I, Peltonen L, Dale PS, Petrill SA, Schalkwyk LS, Craig IW, Lewis CM, Price TS, Donnelly P, Plomin R, Spencer CCA. The correlation between reading and mathematics ability at age twelve has a substantial genetic component. *Nat Commun* 2014;5:4204.
- [273] Harrap SB, Stebbing M, Hopper JL, Hoang HN, Giles GG. Familial patterns of covariation for cardiovascular risk factors in adults: the Victorian Family Heart Study. *Am J Epidemiol* 2000;152:704–15.
- [274] StataCorp. Stata statistical software: release 14. StataCorp LLC College Station, T; 2015.

- [275] Hur YM, Craig JM. Twin registries worldwide: an important resource for scientific research. *Twin Res Hum Genet* 2013;16:1–12.
- [276] Kaprio J. The Finnish twin cohort study: an update. *Twin Res Hum Genet* 2013;16:157–62.
- [277] Skytthe A, Christiansen L, Kyvik KO, Bodker FL, Hvidberg L, Petersen I, Nielsen MM, Bingley P, Hjelmberg J, Tan Q, Holm NV, Vaupel JW, McGue M, Christensen K. The Danish Twin Registry: linking surveys, national registers, and biological information. *Twin Res Hum Genet* 2013;16:104–11.
- [278] Strachan E, Hunt C, Afari N, Duncan G, Noonan C, Schur E, Watson N, Goldberg J, Buchwald D. University of Washington Twin Registry: poised for the next generation of twin research. *Twin Res Hum Genet* 2013;16:455–62.
- [279] Ordonana JR, Rebollo-Mesa I, Carrillo E, Colodro-Conde L, Garcia-Palomo FJ, Gonzalez-Javier F, Sanchez-Romera JF, Aznar Oviedo JM, de Pancorbo MM, Perez-Riquelme F. The Murcia Twin Registry: a population-based registry of adult multiples in Spain. *Twin Res Hum Genet* 2013;16:302–6.
- [280] Ferreira PH, Oliveira VC, Junqueira DR, Cisneros LC, Ferreira LC, Murphy K, Ordoñana JR, Hopper JL, Teixeira-Salmela LF. The Brazilian twin registry. *Twin Res Hum Genet* 2016;19:687–91.
- [281] Moayyeri A, Hammond CJ, Valdes AM, Spector TD. Cohort Profile: TwinsUK and healthy ageing twin study. *Int J Epidemiol* 2013;42:76–85.
- [282] Buchwald D, Kaprio J, Hopper JL, Sung J, Goldberg J, Fortier I, Busjhan A, Sumathipala A, Cozen W, Mack T, Craig JM, Harris JR. International network of twin registries (INTR): building a platform for international collaboration. *Twin Res Hum Genet* 2014;17:574–7.
- [283] Silventoinen K, Jelenkovic A, Sund R, Honda C, Aaltonen S, Yokoyama Y, Kaprio J. The CODATwins Project: The Cohort Description of Collaborative Project of Development of Anthropometrical Measures in Twins to Study Macro-Environmental Variation in Genetic and Environmental Effects on Anthropometric Traits. *Twin Res Hum Gen* 2015;18:348–60.

The Biological Basis of Aging: Implications for Medical Genetics

Junko Oshima¹, Fuki M. Hisama², George M. Martin¹

¹Department of Pathology, University of Washington, Seattle, WA, United States

²Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, United States

15.1 INTRODUCTION

Evolutionary biology has provided a robust theory to explain why we age, but we have much less confidence that we understand how we age, by which we mean the proximal molecular mechanisms of aging. Medical geneticists are in a good position to advance our knowledge of such mechanisms. The first strategy is the time-honored approach of mapping, identifying and characterizing the relevant gene actions underlying important late-onset disorders of aging, such as dementias of the Alzheimer type, atherosclerosis, ocular cataracts, type II diabetes mellitus, osteoporosis, osteoarthritis, and cancer. Significant progress has been made using this strategy, including major advances in our understanding of segmental progeroid syndromes such as the Werner and Hutchinson–Gilford syndromes. The second strategy is to investigate the genetic basis for unusually well-preserved structure and function during the latter half of the usual life span. Unfortunately, medical geneticists have been too preoccupied with disease and, with some notable exceptions (e.g., studies of centenarians), physicians and geneticists have not shown a strong inclination to investigate exceptional well-preserved late-life phenotypes. The new statistical and molecular tools at our disposal are impressive, but they have not been matched by the application of sensitive functional assays [2].

15.2 WHAT IS AGING?

Some gerontologists, particularly those interested in the aging of plants, sharply differentiate between the terms aging and senescing. Senescent changes, they would argue, are those structural and functional changes that occur near the end of the life cycle of a cell, tissue, organ, or organism, and are associated with the impending death of the tissue or organism. By contrast, the term aging would be used for any change in structure or function throughout the life cycle. In other words, some would argue that “aging begins at birth.” Most gerontologists who work with mammals and human subjects, however, use the two terms more or less interchangeably. While different scholars define human aging in various ways, most include exponential increases of age-specific mortality rate and declines of the physiological functions as general characteristics of human aging processes. These parameters were primarily derived from cross-sectional population studies. Researchers of basic biology of human aging at the organismal level have a major difficulty that stems from the fact that, unlike model animals, humans are genetically heterogeneous and undergo behavioral and environmental changes throughout their lifetimes. Moreover, due to the relatively long life span of humans, the cohort studies of longevity, for example, take many years to complete. One approach to testing the validity of the population studies of human longevity and aging is to compare the findings

of other model organisms such as yeast (*Saccharomyces cerevisiae*), fruit fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), and mice, with those from *Homo sapiens* [3]. They continue to create a large segment of the foundations for the progress we have seen in human biology, an example of which is the recent comparative studies on proteostasis under conditions of stress, research that was initiated by the discovery of the remarkable longevities of species of bivalve mollusks [4]. Moreover, their contributions have typically been more incisive, in part because of the experimental tractability of their materials.

None would deny the importance of development in determining the subsequent life history of an organism. A minority of gerontologists in fact embrace the idea that aging is “programmed,” despite cogent arguments that defend what is sometimes referred to as the “classical” evolutionary biological theory of aging, a theory (discussed below) that is based upon the nonadaptive nature of biological aging [5]. Most gerontologists are concerned with declines in structure and function that gradually and insidiously unfold after the organism has achieved the young, mature adult phenotype. At the level of populations, these functional declines translate into an exponential increase in the force of mortality over unit time—the hazard function or instantaneous mortality rate [6,7]. This is the famous Gompertz relationship [8]. This was modified by Makeham [9], who included a constant, A , to account for kinetic departures presumed to have resulted from causes of death during the early life history that were age-independent. The Gompertz–Makeham equation can thus be given as the sum of two types of mortalities, age-independent and age-dependent, the latter exhibiting exponential kinetics over the adult life span:

$$\mu_x = A + Re^{ax}$$

where μ_x is the force of mortality at a given age, x ; A is the Makeham constant; R is the hypothetical value for the force of mortality at birth, the lowest force of mortality, or the Y intercept in a graphic plot of age (X -axis) versus force of mortality (Y -axis) (Fig. 15.1); e is an exponent; and μ is a constant representing the slope of the graphical plot (Fig. 15.1). Fig. 15.1 illustrates differing rates of exponential increases in the force of mortality for two noninbred wildtype murine species despite comparable values for R . These two species, *Peromyscus leucopus* and *Mus musculus*, are of approximately the same size and were

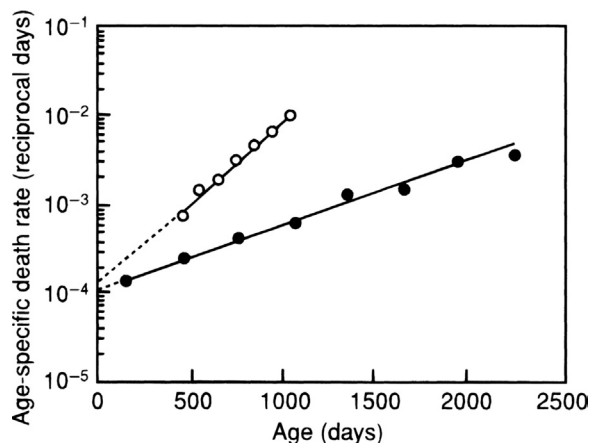


Figure 15.1 Gompertz function plot of the age-specific mortality rates for combined sexes of two different murine species of contrasting maximum life-span potentials but of comparable size. Both species were wild-type and randomly bred from small cohorts captured near the Argonne National Laboratories (Argonne, IL) by the late George A. Sacher. They were housed under essentially identical conditions (caging, bedding, humidity, temperature, diet) in adjacent animal rooms with no special efforts to establish specific pathogen-free conditions (G. A. Sacher, personal communication to G. M. Martin). The longer-lived species (●), *P. leucopus*, was found to have a maximum life span of about 8 years, approximately twice that of *M. musculus* (○). (From Sacher G.A. Evolution of longevity and survival characteristics in mammals. In: Schneider E.L., editor. The genetics of aging. New York: Plenum Press; 1978.)

housed and fed throughout their lifetimes under identical conditions [10]. The maximum life span of *Peromyscus* sp. was found to be about 8 years, about twice that of *Mus* sp. These data illustrate the importance of genetic factors in the determination of approximate life potential. Given the considerable evolutionary distance between these two species (at least 15 million years), this is not a surprising result. Mortality rates of different human populations in the 20th century also followed the Gompertz–Makeham relationship until they reach very old age [6,7,11].

Experiments employing very large populations of aging cohorts of fruit flies and medflies have been reported as showing dramatic departures from Gompertz kinetics within the oldest cohorts, with apparent decreases in the force of mortality at very advanced ages [12]. Very aged flies, however, may become virtually immobilized and may therefore be protected from environmental hazards related, for example, to attempts at flight. Declines in age-specific

mortality rates have also been seen for very aged individuals in human populations [13]. Perhaps this observation might also be explained by behavioral changes (e.g., a more protective environment) in extreme old age. It will nevertheless be prudent to explore various non-Gompertzian models of mortality in human populations that adequately describe mortality at extreme ages. The existence and estimate of the upper limit of the human longevity or maximum life span of humans has been a focus of debates, although it is generally believed to be around 125 years based on the verified longest-lived human [14].

Functional declines can be documented in virtually every organ system starting shortly after sexual maturation. Most physiologic declines, at least in cross-sectional studies, exhibit linear declines, the slopes of which are variable [15]. Declines in the various physiological processes (and underlying molecular and biochemical processes) that maintain optimum functions are likely to “set the stage” for the plethora of late-life disorders and diseases, some 87 of which have recently been tabulated, all of which are subject to both genetic and environmental modulations [3]. Observations of exponential increases in the force of mortality within populations should not lead one to conclude that underlying processes of aging or incidences of geriatric diseases necessarily exhibit exponential kinetics. Consider, for example, the world records of marathon runners, which select for the most robust, physically fit members of our population. This is an attractive assay, as it tests for fitness of multiple organ systems and one’s ability to maintain metabolic homeostasis. Declines are observable during the fourth decade, later than what is the case for sprinters. This probably occurs, in part, because it takes considerably more training and experience in perfecting one’s optimal pacing for a marathon. It also takes years for the gradual development of such compensatory processes as cardiac muscle hypertrophy. For a remarkably wide range of sports, peak activity occurs during the third decade [16].

Many major diseases of late life, however, show exponential increases in age-specific incidence and prevalence, although there may be slight declines in age-specific incidence at very advanced ages, raising the question of selection for genotypic resistance against specific late-life disorders. Alzheimer disease serves as a good example [17,18]. Fig. 15.2 summarizes the results of several community-based studies of the age-specific

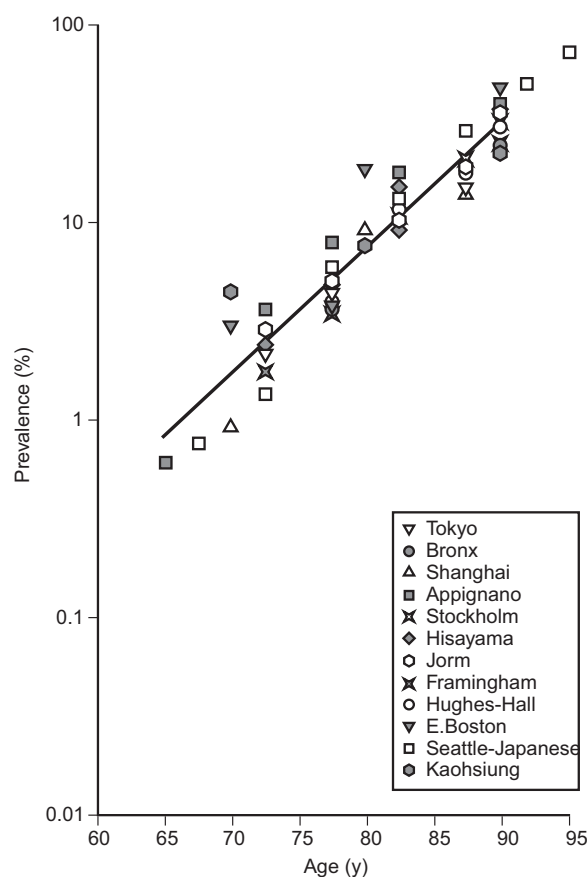


Figure 15.2 Evidence for exponential increases in the age-specific incidence of probable Alzheimer disease in six different community-based studies. (From Breteler MMB, et al. Epidemiology of Alzheimer disease. *Epidemiol Rev* 1992;14:59–82.)

incidence of late-life dementias, most of which are due to dementias of the Alzheimer type [19].

Longitudinal studies of physiologic parameters may exhibit striking variation among individuals [20]. Fig. 15.3 illustrates the case of a measure of renal function. By this measure, some individuals show no evidence of a decline in renal function; some may have superior compensations for structural alterations [21]. Are any of these varied patterns of functional decline (or lack of decline) in apparently normal aging human subjects determined, in part, by constitutional differences in the genotype? Essentially no research has been carried out to address this important question. Medical geneticists have an obvious bias in favor of the discovery of deleterious allelic variants. There is a great need to define allelic variants that, in ordinary environments, are associated with

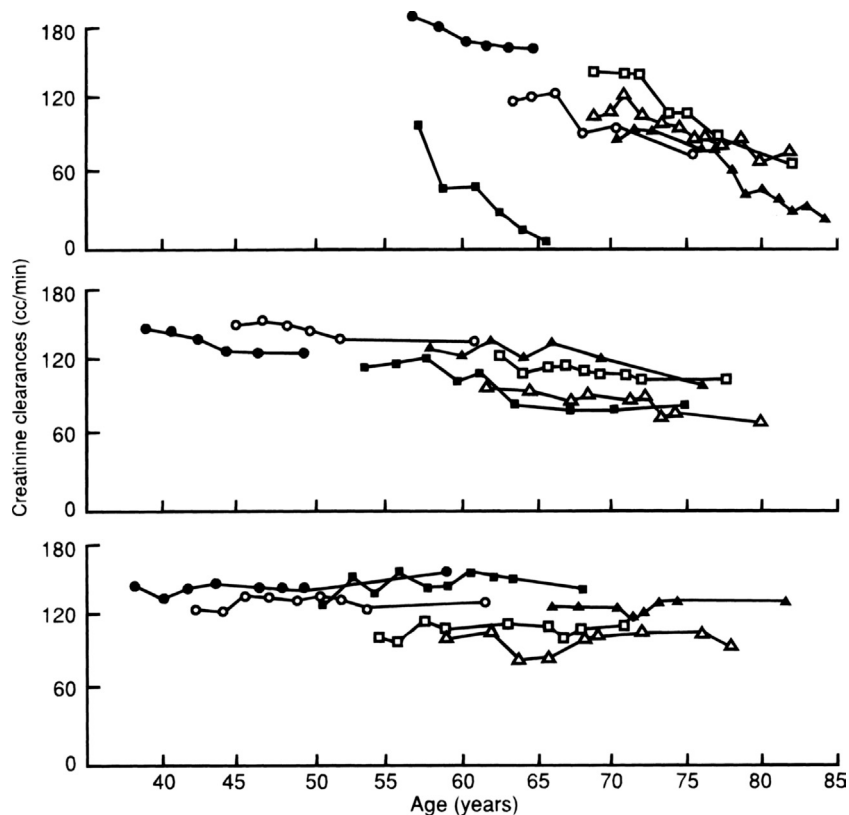


Figure 15.3 Longitudinal studies of creatinine clearance (an approximate measure of the glomerular filtration rate) for a representative sample of a subset of 446 clinically normal male volunteers in the Baltimore Longitudinal Study of Aging of the National Institute on Aging followed between 1958 and 1981. The results could be classified in one of three major patterns. The top panel illustrates substantial rates of decline in this measure of renal function for six representative subjects who were followed for 8–14 years. The middle panel illustrates a pattern of slight, but significant, decline for six representative subjects followed for 11–22 years. For the six representative subjects in the bottom panel, who were followed for periods of 15–21 years, there were no apparent declines in this measure of renal function. (From Lindeman RD, et al. Longitudinal studies on the rate of decline in renal function with age. *J Am Geriatr Soc* 1985;33:278–85.)

the maintenance of enhanced structure and function during aging. One such example may be the ApoE2 allele, the prevalence of which is significantly increased in centenarians [22]. The APOE2 results are understandable in that there is evidence that carriers are provided with some protection against Alzheimer disease [23].

There is very strong evidence indicating a major role for the constitutional genotype in the susceptibility to various familial and “sporadic” forms of Alzheimer disease. In addition to the apparent protection by the E2 allele of ApoE noted above, individuals carrying the E4 allele, particularly homozygotes, are at elevated risk to develop the disease [24]. ApoE4 may act

as an age-of-onset modifier for the common relatively late-onset forms of the disorder. Although deserving of additional research, there is evidence for multiple mechanisms underlying the effects of the E4 allele [25]. This important subject is considered in much more detail in a later volume. The lay perception that aging is accompanied by a global loss of cognitive function is certainly incorrect. Only selected regions of the nervous system appear to be particularly susceptible.

The other type of terminally differentiated cell receiving special attention from gerontologists is the multinucleated skeletal muscle cell. Structural and functional declines in skeletal muscle vary from muscle to muscle,

with weight-bearing muscles being more susceptible; the rates of these declines accelerate after about age 70 [26]. At least some proportion of the pathology is likely to be related to denervation atrophy [27]. Disuse atrophy is also an important component [28].

Many postreplicative aging cell types gradually accumulate a mixture of complex fluorescent pigments called lipofuscins. These are likely to vary in composition from tissue to tissue. Most investigators believe that all lipofuscins are the products of lipid peroxidation reactions. They could therefore be regarded as evidence in support of theories of aging that invoke oxidative alterations of macromolecules. That lipofuscins are markers of some underlying aging process is a theory supported by three lines of evidence. First, they appear to be almost invariable features of aging in an amazing variety of organisms, including certain strains of fungi under certain growth conditions (e.g., *Podospira anserina* and *Neurospora crassa*), paramecia, nematodes, snails, fruit flies, houseflies, frogs, parrots, house mice, rats, guinea pigs, cats, dogs, pigs, monkeys, and humans [29,30]. Second, quantitative studies of lipofuscin rates of accumulation in the hearts of dogs and humans indicate appropriate correlations with the lifespan potentials of those species [31]. (No such correlations have been observed, however, among cardiac tissues from a group of primates of contrasting life spans [32].) Third, age-related increases in concentrations of at least some classes of lipofuscins are blunted by caloric restriction, an intervention known to increase life span in mammals [33,34].

The importance of extracellular aging has been emphasized by the late Robert R. Kohn [35]. Long-lived proteins, such as lens crystallines and collagens, are particularly susceptible to a variety of posttranslational alterations; these can result in changes of amino acids [36]. Diabetics, who have many progeroid features, are particularly susceptible to glycation of proteins [37]. Advanced glycation end products may also play an important role in the genesis of osteoarthritis [38]. Modified matrix components could perturb cell–matrix interactions and hence change cell function. Such a scenario has been suggested to play a role in the genesis of atherosclerosis [39,40].

15.3 WHY DO WE AGE?

Evolutionary biologists believe that they have an answer to the ultimate cause of aging in age-structured

populations (i.e., populations that consist, at any given time, of cohorts of varying chronological ages) [41]. Simply put, we age because senescent phenotypes escape the force of natural selection [42]. This theory was developed for the case of species with age-structured populations, a situation that occurs when there are serial episodes of reproduction in an individual's lifetime, as opposed to one massive “big-bang” production of progeny in short-lived animals [43]. For human populations, the late William Hamilton showed that the force of natural selection for or against alleles that do not reach phenotypic expression until about the age around 45 years is essentially nil [44,45]. More recently, a mathematically rigorous challenge to Hamilton's theory has been published [46]. It describes a scenario whereby the force of natural selection can actually increase during aging. There are in fact some species of fish that continue to grow and, as such, become more like predators than prey. Under those circumstances, it is easy to imagine declines in the force of natural selection with age.

August Weissmann, one of the giants of 19th century biology, postulated a limited life span of somatic cells, from which he proposed programmed death theory with the idea that aging is good for the species in that it results in enhanced resources for the young. While a finite replicative capacity of somatic cells was later confirmed experimentally [47], there has been no good evidence that aging evolved because it was adaptive for the species or the individual. Essentially all population geneticists who have considered this issue have concluded that aging is nonadaptive. A striking demonstration that single gene mutations can extend the life spans of nematodes, fruit flies, and mice—sometimes dramatically [1,48]—can be interpreted as providing evidence against the classical evolutionary biologic theory of aging, which would predict a highly polygenic modulation. We have since learned, however, that a remarkable number of mutations at single loci can enhance the life spans of model organisms, notably of *C. elegans*, where numerous single gene mutations and genes that have been downregulated by RNAi have been shown to provide substantial increases in life span [49]. A caveat in the interpretation of such studies, however, is that these increases in life span have typically been only examined under conventional laboratory conditions. When challenged by competition experiments with wildtype organisms under other conditions, such as “feast or famine” conditions of the availability of their bacterial

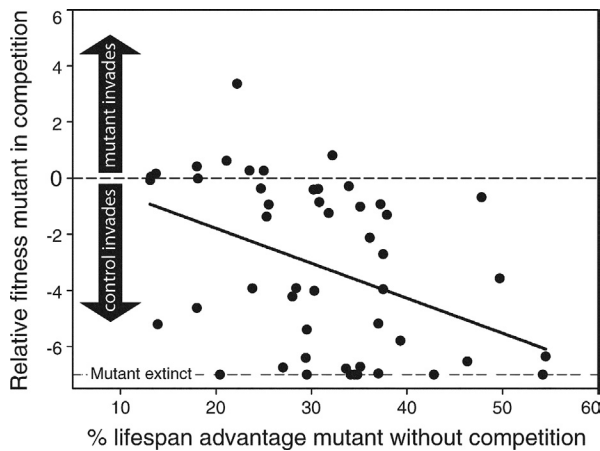


Figure 15.4 Genetic variants of model organisms (in this case, yeast) that exhibit enhanced life spans typically exhibit reductions in relative fitness in competition experiments. (From Briga M, Verhulst S. What can long-lived mutants tell us about mechanisms causing aging and lifespan variation in natural environments? *Exp Gerontol* 2015;71:21–6.)

diets, that advantage was shown to have disappeared in *age1*, the first mutation shown to enhance life span in *C. elegans* [50]. In a systematic competition study involving long-lived yeast mutants versus the parental wildtype strain, the wildtype typically exhibited greater fitness [51] (Fig. 15.4). There is a lesson here for clinical investigators interested in translational research towards the enhancements of human health spans and life spans based upon laboratory experiments in model organisms. Human subjects not only show enormous differences in their genomes and epigenomes, they may often share strikingly different environmental exposures and may lose fitness even within comparable environments when competed with nonmutant individuals.

We have known for many decades that a single environmental manipulation, caloric restriction (or, more conservatively, dietary restriction, since restriction of a single amino acid, methionine, can give comparable results) [52], can substantially enhance life span in a remarkably wide range of species [53]. Once aging, some would argue that these observations support a “programmed” mechanism of aging—that aging involves sequential, determinative changes in gene expression that actively produce aging. One interpretation of both the single gene mutation and caloric restriction experiments, however, is that all or many of them are examples of *diapause*s—time-outs from the business of

reproduction during “bad times”—be they nutritional, climatic, or other environmental challenges.

In terms of genetic mechanisms that form the basis of the classical evolutionary theory of aging, two ideas currently dominate the field. The first, championed by the late Peter Medawar, is generally referred to as “mutation accumulation” [54]. This is an unfortunate name, as the mutations in question are not somatic mutations developing during the life span, but germline mutations that do not reach phenotypic expression until late in the life course, when the force of natural selection would be attenuated. Huntington disease is the prototypical example. Haldane was puzzled by the surprisingly high prevalence of this disorder, which exceeds 15 per 100,000 in some western European populations [55], while germline mutations typically have frequencies of about one in a million. Haldane suggested that the reason the mutation survived in the population was because of its delayed manifestations, thus escaping the force of natural selection [56]. If that were the case, there would be selection for “suppressor alleles” that progressively delayed the age of phenotypic expression. Medawar concluded that many such suppressors might only delay these deleterious effects [54]; eventually, however, the delayed age of expression would be such that there could be little or no influence of natural selection. This scenario, especially when coupled with the other mechanisms discussed below, would result in an enormous degree of heterogeneity in patterns of aging among individuals in out-breeding populations. Thus, each of us may be essentially unique in precisely how we age.

The second dominant idea was first clearly enunciated by George C. Williams [57], and has been referred to as antagonistic pleiotropy; it has been elaborated by Michael R. Rose [42]. By this view, some varieties of genes might have been selected because of good effects early in the life span, but may also have deleterious effects late in the life span, thus contributing to aging phenotypes. As one potential example, Williams speculated that alleles selected because of enhanced incorporation of calcium into bones might be responsible for forms of calcific arteriosclerosis, when acting over long periods. There have been many suggested examples in the literature, but they have been hard to definitively establish. Potential examples include atherosclerosis [58], the role of the apolipoprotein E4 allele in Alzheimer disease [59], common late-life cancers [60], and

immunosenesence [61]. Surprisingly, a single example of what has been called “reverse” antagonistic pleiotropy has now been documented in mice, at least under standard laboratory conditions [62]. Medical geneticists are in a good position to suggest a number of other examples, and perhaps to provide supportive evidence. Such research has the potential to illuminate the most basic aspects of the aging problem. It may also have important translational implications.

Another conceptual formulation, one that overlaps with what has been discussed above, is that there is, inevitably, a trade-off of energetic resources expended by an organism for purposes of reproduction and resources devoted to the maintenance of the macromolecular integrity of the organism. Examples include repair of DNA, scavenging of abnormal proteins, and replacement of effete somatic cells; this is the disposable soma theory of Tom Kirkwood [63]. These ideas can be generalized as life history “optimization” theories of aging [64]. Experimental evidence in *Drosophila* sp. supports both optimization theories and the mutation accumulation theory of aging [64,65]. The relative quantitative contributions of each theory, however, particularly in *H. sapiens*, are completely unknown. As pointed out by Partridge and Barton [64], the resolution of this issue has potentially profound implications for the future life history of our species. If optimization mechanisms predominate, any lifespan extensions may be offset by the trade-off of lower early fertility, delayed maturation, and potential increases in early life history morbidity and mortality. If mutation accumulation mechanisms prevail, enhanced life span attributable to the elimination of such constitutional mutations would presumably have few effects on early lifespan structure and function. Given a major role for optimization theories, a continuation of the present secular trends of elective delays in the ages of reproduction in the developed societies would predict the emergence, by indirect selection, of increased life spans and related declines in early fertility after several centuries of continued evolution of our species. It has been well documented that advanced parental age is associated with increases in germline mutations. Of particular relevance to our interest in late-life disorders is the evidence that there is a large paternal age effect for point mutations [66]. Such secular trends could therefore be associated with increases in germline mutations, with potential deleterious effects in subsequent generations. It has been difficult, however, to confirm a relationship of paternal

age and the occurrence of nonfamilial varieties of such common polygenic late-life disorders such as Alzheimer disease or prostate cancer [67].

At a more fundamental level, one can ask why one observes such striking variations in the life spans of various mammalian species. While such variation is obviously related to the constitutional genotype, it does not necessarily follow that aging is “programmed”—at least in the sense of concerted, determinative, sequential gene action comparable to what one observes in development. The most satisfying idea invokes differential impacts of environmental hazards (e.g., accidents, predation, drought, starvation, infectious diseases) during the emergence and maintenance of various species. This is nicely articulated in a popular book on aging [68]. Species with comparatively high hazard functions would be expected to evolve life history strategies that emphasize rapid maturation, high fecundity, early fecundity, and short life spans. An attenuation of those hazards could set the stage for the emergence of sibling species with a more leisurely rate of maturation, lower early fecundity, and longer life spans. One of the few field biology studies to examine this idea has in fact provided strong support for that hypothesis [69]. Contrary findings have been reported, however, for different species in different ecologies [70].

The evolutionary formulations of the nature of aging have a number of interesting and important implications, in addition to those noted above. Let us summarize some of these propositions:

1. Stochastic processes: These are likely to play a major role in senescence. This follows from the conclusion that one is not dealing with a determinative sequence of concerted gene action but rather with an epiphenomenon of selection for gene action designed for reproductive fitness. Consider the analogy with a spacecraft engineered to function for a given period of time in order to complete a specific mission. Engineering specifications for indefinite maintenance of the craft would be prohibitively expensive or impossible. One would therefore anticipate an element of chance as to which components will initially exhibit structural and functional failures and when such failures will be detectable with the available diagnostic facilities. Many major geriatric diseases of humans (e.g., cancer, strokes, coronary thrombosis) are surely based on stochastic events. In the case of malignant neoplasms, selection for a series of random somatic

mutations is the key to the understanding of the pathogenesis, although arguments can be made that the first step in neoplasia may be age-related hyperplasia resulting from variegated gene expression involving cell cycle regulatory loci [71]. Overall longevity also is subject, in part, to stochastic laws. There are numerous examples in which investigators have rigorously controlled both environment and genotype, yet have observed marked variations in longevity. The most convincing example comes from studies of *C. elegans*, which can be grown (except for yeast extracts) in chemically defined media in suspension cultures, thus ensuring rigorous control of the environment [72].

2. Polygenic basis: There is a polygenic basis for aging. There is no reason to believe that the optimization theories or mutation accumulation theories involve only a few genes. Indeed, for the case of the successful experiments involving indirect selection for increased life spans in genetically heterogeneous wildtype stocks of *D. melanogaster*, genetic analysis indicated genes on all of the major chromosomes [73]. Martin [74] estimated (as an upper limit) that allelic variation or mutation at close to 7% of loci of the human genome has the potential to modulate varying aspects of the senescent phenotype. A different and more conservative estimate—the number of genes likely to have evolved in the hominoid lineage leading to humans (and thus could be associated with the increased life span of *H. sapiens*)—gave a figure of 0.6% of functional genes [75]. Neither estimate can be characterized as oligogenic.
3. Multiple mechanisms: There are likely to be multiple mechanisms of aging, although there would be selective pressure toward some degree of synchronization of the ages of expression of phenotypic effects resulting from independent mechanisms. This proposition follows from the randomness of the accumulated constitutional mutations and from the great variety of types of gene action that could be involved in “trade-off” types of gene action. Against this proposition, however, is the fact that a single environmental manipulation—dietary or caloric restriction—regularly leads to lifespan extension in rodents, or at least those that have been selected for the easy life of the laboratory setting [53]. We have little information on the effects of caloric restriction in the wild, however. The life course histories of organisms, whose evolutionary history reflected exceptionally high environmental hazard functions, such as the mice and rats used for most of the calorie-restriction experiments, are quite distinct from those of the higher primates [68]. It is therefore not at all clear that dietary restriction would make a significant impact on the lifespan potentials of human subjects. Evidence is pointing toward the conclusion that dietary restriction may delay the onset of age-associated pathologies and reduce the incidence of common age-related disorders and age-related deaths [76]. We shall have to await the final outcome of current research in rhesus and squirrel monkeys (reviewed by Roth et al. [77]) to know how likely such an effect will obtain for our species. The most recent analysis of the data from two major studies supports the conclusion that this intervention enhances both life span and health span in Rhesus monkeys [78]. Meanwhile, common sense tells us that we should avoid gluttony!
4. Species specificity: There is likely to be a degree of species specificity in relevant gene actions. We have already developed certain of these arguments, but let us consider an extended argument. If aging is in fact an epiphenomenon—a byproduct of selection for alleles ensuring an optimal degree of reproductive fitness in a given environment—there is no a priori reason to expect identical scenarios of gene action among very different species. Consider the striking differences in the behavioral patterns among different species that lead to successful matings. There is surely a wide variety of different loci involved, and those that are operative in fruit flies must surely differ from those that are relevant for man! Nevertheless, it is quite possible that there are a number of common mechanisms among groups of related species (including all mammals), and it is even conceivable that such global mechanisms as oxidative damage to macromolecules underlie the aging of all or most organisms. This is the rationale for carrying out comparative gerontological research. The first “public” mechanisms of aging were documented in experiments in *C. elegans*, *D. melanogaster*, and *M. musculus domesticus* (Fig. 15.5) [48,65]. Remarkably, “leaky” mutations in comparable neuroendocrine signal transduction pathways involving insulin-like growth factor (IGF1) receptors and the nuclear translocation of a transcription factor can

Conserved Nutrient Signaling Pathways Regulating Longevity

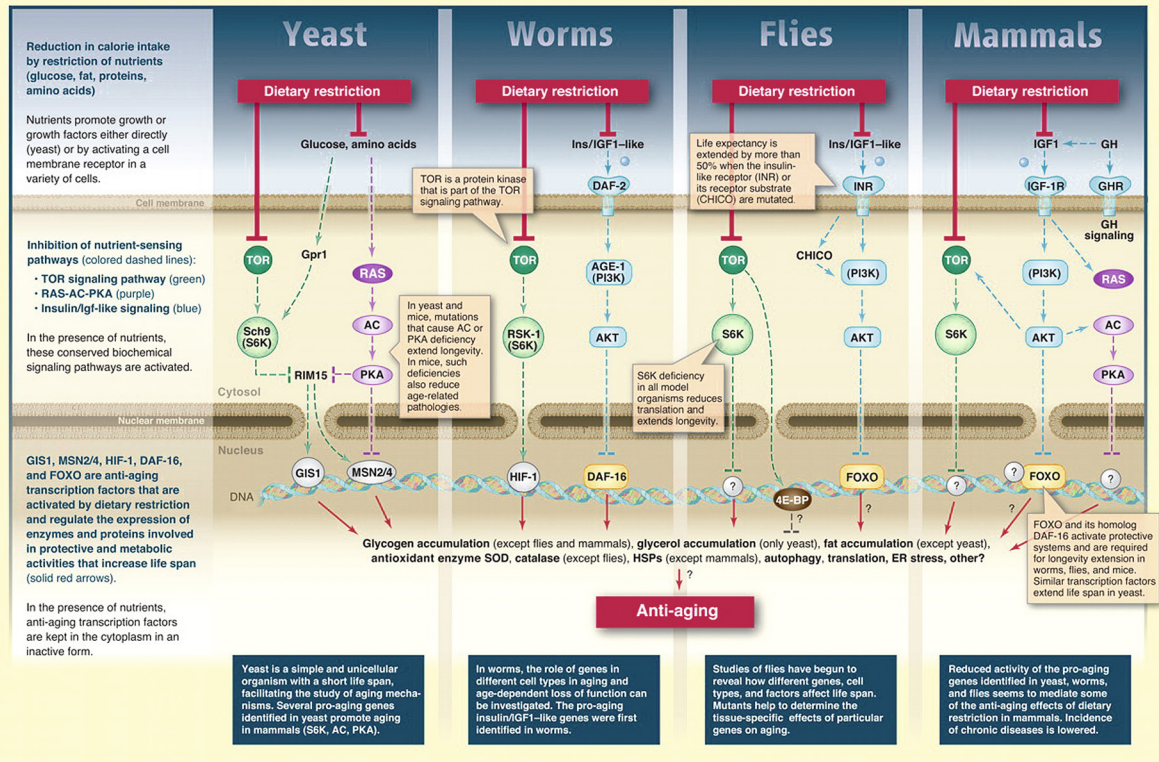


Figure 15.5 Evidence for a partially conserved “public” mechanism for the modulation of aging in yeast, worms, fruit flies, laboratory mice, and, possibly humans. These signal transduction pathways evolved to modulate metabolism under transient adverse environmental conditions, during which gene actions that enhance the protection of somatic cells are up regulated while further growth, development, or reproduction is postponed. Certain mutations in these pathways were found to substantially enhance the life spans of laboratory strains of model organisms amenable to genetic analysis. Mutations so far studied in this general pathway in humans result in pathology. The possible effects of a wider spectrum of such mutations upon the longevity of humans are not yet understood. (From Fontana L, Partridge L, Longo VD. Extending healthy life span from yeast to humans. *Science* 2010;328:321–6.)

lead to substantial extensions of life spans [1]. At first blush, these observations would appear to contradict the conclusion, discussed above, that aging is under polygenic controls and that multiple mechanisms are at work. The authors’ interpretation of these important discoveries, however, is that they are examples of diapauses—nature’s “time-outs” from the business of reproduction when faced with conditions of nutritional, climatic, or other environmental challenges [3]. These can best be considered as “reprieves”; they will be eventually trumped by other gene actions that, unlike diapauses, have escaped the force of natural selection.

5. **Intraspecific variations:** There are likely to be significant intraspecific variations in phenotypic patterns of aging, particularly in humans. This also follows from many of the arguments discussed above. Given the polygenic nature of aging, the likelihood of a variety of mechanisms, a strong stochastic component, the realization that one is dealing with alterations in all body systems, and the enormous genetic and environmental heterogeneity in our species, one would certainly predict substantial differences in the way it plays out in individual subjects. Every clinician has witnessed this phenomenon first-hand. While differential impacts of the environment are likely to

be partially responsible for such variations, the challenge for medical geneticists is to dissect out specific major and minor genetic factors responsible for particularly favorable or unfavorable nature–nurture interactions.

6. Plasticity: The life span of a species should exhibit a degree of plasticity. This follows directly from the arguments on the nature of gene action in aging discussed above and from the experimental results in *Drosophila* sp. Nonetheless, there are likely to be some severe constraints on such plasticity—constraints related to the basic architecture of the organism. We do not expect a fruit fly to live as long as a mouse, without essentially creating a new species.

15.4 HOW DO WE AGE?

We now turn to a more systematic consideration of the present state of our knowledge concerning the underlying molecular mechanisms of aging. In contrast to the reasonably satisfying evolutionary explanations for why we age, there is no consensus as to how we age, although the research programs of a growing number of investigators appear to be motivated by the theory that oxidative damage to macromolecules, including those mediated by chemical-free radicals (the “free radical theory of aging”) (reviewed by Muller et al. [79]), are of paramount importance. Even that canonical theory has likely been oversimplified, however, as alterations in redox signaling in mitochondria are considered to be of importance to mitochondrial dysfunction [80,81] (Fig. 15.6).

15.4.1 Alterations in Proteins

In 1963, Orgel introduced the protein synthesis error catastrophe theory of aging [82,83]. It was proposed that transcriptional and/or translational errors in the synthesis of proteins that were themselves used for the synthesis of proteins (e.g., DNA-dependent RNA polymerases, ribosomal proteins, etc.) could result in an exponential cascade of errors involving essentially all proteins, leading to cell and organismal death. Biosynthetic errors in protein synthesis appear to be rare, however, even in old organisms [84]. Although most gerontologists have abandoned this theory, very few tests of the theory have been carried out with postreplicative cells in vivo [85]. By contrast, there is a growing body of evidence indicating the prevalence of posttranslational modifications in

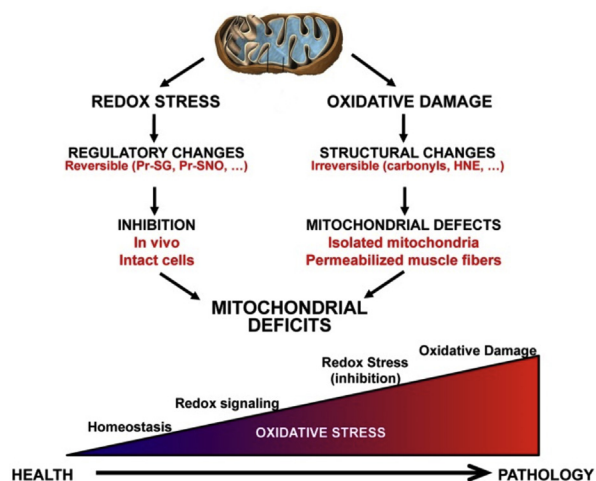


Figure 15.6 A summary of two pathways leading to age-related mitochondrial dysfunction and disease via oxidative stress: Regulatory aberrations may be reversible, whereas structural alterations in macromolecules are likely to be irreversible. (From Marcinek DJ, Siegel MP. Targeting redox biology to reverse mitochondrial dysfunction. *Aging* (Albany NY) 2013;5:588–9.)

proteins in aging tissues; although, of the more than 140 major and minor known modifications of proteins, only a few have been studied in aging cells, tissues, and organisms [86,87]. Beginning with a classic paper on senescent nematodes by Gershon and Gershon [88], many studies have demonstrated an accumulation of immunologically detectable, but enzymatically inactive, enzyme molecules in various mammalian tissues. These may result from a variety of posttranslational modifications, including subtle conformational changes [84]. There is currently a great deal of interest in oxidative alterations [89]. Metal-catalyzed oxidation systems have the potential to inactivate enzymes oxidatively via attacks on the side chains of certain amino acids, with the formation of carbonyl derivatives. The side chains of histidine, arginine, lysine, and proline are particularly susceptible. The sulfhydryl groups of methionine are also susceptible to oxidation. Other posttranslational changes that can be observed in aging cells include racemization, deamidation, isomerization, phosphorylation, and glycation.

Many gerontologists believe that glycation, the spontaneous nonenzymatic reaction of glucose with proteins and nucleic acids, may be a major factor in the development of certain age decrements, as well as complications of diabetes mellitus. Glycation is the slow, spontaneous

reaction of the aldehydic form of glucose with free amino groups to form a Schiff base, which subsequently rearranges to form a stable Amadori product. Subsequent reactions, possibly involving oxygen radicals, generate more complex products referred to as advanced glycosylation end (AGE) products. Some of these compounds, including pentosidine, have been characterized. Antibodies to the AGE products have been generated and used to map their distribution to neuritic plaques and tangles and to other sites [37]. Because the levels of AGE products increase with age and with elevated blood glucose, crosslink proteins, and change their physical and biologic properties, they are thought to underlie the development of atherosclerosis, cataracts, and peripheral neuropathies. In addition, macrophage receptors bind to the AGE products and initiate the secretion of inflammatory cytokines such as the tumor necrosis factor [90]. Thus, glycation represents a progressive age change linked to age-associated disabilities. Support for these ideas has come from experiments in aging dogs, in which it was possible to reverse myocardial stiffness and improve cardiac function by the administration of an experimental compound known to break the crosslinks associated with the formation of advanced glycation end products [91].

Calorically restricted rodents, which have substantially increased life spans, exhibit evidence of both enhanced defenses against reactive oxygen species and reduced levels of protein glycation (associated with decreased levels of plasma glucose). Such results suggest that both the free radical theory of aging and the glycation theory of aging may be operative and potentially synergistic [92]. A number of different types of amyloids accumulate in mammalian tissues during aging [93]. In their advanced states, they are detected extracellularly as protein aggregates associated with proteoglycans and other proteins. Each type is derived from a different precursor protein. These include the beta-amyloid protein of Alzheimer disease and the aging brain; a transthyretin-derived amyloid in peripheral nerve tissues, autonomic nervous system, choroid plexus, cardiovascular system, and kidneys; atrial amyloid derived from the atrial natriuretic peptide; the amylin-derived amyloid in the pancreatic islets of Langerhans [94]; systemic amyloid AA derived from apolipoprotein A-II [95]; and possibly unique types of amyloid in the anterior pituitary gland, intervertebral discs, the aortic intima and media, aortic heart valves, and the adrenal cortex. In certain

of these conditions, variants in the precursor protein greatly accelerate the rates of deposition of the derivative amyloids. This has been particularly well demonstrated for the case of beta amyloid [96].

It is a challenge for the future to discover common denominators underlying this remarkable propensity of mammalian tissues to accumulate these different types of abnormal proteins. Obvious approaches would include more detailed studies of alterations in protein turnover with age (including the turnover of amyloid deposits) and how such turnover might be modulated by endocrine and neuroendocrine factors. Another promising and relatively new area of research seeks to define gene products that function in the repair of altered proteins. An example is the catalysis of the transfer of a methyl group from S-adenosylmethionine to L-aspartyl and D-aspartyl residues by protein carboxyl methyltransferases (ED 2.1.1.77). These enzymes have the potential to repair abnormal proteins via the conversion of L-isoadpartyl residues to L-aspartyl residues [97]. This enzyme is polymorphic in humans, raising the question of the differential repair of such classes of altered proteins during aging in human populations [98].

Research on the maintenance of the integrity of proteins and protein complexes (protein homeostasis or “proteostasis”) is among the fastest-growing fields in geroscience. It is now being pursued in terms of networks of gene actions that modulate protein synthesis, folding, transport, heteromeric protein complex formations, and degradation [99].

15.4.2 Alterations in DNA

15.4.2.1 Nuclear DNA—Epigenetic Events

Given the fact that, for most genetic loci, only two alleles are present, nuclear DNA would appear to be a particularly vulnerable target for damage during aging. Historically, the first specific type of somatic “mutational” theories of aging was proposed by a physicist, Leo Szilard [100]. He envisioned random “hits” that would inactivate entire chromosomes or chromosome arms. In modern terms, such inactivations could conceivably be associated with epigenetic events, as for the case of the random inactivation of one of the two X chromosomes of the human female during embryonic development and the processes of parental genomic imprinting. There is no good evidence of widespread heterochromatinizations or inactivation of large chromatin domains during aging. In fact, at least for the case of mice, there is evidence of a

reactivation of certain gene loci on a previously inactive X chromosome during aging [101–103]. No such reactivation could be demonstrated for the case of the *HPRT* locus of heterozygous human females [104]. Reactivation has also been demonstrated for a genomically imprinted autosomal locus in mice [101]. Global losses of 5-methyl cytosine have been demonstrated in aging fibroblast cultures [105] and in tissues of two species of aging rodents [106], but there have been few studies of altered methylation in specific domains of specific genes during aging. In one such study, hypermethylation was mapped to the proximal 5' spacer domain of ribosomal DNA genes of aging mice; silver stains of cytogenetic preparations revealed that the ribosomal gene cluster on chromosome 16 was preferentially inactivated [106]. It remains to be seen, however, whether this remarkable result reflects some developmental, adaptive process in laboratory mice or in the particular strain of mice investigated, as the biochemical changes were observed as early as 6 months. A form of gene-specific methylation of CpG islands has clearly been established to progress steadily into old age in human subjects. It is associated with the silencing of the estrogen receptor gene of a subset of cells of the colonic mucosa [107]. A striking finding in that study was that, of a set of 45 colorectal human tumors examined, including those in very early stages of oncogenesis, estrogen receptor expression was either diminished or absent. Moreover, the introduction and expression of an estrogen receptor gene in a line of colon carcinoma cells resulted in marked growth suppression. This important paper therefore demonstrates a link between a presumably epigenetically based progressive repression of a specific gene during aging and the susceptibility to the development of a common type of cancer of aging.

Using the yeast model of replicative aging, Lenny Guarente and his colleagues highlighted a key role of NAD-dependent histone deacetylation in the regulation of energy metabolism, genomic silencing, and aging. There continues to be a great deal of current research in various organisms on homologs (sirtuins) of the yeast Sir2 gene responsible for histone deacetylations [108,109]. A variety of other changes in gene expression occur throughout the life span, notably changes in the methylome, but it remains to be seen which of these alterations are of primary significance to one or more aging processes and which are merely epiphenomena. One such approach is to explore the effects of

caloric restriction [110]. A marked transcriptional stress response, with lowered expression of metabolic and biosynthetic genes, is found in aging mouse tissues. These alterations are ameliorated in calorically restricted mice.

Age-related changes of methylation at CpG regions of the genome have received a great deal of attention because of the pioneering research by Steve Horvath and colleagues on the concept of the “Epigenetic Clock” [111]. Patterns of genomic methylation (the “methylome”) predicted longevity within several ethnically distinct populations independent of chronological ages. These results were even more robust when incorporating data on age-related shifts in peripheral blood cell compositions [111].

15.4.2.2 Nuclear DNA—Mutational Events

In 1961, a now classic paper appeared, casting doubt on the validity of somatic mutational theories of aging [112]. Taking advantage of the occurrence of a species of wasp in nature—the males of which exist as either haploid or diploid organisms—it was found that there was no difference in the life spans of these organisms. As expected, however, the haploid wasps were much more susceptible to the effects of ionizing radiation. These results were strong evidence against a role for recessive mutations in insects. They did not rule out, however, some role for a combination of dominant and recessive mutations in the aging of such organisms. Moreover, the interpretations are complicated by the occurrence of polyploid cell types. Finally, those experiments told nothing about the role of somatic mutations of replicating populations of cells in the limitation of life span, since wasps and other insects, with the exception of gonadal tissues and, in *Drosophila*, certain intestinal cells, consist of postreplicative cells. For the case of such replicating populations of cells, there is now compelling evidence that somatic mutations constitute a link between the biology of aging and the biology of cancer. Thus, while more data are required, there are reasonable correlations of species-specific life spans and rates of development of various neoplasms (e.g., [112]). Among those genes that evolved in association with the relatively long life spans of *H. sapiens* (the longest lived of all mammalian species), there must be loci conferring enhanced genomic stability in comparison, for example, with those of *M. musculus domesticus*. Moreover, there may be considerable species differences in the patterns of somatic mutation. For mice, for example,

there is evidence of a marked susceptibility to cytologically detectable chromosomal mutations during aging [113], while in comparable cell types (renal tubular epithelial cells), there is little evidence for the accumulation of mutations (presumably intragenic) at the *HPRT* locus, a target chosen because of the lack of evidence for selection against such mutations in renal tissue [114]. Mutations have been shown to accumulate at the *HPRT* locus of T lymphocytes and in the renal tubular epithelial cells of aging human subjects, however, with higher frequencies of mutation being observed in the epithelial cell type [115,116]. The lymphocyte study showed that deletions were relatively common [117]. Such accumulations could be attributable to chronology rather than to intrinsic biologic aging. We will require additional research in mammals of contrasting lifespan potentials to address this question; an approach using comparable transgenic reporter constructs may be promising, if these could be comparably buffered from position effects [118].

15.4.2.3 Nuclear DNA—Molecular Misreading

What originally appeared to be the accumulation of frameshift mutations in DNA in aging mammals now appears to be the result of transcriptional errors at particularly vulnerable sites, especially those with runs of GAGAG. van Leeuwen has named this phenomenon “molecular misreading.” The process can impact upon the fidelity of transcription of such important loci as the beta-amyloid precursor protein and ubiquitin B [119].

15.4.2.4 Telomeric DNA

Perhaps the most robust age changes noted in the nuclear genome of normal somatic cells are alterations in telomere length, leading to replicative senescence (reviewed by de Lange [120]). Telomeres have a highly repetitive structure (TTAGGG in humans and mice) that extends for many thousands of nucleotides at the ends of chromosomes. The telomeres stabilize chromosomal structure and their loss leads to various cytologic aberrations and the arrest of cell division. Current concepts suppose that telomeres in cells of the germline and in many neoplastic cells are added to the ends of chromosomes *de novo* by a unique enzyme referred to as telomerase, which uses an associated RNA to code for the hexanucleotide repeats. It appears that telomerase is lost or its concentration greatly diminished in the progeny of somatic stem cells (but not in cells of the germline). Somatic cell telomeres are then

duplicated during cell division by DNA polymerases without the assistance of telomerase. It is characteristic of DNA polymerases that they fail to copy some 50–200 terminal bases of the trailing strand and the telomeres are shortened by this amount with every cell division. The shortening of telomeres is strikingly apparent when one examines the telomeric/subtelomeric DNA isolated by appropriate restriction enzyme cleavage from normal human fibroblasts that have undergone large numbers of cell divisions in culture. Exit from the cell cycle may occur as when only a few chromosome arms reach a critical level of shortening, thus activating cell cycle checkpoints [121].

The development of a PCR-based method for the assay of telomere lengths has led to a remarkable association of short telomeres in DNA from peripheral blood of elderly human subjects with mortality; the results were largely attributable to earlier deaths from cardiovascular and infectious diseases [122]. Shorter telomere lengths have also been reported in mothers of chronically ill children; the authors surmised that life stress could shorten life span [123]. An alternative interpretation, however, is that the mothers of many chronically ill children are more frequently exposed to infectious agents, thus driving proliferation of lymphocytes, resulting in shorter telomeres. The relationships between various forms of stress and telomere lengths are the subject of continuing research [124].

15.4.2.5 Mitochondrial DNA

The venerable “Free Radical Theory of Aging” had its origins with the work of Rachel Gershan and Daniel Gilbert on “oxygen toxicity” in 1954 [125,126] and subsequent papers. Harman proposed that aging resulted from the cumulative damaging effects of the byproducts of aerobic metabolism, namely reactive oxygen species (ROS), on mitochondrial DNA itself as well as proteins, leading to cellular deterioration and organ dysfunction. Over the years, advances in understanding mitochondrial genetics and biology has led to a more multifaceted and complex picture of mitochondrial dysfunction and aging, including the role of oxygen-derived free radicals in redox signaling as discussed above and summarized in Fig. 15.6.

mtDNA is closed circular DNA of some 16,569 nucleotides that codes for some of the mitochondrial proteins plus the tRNAs and rRNAs used for mitochondrial protein synthesis. Other components of the mitochondria are coded for by nuclear genes and are transported to

the mitochondria. Essentially, all mtDNA molecules are maternal in origin; thus, mtDNA genetic diseases are maternally transmitted.

It is well known that mtDNA rearrangements (single deletions or multiple deletions) cause numerous mitochondrial diseases, for example, Leber hereditary optic neuropathy (LHON), Kearns–Sayre syndrome (KSS), or progressive external ophthalmoplegia (PEO). In addition, mitochondrial rearrangements are found at low levels in healthy tissues with no signs of mitochondrial disease, and accumulate with aging in postmitotic tissues, where their relationship to aging remains an area of debate. Based on the frequency and consistent location of common specific age-related deletions, one can postulate that the sequences between the direct repeats are looped out following damage to the DNA by a slip replication mechanism. The damage to mtDNA molecules may be initiated by oxygen radicals generated as a byproduct of the oxidative phosphorylation reactions carried out by the mitochondria. The proximity of mtDNA to the sources of oxygen radicals, plus the lack of associated histones, would make mtDNA more vulnerable than nuclear DNA. Thus, the age-related changes observed could be due to increased damage and/or reduced repair. One important mechanism for repair is the proofreading domain of DNA polymerase gamma, the enzyme that replicates mtDNA. Support for the importance of this function has come from the synthesis of mice with knock-in mutations in that domain; these transgenic mice exhibited progeroid features [127].

More recently, a compelling case for a relationship between mitochondrial function and neurodegeneration has been highlighted by the discovery of a number of genes, in which pathogenic variants result in hereditary forms of Parkinson disease. Parkinson disease (PD) is the second most common neurodegenerative disorder after Alzheimer disease, and over 90% of cases are sporadic. Pathogenic variants were first identified in parkin in Japanese families with juvenile PD, and are the most common cause of autosomal recessive PD. Parkin encodes an E3 ubiquitin ligase which is recruited to regions of mitochondrial damage where it ubiquitinates outer mitochondrial membrane proteins. The discovery of *PINK1*, a serine/threonine kinase which phosphorylates a key protein in mitochondrial trafficking, and regulates mitochondrial quality control, added further evidence to the link between mitochondrial dysfunction and Parkinson disease. Specifically, *PINK1* and parkin function to target

and degrade damaged mitochondria through a specific form of autophagy termed mitophagy. A more detailed picture of mitochondrial function in PD pathogenesis has emerged with the discovery of additional hereditary forms of PD which are caused by pathogenic variants in *DJ-1*, *ATP13A2*, *FBX07*, *DNAJC6*, *SYNJ1*, and *PLA2G6* [128]. Collectively, these have implicated mitochondrial transport, fission/fusion, biogenesis, quality control, and mitophagy as underlying causes of neurodegeneration, certainly in hereditary forms, and potentially in sporadic forms of PD. Strategies which enhance parkin expression or increase *PINK1* activity are emerging as promising targets for new therapeutics.

In summary, the link between mitochondrial biology and aging continues to be an important topic, and has expanded beyond the role of reactive oxygen species; currently, multiple aspects of mitochondrial physiology are under investigation including apoptosis, senescence, calcium-dependent signaling, and mitochondrial trafficking and turnover [129]. Age-related alterations in the proteostasis of the heteromeric protein complexes are also worthy of research, particularly given their complex origins from both nuclear and mitochondrial gene products.

15.4.2.6 Germline Mutations

Medical geneticists are well aware of the increased risk to the conceptus of chromosomal types of mutations (mainly aneuploidies) as functions of maternal age. This is, of course, the basis for the clinical practice of counseling women of the availability of prenatal diagnosis. The relationship of paternal age to the increased risk of certain types of mutations has also been well documented and considered to be driven by a variety of mechanisms [130]. These important subjects have in fact received substantially less attention by the gerontological community than the question of somatic mutation and aging.

15.4.3 Alterations in Lipids

Given the seminal importance of membranes in cell biology, alterations in the structure of membrane lipids could constitute a primary mechanism of age-related cellular dysfunction and cell death. Most research in this field has addressed the issue of lipid peroxidation, an integral component of the free radical theory of aging. Aspects of this idea have been discussed above, including the ubiquitous nature of lipofuscin pigments as a biologic marker of aging. A second line of research

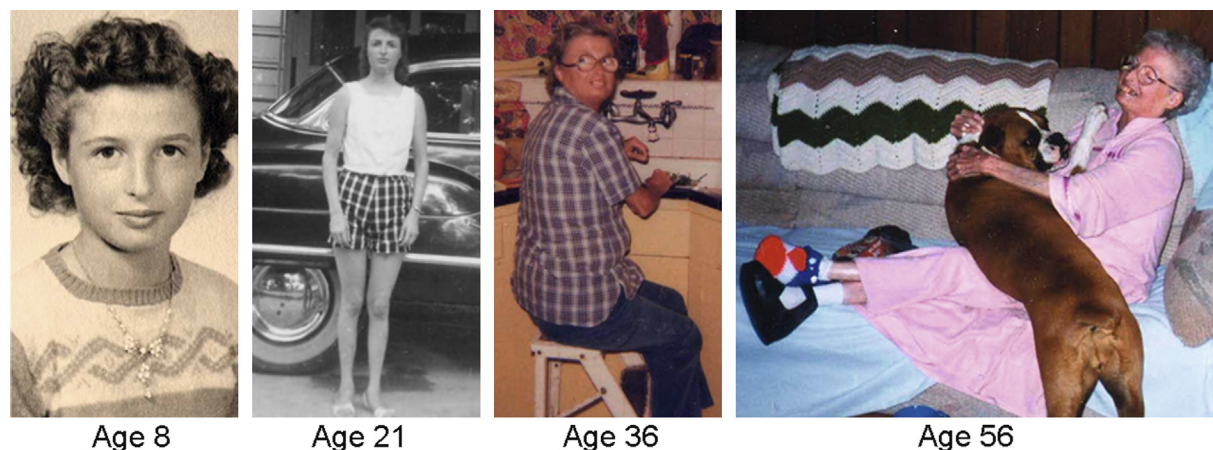


Figure 15.7 Werner syndrome patients with homozygous *WRN* mutations. (From Hisama FM, Bohr VA, Oshima J. *WRN*'s tenth anniversary. *Sci Aging Knowledge Environ* 2006;28:pe18.)

in this field has emphasized age-related increases, in various cell types, of the cholesterol-to-phospholipid ratios of plasma cell membranes, with a consequent decrease in membrane fluidity. At least in some cell types, such as neurons of the dorsal root ganglia, the decline in membrane fluidity, as measured by lateral diffusion coefficients, is related to development rather than to postmaturation aging [131]. Studies of replicative senescence in yeast have emphasized the importance of alterations in the lipid membranes of mitochondria [132], thus integrating this topic with the preceding topic on mitochondrial alterations in aging. Such overlaps of various mechanisms of aging are likely to be the rule rather than the exception.

15.5 PROGEROID SYNDROMES OF HUMANS

Having reviewed the state of our knowledge of the biology and pathobiology of aging, we can now consider spontaneous mutations in humans that may modulate the aging phenotype. As we have seen, however, evolutionary theory would argue that no single mutation or polymorphism is likely to modulate all aspects of the senescent phenotype. In a systematic survey of several editions of McKusick's catalog of the Mendelian inheritance of humans, one of us (GMM) indeed concluded that no single mutation has yet been identified that could be characterized as a global progeria [74]. A number of mutations, however, could be characterized as "segmental progeroid mutations," in

that multiple segments of the complex senescent phenotype appear to have been affected, whereas unimodal syndromes predominantly affect a single organ (e.g., dementias of the Alzheimer type) [74,133]. The responsible mutations include those that impact genomic stability, nuclear structure, numbers of triplet-repeats, and alterations in lipid and carbohydrate metabolism. Some chromosomal aneuploidies (e.g., trisomy 21) also exhibit segmental progeroid features [74]. The two best-known examples of segmental progeroid syndromes are Werner syndrome (WS) and Hutchinson–Gilford progeria syndrome (HGPS), which are discussed in more detail in the following.

15.5.1 Werner Syndrome

The clinical phenotype of WS (OMIM# 277700) has been succinctly summarized as a "caricature of aging" (Fig. 15.7) [134,135]. WS patients usually develop normally until they reach the second decade of life. The first clinical sign is a lack of the pubertal growth spurt during the teen years. In their 20s and 30s, patients begin to exhibit a general appearance of accelerated aging with skin atrophy, loss of subcutaneous fat, and loss and graying of hair. They also develop common age-related disorders including type II diabetes mellitus, bilateral ocular cataracts (requiring surgery at a median age of 30), osteoporosis; gonadal atrophy (with early loss of fertility), premature and severe forms of arteriosclerosis (including atherosclerosis, arteriolosclerosis, and medial calcinosis); and peripheral neuropathy. Multiple

cancers can be observed by middle age [136]. Our survey of WS patients with a molecularly confirmed diagnosis revealed that the prevalence of cataracts was 100% (87/87) [137]. The prevalence of osteoporosis was 91%, hypogonadism 80%, diabetes mellitus 71%, and atherosclerosis 40% at the time of diagnosis. Median age of death in the most recent study was 54 years, a significant increase over what had been observed several decades ago [137–140], perhaps the result of improved medical management. The most common causes of death are myocardial infarction and cancers [137].

Although many clinical features of WS are similar to those observed during “normal” aging, there are significant distinctions. There are a disproportionate number of sarcomas in WS patients: the ratio of mesenchymal cancers to epithelial cancers in WS is approximately 1:1 as compared to 1:10 in the general population [136]. Alzheimer-type dementia is not common in WS [141]. The long bones of the limbs, especially of the lower limbs, are particularly vulnerable to osteoporosis, whereas in ordinary aging the vertebral column is particularly vulnerable, especially in females [142]. There is also a peculiar osteosclerosis of the distal phalanges that is not seen during ordinary aging [143]. Necrotic skin ulcers and necrosis around ankles and occasionally around elbows, which eventually may require amputation, are characteristic to WS, but rarely seen during usual aging.

Classical WS is caused by mutations of the *WRN* gene on chromosome 8. The locus spans approximately 250 kb and consists of 35 exons, 34 of which are protein coding [144]. *WRN* encodes a 180-kDa multifunctional nuclear protein that belongs to the RecQ family of helicases [145]. A structural study revealed a unique interaction between the RecQC-terminal domain of *WRN* protein and the DNA substrates during base separation [146]. In contrast to other members of the RecQ family, *WRN* protein includes an N-terminal domain that codes for exonuclease activity [147]. A single-strand DNA annealing activity in the C-terminal region has also been reported [148]. Its preferred substrates resemble various DNA metabolic intermediates, substrates for which its helicase and exonuclease activities function in a coordinated manner, suggestive of roles in DNA repair, recombination, and replication [149,150]. The *WRN* protein is also involved in telomere maintenance [151], which explains the accelerated telomere shortening of fibroblasts derived from WS patients [152].

To date, more than 80 different *WRN* mutations have been reported, some of which appear to be specific to certain ethnic groups [153]. The majority of these disease mutations result in the truncation of the nuclear localization signal at the C-terminus of the *WRN* protein [154], which makes mutant *WRN* proteins unable to enter the nuclei. This seems to be a satisfactory explanation of why we do not observe noticeable phenotypic differences among various common *WRN* mutations. WS cases are most frequently reported in Japan, where the prevalence of heterozygotic carriers, as estimated from the most common Japanese mutation, was approximately 1/167 [155]. Another region with high incidence of WS is Sardinia, where the prevalence of heterozygous carriers was estimated to be of the order of 1/120 [156]. Frequencies of *WRN* mutations in other populations are unknown, as WS may often escape diagnosis.

Some evolutionary biologists would argue that WS is a poor model of aging, in that it is clear that it would not fit the definition of a set of phenotypes that have escaped the force of natural selection [42].

15.5.2 Hutchinson–Gilford Progeria Syndrome

HGPS (OMIM# 176670) is a childhood-onset progeria (Fig. 15.8). It was first described by Jonathan Hutchinson [157] in a boy with baldness and atrophic skin. Hastings Gilford [158] then described a patient with accelerated aging and ateleiosis who died with symptoms of angina pectoris at age 18. HGPS patients are typically normal at birth, but their growth soon falls below the normal range within the first 3–6 months of life. In addition, the children develop accelerated, degenerative changes of the cutaneous, musculoskeletal, and cardiovascular systems [159–161]. The pathognomonic appearance develops usually by age 2 years, and includes: baldness with loss of eyelashes and eyebrows, prominent eyes, convex nasal bridge, small jaw, and generalized loss of subcutaneous fat, resulting in an overall prematurely aged appearance. Historically, most patients survived only to the early teens, with a median age of death in HGPS patients of 16.4 years. Virtually all the patients succumb to myocardial infarction, strokes, or congestive cardiac failure [162]. This is in contrast to patients with the adult-onset progeroid syndrome, WS, whose onset is after puberty and who live until the sixth decade or beyond. Malignancies, ocular cataracts, and Alzheimer-type dementia are also not commonly seen in HGPS, perhaps because they die at such early ages.



Figure 15.8 Hutchinson–Gilford progeria syndrome. (Courtesy of the Progeria Research Foundation.)

HGPS is caused in nearly all cases by *de novo* heterozygous mutations in *LMNA*, which encodes nuclear intermediate filaments, lamin A and C [163,164]. Lamin A and C, generated by alternative splicing of *LMNA*, undergo dimerization and head-to-tail assembly to form nuclear lamina that lies on the inner surface of the inner nuclear membrane [165]. Point mutations within *LMNA* exon 11 found in HGPS create a cryptic splicing site and generate a 50-amino-acid in-frame deletion that includes the proteolytic site required for the maturation of progerin to lamin A [163,164]. Unlike wild-type lamin A, this in-frame deletion mutant, termed progerin, retains the farnesyl moiety at its C-terminus. The resulting accumulation of progerin is thought to be responsible for the phenotypic presentation of HGPS [166]. At a cellular level, the presence of progerin is shown to cause structural abnormalities and/or fragility of nuclei [167], aberrant reorganization of the heterochromatin and epigenetic changes [168], genomic instability [169], and impaired telomere maintenance [170]. Age-associated accumulation of small amounts of progerin has been demonstrated in human fibroblasts and coronary arteries, suggesting the possibility that progerin may be, in part, involved in development of the age-related pathologies in normal individuals [162,171].

LMNA mutations are also responsible for a group of disorders, termed laminopathies, including Emery–Dreifuss muscular dystrophy, dilated cardiomyopathy type 1A (DCM1A) with or without atrioventricular

conduction disturbance, limb-girdle muscular dystrophy type 1B (LGMD1B), Charcot–Marie–Tooth disease type 2 (CMT2), Dunnigan-type familial partial lipodystrophy, mandibuloacral dysplasia, restrictive dermopathy (RD), and atypical forms of Werner syndrome [172]. The atypical Werner syndrome patients have tested negative for biallelic pathogenic *WRN* variants, have short stature, and adult onset of progeroid features, but with accelerated cardiovascular disease in adulthood. Two such patients were investigated and found to express progerin, albeit at much lower levels than found in classic HGPS patients [173].

Farnesyltransferase inhibitors have been shown to ameliorate HGPS phenotypes in cell cultures and in mouse models [174–177]. Results of a clinical trial in humans were published in 2018 [178,179]. The rationale for this approach has been challenged, however, by the finding that nonfarnesylated progerin can elicit HGPS-like phenotypes in mice [180].

15.5.3 MDPL Syndrome

MDPL syndrome (mandibular hypoplasia, deafness, progeroid features, lipodystrophy) usually presents in the first or second decades of life [181]. MDPL patients begin to develop prominent loss of subcutaneous fat, a characteristic facial appearance, metabolic abnormalities, including diabetes mellitus, and progeroid features. Sensorineural deafness is seen in most cases. Undescended testes and hypogonadism have been reported in males but females may be fertile.

MDPL syndrome is an autosomal dominant disorder caused by heterozygous germline pathogenic variants in the *POLD1* gene. *POLD1* encodes one of the main replicative polymerases, which contains an intrinsic exonuclease domain and interacts with WRN protein [182]. Its additional role is in postreplication repair of the lagging strand as a translesion synthesis (TLS) polymerase. The most common *POLD1* variant found in MDPL patients is a deletion (p.S605del) within the polymerase domain. A single missense mutation located in its exonuclease domain has also been identified [181,183,184]. Interestingly, several germline pathogenic variants in *POLD1* exonuclease domains are also known to predispose to cancers, particularly familial colorectal cancers [185].

15.5.4 Rare Genomic Instability Disorders Resulting in Segmental Progeroid Phenotypes

The role of genomic instability in producing a wide variety of segmental progeroid phenotypes has been revealed by additional rare genetic syndromes, a few illustrative examples of which are given here.

Ataxia-telangiectasia (AT). The progeroid features caused by biallelic pathogenic ATM gene variants in AT individuals include: an increased cancer risk, graying hair, immunodeficiencies, reduced fertility, and neurological signs, such as ataxia and oculomotor apraxia. Women who are heterozygous for ATM gene pathogenic variants are at usually moderately increased risk of breast cancer. Cultured cells from AT patients exhibit an increased frequency of spontaneous chromosomal anomalies including double-stranded breaks and telomeric abnormalities, and demonstrate a characteristic and marked sensitivity to the cytotoxic effects of ionizing radiation.

Cockayne syndrome (CS). Loss of subcutaneous adipose tissue, hypertension, atherosclerosis and arteriolosclerosis, age-related renal pathology, and cognitive decline are among the clinical features of Cockayne syndrome, which contribute to its inclusion as a segmental progeroid syndrome [186]. Mutations in at least five loci have been associated with CS—CSA, CSB, XPB, XPD, and XPG—thus documenting pathogenic overlaps with xeroderma pigmentosa [187]. About two-thirds of Cockayne patients have mutations at CSB and about one-third at the CSA locus [186]. The underlying pathogenesis involves defects in

transcription-coupled excision repair of DNA [188]. Oxidative stress and mitochondrial dysfunction in Cockayne syndrome have been shown to be related to deletion of the catalytic subunit of DNA polymerase gamma, the enzyme responsible for replicating mitochondrial DNA. That deletion was associated with the accumulation of a serine protease; of great potential therapeutic significance, the phenotype could be reversed by a serine protease inhibitor [189].

Xeroderma pigmentosum. Xeroderma pigmentosum (XP) is a group of autosomal recessive disorders with cardinal features of sensitivity to sunlight, marked predisposition to skin cancer (>1000-fold increased risk of both melanoma and nonmelanoma-type cancers) beginning in childhood, and cutaneous abnormalities including atrophy, telangiectasias, actinic keratosis, and pigmentary changes. Ocular features include: corneal abnormalities, visual impairment, and tumors. Neurological abnormalities are seen in a small subset of patients, and features include microcephaly, intellectual disability, hearing loss, and impaired motor function.

XP results in most cases from a defect in nucleotide excision repair (NER), the major DNA repair mechanism to remove helix distorting lesions such as UV-induced pyrimidine dimers, and bulky adducts induced by certain chemicals. NER requires the coordination of more than 30 polypeptides acting in two NER subpathways: global genome repair, and transcription-coupled repair.

Xeroderma pigmentosum was defined classically by eight complementation groups (XPA–XPG, and XPV), however, the discovery of the underlying molecular genetic causes of XP has revealed significant phenotypic overlap with other defined disorders of DNA repair such as Cockayne syndrome, Fanconi anemia, trichothiodystrophy, and cerebro-oculo-facio-skeletal syndrome (COFS).

XPF/progeroid syndrome. A single patient has been reported who was normal at birth except for marked sun sensitivity, but then had learning disability, hearing loss, optic atrophy by age 6 years, a progeroid appearance by age 10 years [190]. As a teen, he was found to have microcephaly, renal insufficiency, hypertension, and dry atrophic, irregularly pigmented skin with sunburn, but without skin cancer. Skin fibroblasts showed a severe reduction in UV-induced DNA damage repair, consistent with a diagnosis of xeroderma pigmentosum, but the mild

skin findings and progeroid features with multiorgan involvement were not characteristic. The parents were consanguineous, and the patient was found to be homozygous for p.R153P in the ERCC4 gene. It is thought that the p.R153P variant results in cell death, which allows progeroid features, but milder variants result in accumulation of somatic mutations, and the development of cancer, rather than cell death.

COATS plus disease. COATS plus disease, also called cerebroretinal microangiopathy with calcifications and cysts-1 (CRMCC1), is a developmental disorder characterized by intracranial calcifications, leukoencephalopathy, and retinal telangiectasia (COATS disease). It is caused by biallelic mutations of CTC1 gene that encodes the conserved telomere maintenance component 1 [191]. While null or truncated CTC1 mutations are generally associated with lethality, patients with missense mutations may present with a combination of milder phenotypes including progeroid features and recurrent fractures [192].

15.5.5 Disorders of Lipid and Carbohydrate Metabolism Resulting in Segmental Progeroid Phenotypes

Generalized lipodystrophies can be genetic or acquired, and are associated with profound metabolic disturbances that increase in prevalence with age in the normal population including: insulin resistance, fatty liver, hypertriglyceridemia, and type II diabetes mellitus. Lipodystrophies therefore warrant consideration as segmental progeroid syndromes.

A valuable review of the pathophysiologies of a wide range of both genetic and acquired lipodystrophies has recently been published, including several genetic variants of the Seip syndrome [193]. All currently recognized forms of the latter are autosomal recessive in nature. The type 1 disorder is due to mutations at *AGPAT2*, which codes for 1-acylglycerol-3-phosphate O-acyltransferase-2, an enzyme involved in de novo phospholipid biosynthesis. Type 2 is caused by mutations at *BSCL2* (Berardinelli–Seip congenital lipodystrophy 2, also known as Seipin), which codes for a transmembrane protein that, like *AGPAT2*, is localized to the endoplasmic reticulum and participates in the control of lipid droplet formation and adipocyte differentiation. The type 3 disorder involves mutations at *CAVI*, or caveolin 1, a plasma membrane scaffolding protein and oncogene. Finally, the type 4 disorder

is associated with mutations at *PTRF*, coding for the polymerase I and transcript release factor, which is required for dissociation of a transcription complex and is also involved in the organization of the caveolae of plasma membranes. These mutations result in variable expressions of striking losses of normal adipose tissue and abnormal accumulations of lipids in various viscera, including skeletal muscle, liver, and heart. In addition to the regional atrophy of subcutaneous tissues that is so common in normative aging, one also observes type II diabetes mellitus [194], cardiovascular lesions [195] often associated with lipid abnormalities, sometimes with multiple xanthomas [196], psychomotor abnormalities [197] and what some regard as secondary abnormalities of mitochondrial oxidative phosphorylation [198]. Gastrointestinal polyps, a common benign feature of normative aging, have also been observed [199].

Like all segmental progeroid syndromes, there are of course discordances with what one observes in normative aging, the most dramatic of which is the striking muscular hypertrophy associated with Berardinelli–Seip congenital lipodystrophy.

15.5.6 Miscellaneous Disorders Resulting in Segmental Progeroid Phenotypes

A number of additional genetic disorders result in a segmental progeroid phenotype, including myotonic dystrophy (cataracts and muscle atrophy), and trisomy 21 (premature Alzheimer disease and prematurely aged appearance). Type II diabetes is a common, complex disorder with a dramatic increase in prevalence in the past two generations in industrialized countries, and in recent years, is becoming an increasingly common diagnosis in the pediatric population. The rapid change in the prevalence cannot be attributed to underlying genetic alterations, but rather largely to lifestyle and environmental factors, including sugar-laden foods and beverages, and sedentary lifestyle. The reason it is worth bringing to attention here is that type II diabetes has premature aging effects on many organs including: the cardiovascular system, the renal system, the nervous system, and causes retinopathy. It is entirely possible that the cumulative effects of a large population of children and young adults with many more years of type II diabetes than previously could result in shortened life spans and health spans in the next few generations.

15.6 PRO-LONGEVITY LOCI AND “ANTIGEROID” SYNDROMES

Similar to the absence of a global “progeroid syndrome”, i.e., one that recapitulates *all* of the features of usual aging, there is also no global “antigeroid” syndrome yet discovered in humans. Given the discussion above on the polygenic mechanisms of aging, it would seem unlikely that allelic variants at a single locus would lead to such a syndrome.

It is the case, however, that there are many human subjects who remain healthy, cognitively and physically active, well into their 80s, 90s, and beyond. These have included research subjects recruited by the New England Centenarian study and the Institute of Aging Research Longevity Genes Project of the Albert Einstein College of Medicine.

Unlike the case for model organisms such as *C. elegans*, in which single gene variants can lead to a doubling of life span, no such rare variants of large effect have been discovered to date among populations of human centenarians. Nonetheless, there have been some tantalizing genetic associations. For example, a deletion of exon 3 (d3) in the human growth hormone receptor is a polymorphism found in approximately 25% of the population. The frequency of homozygosity for this allele is tripled from 4% in controls to 12% in male centenarians. Multivariate regression analysis indicated that the d3/d3 genotype increased life span by 10 years. Given the role of growth hormone and insulin-like growth factor signaling in the regulation of life spans of a number of species, this finding supports a similar role in our own species. Surprisingly, however, no enrichment of this allele was found among female centenarians [200].

15.6.1 Dementias of the Alzheimer Type (DAT)

The presence of one or two *APOE4* alleles has been shown to be by far the major genetic risk factor for sporadic, late-onset forms of DAT. The complementary observation that the *APOE2* allele is protective, however [201], has received comparatively much less attention, as evidenced by searches of PubMed for “*APOE4* and Alzheimer’s disease” versus “*APOE2* and Alzheimer’s disease.” As of this writing, those numbers are, respectively, 3201 versus 432. This, plus the fact that it is so difficult to find other well-established examples of unimodal antigeroid alleles, supports the need for much more research on this topic. For the present example, the

importance of such research is not only relevant to the disease entities in question (DAT), it is also relevant to the broader issue of gene actions related to the heritability of longevity. The *APOE2* allele is among those which contribute to this heritability, which has been estimated to be of the order of 25–33% [202].

A number of suggested mechanisms of gene action have been reviewed for the *APOE2* allele [202b]. Of particular interest are gene actions that impact upon the structure and function of dendrites and possible antioxidant functions.

15.6.2 Atherosclerosis

Although the *APOE2* allele has the potential to enhance longevity (Suri et al., 2013), it is perhaps surprising that homozygosity for that allele has been associated with dysbetalipoproteinemia (type III hyperlipoproteinemia), a disorder that accelerates atherogenesis. The common E3 allele might therefore be considered to protect from atherogenesis.

There are many potentially fruitful areas of research regarding gene actions that protect human subjects from atherosclerosis, a major contributor to death from cardiovascular diseases. This has been well established, most notably at the *PCSK9* locus. Pathogenic missense variants in *PCSK9*, which encodes proprotein convertase subtilisin kexin type 9, a serine protease, were reported to result in a rare familial form of hypercholesterolemia in 2003 [203]. The mechanism was thought to be a gain of function which promotes degradation of the hepatic low-density lipoprotein (LDL) receptors, which normally act to clear circulating LDL. This observation suggested the possibility of loss-of-function variants in *PCSK9* which would be predicted to increase total cell surface LDL receptors and result in decreased cholesterol. In fact, two premature truncating variants were discovered in ~2% of African-Americans with low plasma LDL, and associated with 30–40% reductions in LDL and an 80% reduction in coronary artery disease [204]. In a relatively short period of time, this discovery led to the development of a new class of effective (albeit expensive) cholesterol-lowering drugs: monoclonal antibodies against *PCSK9*.

This success story of the role of human genetics in leading to development of a novel drug focused attention on another potential target: *ANGPTL3*, which encodes angiopoietin-like protein 3. Investigation of a family with loss-of-function variants in *ANGPTL3*

found affected members had combined hypolipidemia with reduced triglycerides, as well as low LDL and HDL cholesterol, and resistance to atherosclerosis [205]. Subsequently, in clinical trials with monoclonal antibodies against ANGPTL3 or antisense oligonucleotides against ANGPTL3 mRNA, both approaches resulted in dramatic reductions in triglycerides, and significant reductions in cholesterol [206,207].

Finally, systematic studies of “human knockouts” in consanguineous populations have identified individuals homozygous for loss-of-function (LOF) variants in APOC3 encoding apolipoprotein C3. Deep phenotyping of one family in which both parents and all of their children were homozygous for LOF APOC3 showed absent plasma apoC3 protein, lower triglycerides, higher HDL cholesterol and similar LDL cholesterol, and blunting of the postprandial rise in triglycerides after a fatty meal [208].

15.6.3 Genetic Resistance to Environmental Carcinogens

There is no doubt that one of the secrets to the avoidance of cancer (especially lung cancer) and to increasing one's chances of living to the 10th and 11th decades of life is to avoid cigarette smoke [209]. But why do some heavy cigarette smokers live well into their 10th and 11th decades free of lung cancer [210]? There is a very large literature on candidate polymorphic variants that can provide protection against some of the large numbers of carcinogenic compounds in cigarette smoke, a review of which is beyond the scope of this chapter. Polygenic models [211,212] are likely to be the most satisfactory approaches to uncovering various patterns of resistance and susceptibility as we approach the era of whole-genome sequencing and precision medicine [213].

15.6.4 Human Allelic Variants Homologous to Pro-Longevity Genes in Model Organisms

There has been a surge of interest in testing the hypothesis that the ability to achieve remarkable longevity in centenarians is due to the inheritance of alleles at a few loci of major relevance. A priori, one would predict that such research would be quite risky, given the arguments made earlier in this chapter that life span is under highly polygenic modulations and that it is also determined, in part, by stochastic events. It is the case, however, also noted above, that atherosclerosis (and the associated heart attacks and strokes) is a major limitation of human

life span in developed societies. Therefore, it is perhaps not surprising that an association of unusual longevity with variant alleles for lipoprotein metabolism has been observed [214]. More recent studies demonstrated the association of polymorphisms in the forkhead box class O (FOXO) family of transcription factors among several independent centenarian populations [215]. The FOXO genes are key regulators of the insulin-IGF1 signaling pathway (Fig. 15.5).

There has also been considerable interest lately in Laron dwarfism because of their mutations in the growth hormone pathway [216,217]. People with Laron dwarfism in Ecuadorian villages are resistant to cancer and diabetes and are somewhat protected against aging. This is consistent with findings in mice with a defective growth hormone receptor gene, suggesting this “public mechanism” of aging may apply to our species [218].

15.7 CONCLUSIONS AND FUTURE DIRECTIONS

The careful phenotypic characterization of both segmental progeroid syndromes [74] and unimodal progeroid syndromes [219] by medical geneticists and others have greatly contributed to the development of various hypotheses of gene actions involved in biological aging and geriatric disorders. Regarding fundamental mechanisms of aging, the segmental progeroid syndromes have provided particularly strong support for the role of genomic instability [220]. This mechanism of aging may have deep evolutionary roots [221]. Regarding specific geriatric disorders, very specific pathogenetic pathways have been discovered, a cogent example of which is the elucidation of the role of beta amyloid in the pathogenesis of all forms of dementias of the Alzheimer type [222]. It is important to point out, however, that the origins of that hypothesis did not come from the study of the common sporadic, late-onset forms of the disorder, but were the result of studies by medical geneticists of rare pedigrees with autosomal dominant mutations in that pathway leading to early-onset forms of the disorder [96]. There is an important lesson here for investigators interested in other common geriatric disorders.

Regarding future directions by medical geneticists interested in the pathobiology of aging and age-directed disorders, we wish to emphasize what has been a comparatively neglected approach—a genetic analysis of individuals exhibiting unusual resistance to segmental

or unimodal patterns of aging and age-related disorders—i.e., the search for antigeroid allelic variants. Given the classical evolutionary biological theories of aging [54,57], it would seem prudent to carry longitudinal studies of phenotypes that begin to emerge after the steep decline in the force of natural selection—i.e., beginning in early middle age [2].

Medical geneticists have great opportunities to capitalize upon the increasing pace of our understanding of the epigenome, including the development of tools for the epigenetic analysis of single cells [223]. These have the potential to elucidate stochastic variations in gene expression during aging [71].

Finally, although we have been focusing upon phenotypes in middle and old age, it will be important to keep in mind that fact that how well one builds an organism makes a great deal of difference in how well that organism functions and how long it lasts. Geroscientists, including geneticists interested in variations in rates and patterns of aging, should certainly not neglect developmental biology. This research should enthusiastically embrace epigenetic research on intergenerational and transgenerational inheritance [224]; it would be hard to overemphasize the public health significance of such research.

REFERENCES

- [1] Mazucanti CH, Cabral-Costa JV, Vasconcelos AR, Andreotti DZ, Scavone C, Kawamoto EM. Longevity pathways (mTOR, SIRT, insulin/IGF-1) as key modulatory targets on aging and neurodegeneration. *Curr Top Med Chem* 2015;15:2116–38.
- [2] Martin GM. Help wanted: physiologists for research on aging. *Sci Aging Knowledge Environ* 2002;vp2.
- [3] Martin GM. Modalities of gene action predicted by the classical evolutionary biological theory of aging. *Ann N Y Acad Sci* 2007;1100:14–20.
- [4] Treaster SB, Chaudhuri AR, Austad SN. Longevity and GAPDH stability in bivalves and mammals: a convenient marker for comparative gerontology and proteostasis. *PLoS One* 2015;10:e0143680.
- [5] Kowald A, Kirkwood TB. Can aging be programmed? A critical literature review. *Aging Cell* 2016.
- [6] Gavrilov LA, Gavrilova NS. The biology of life span: a quantitative approach. New York: Harwood Academic; 1991.
- [7] Gavrilov LA, Gavrilova NS. The quest for a general theory of aging and longevity. *Sci Aging Knowl Environ* 2003;2003:RE5.
- [8] Gompertz B. On the nature of the function expressive of the law of human mortality and on a new mode of determining life contingencies. *Philos Trans R Soc Lond Biol Ser A* 1825;115:513–85.
- [9] Makeham WM. On the law of mortality and the construction of annuity tables. *J Inst Actuar* 1860;8: 301–10.
- [10] Sacher GA. Evolution of longevity and survival characteristics in mammals. In: Schneider EL, editor. The genetics of aging. New York: Plenum Press; 1978.
- [11] Milne EM. Dynamics of human mortality. *Exp Gerontol* 2010;45:180–7.
- [12] Carey JR, Liedo P, Muller HG, Wang JL, Vaupel JW. Dual modes of aging in Mediterranean fruit fly females. *Science* 1998;281:996–8.
- [13] Vaupel JW, Carey JR, Christensen K, Johnson TE, Yashin AI, Holm NV, Iachine IA, Kannisto V, Khazaeli AA, Liedo P, Longo VD, Zeng Y, Manton KG, Curtsinger JW. Biodemographic trajectories of longevity. *Science* 1998;280:855–60.
- [14] Robine JM, Allard M. The oldest human. *Science* 1998;279:1834–5.
- [15] Scock NW. Systems integration. In: Finch CE, Hayflick L, editors. Handbook of the biology of aging. New York: Van Nostrand-Reinhold; 1977.
- [16] Schulz R, Curnow C. Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf. *J Gerontol* 1988;43:P113–20.
- [17] Breteler MM, Claus JJ, van Duijn CM, Launer LJ, Hofman A. Epidemiology of Alzheimer's disease. *Epidemiol Rev* 1992;14:59–82.
- [18] Ritchie K, Kildea D. Is senile dementia “age-related” or “ageing-related”?—evidence from meta-analysis of dementia prevalence in the oldest old. *Lancet* 1995;346:931–4.
- [19] Katzman R, Kawas C. Risk factors for Alzheimer's disease. *Neurosci News* 1998;1:27–34.
- [20] Nelson EA, Dannefer D. Aged heterogeneity: fact or fiction? The fate of diversity in gerontological research. *Gerontol* 1992;32:17–23.
- [21] Lindeman RD. Is the decline in renal function with normal aging inevitable? *Geriatr Nephrol Urol* 1998;8:7–9.
- [22] Schachter F, Faure-Delanef L, Guenot F, Rouger H, Froguel P, Lesueur-Ginot L, Cohen D. Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet* 1994;6:29–32.
- [23] Higgins GA, Large CH, Rupniak HT, Barnes JC. Apolipoprotein E and Alzheimer's disease: a review of recent studies. *Pharmacol Biochem Behav* 1997;56: 675–85.

- [24] Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 1993;261:921–3.
- [25] Yu JT, Tan L, Hardy J. Apolipoprotein E in Alzheimer's disease: an update. *Annu Rev Neurosci* 2014;37:79–100.
- [26] Carmeli E, Reznick AZ. The physiology and biochemistry of skeletal muscle atrophy as a function of age. *Proc Soc Exp Biol Med* 1994;206:103–13.
- [27] Gonzalez-Freire M, de Cabo R, Studenski SA, Ferrucci L. The neuromuscular junction: aging at the crossroad between nerves and muscle. *Front Aging Neurosci* 2014;6:208.
- [28] Wall BT, Dirks ML, van Loon LJ. Skeletal muscle atrophy during short-term disuse: implications for age-related sarcopenia. *Ageing Res Rev* 2013;12:898–906.
- [29] Lopez-Torres M, Perez-Campo R, Fernandez A, Barba C, Barja de Quiroga G. Brain glutathione reductase induction increases early survival and decreases lipofuscin accumulation in aging frogs. *J Neurosci Res* 1993;34:233–42.
- [30] Martin GM. Interactions of aging and environmental agents: the gerontological perspective. *Prog Clin Biol Res* 1987;228:25–80.
- [31] Martin GM. Cellular aging—postreplicative cells. A review (Part II). *Am J Pathol* 1977;89:513–30.
- [32] Nakano M, Mizuno T, Gotoh S. Accumulation of cardiac lipofuscin in crab-eating monkeys (*Macaca fascicularis*): the same rate of lipofuscin accumulation in several species of primates. *Mech Ageing Dev* 1993;66:243–8.
- [33] Katz ML, White HA, Gao CL, Roth GS, Knapka JJ, Ingram DK. Dietary restriction slows age pigment accumulation in the retinal pigment epithelium. *Invest Ophthalmol Vis Sci* 1993;34:3297–302.
- [34] Rao G, Xia E, Nadakavukaren MJ, Richardson A. Effect of dietary restriction on the age-dependent changes in the expression of antioxidant enzymes in rat liver. *J Nutr* 1990;120:602–9.
- [35] Kohn RR. Extracellular aging. In: Kohn RR, editor. *Principles of mammalian aging*. Englewood Cliffs, NJ: Prentice-Hall; 1978.
- [36] Sell DR, Monnier VM. Conversion of arginine into ornithine by advanced glycation in senescent human collagen and lens crystallins. *J Biol Chem* 2004;279:54173–84.
- [37] Yan SF, Ramasamy R, Naka Y, Schmidt AM. Glycation, inflammation, and RAGE: a scaffold for the macrovascular complications of diabetes and beyond. *Circ Res* 2003;93:1159–69.
- [38] Saudek DM, Kay J. Advanced glycation endproducts and osteoarthritis. *Curr Rheumatol Rep* 2003;5:33–40.
- [39] Barnes 2nd RH, Akama T, Ohman MK, Woo MS, Bahr J, Weiss SJ, Eitzman DT, Chun TH. Membrane-tethered metalloproteinase expressed by vascular smooth muscle cells limits the progression of proliferative atherosclerotic lesions. *J Am Heart Assoc* 2017;6.
- [40] Bilato C, Crow MT. Atherosclerosis and the vascular biology of aging. *Ageing* 1996;8:221–34.
- [41] Charlesworth B. *Evolution in age-structured populations*. Cambridge: Cambridge University Press; 1980.
- [42] Rose MR. *Evolutionary biology of aging*. New York: Oxford University Press; 1991.
- [43] Diamond JM. Big-bang reproduction and ageing in male marsupial mice. *Nature* 1982;298:115–6.
- [44] Hamilton WD. The moulding of senescence by natural selection. *J Theor Biol* 1966;12:12–45.
- [45] Martin GM, Austad SN, Johnson TE. Genetic analysis of ageing: role of oxidative damage and environmental stresses. *Nat Genet* 1996;13:25–34.
- [46] Baudisch A. Hamilton's indicators of the force of selection. *Proc Natl Acad Sci U S A* 2005;102:8263–8.
- [47] Hayflick L, Moorhead PS. The serial cultivation of human diploid cell strains. *Exp Cell Res* 1961;25:585–621.
- [48] Fontana L, Partridge L, Longo VD. Extending healthy life span—from yeast to humans. *Science* 2010;328:321–6.
- [49] Uno M, Nishida E. Lifespan-regulating genes in *C. elegans*. *NPJ Aging Mech Dis* 2016;2:16010.
- [50] Walker DW, McColl G, Jenkins NL, Harris J, Lithgow GJ. Evolution of lifespan in *C. elegans*. *Nature* 2000;405:296–7.
- [51] Briga M, Verhulst S. What can long-lived mutants tell us about mechanisms causing aging and lifespan variation in natural environments? *Exp Gerontol* 2015;71:21–6.
- [52] Johnson JE, Johnson FB. Methionine restriction activates the retrograde response and confers both stress tolerance and lifespan extension to yeast, mouse and human cells. *PLoS One* 2014;9:e97729.
- [53] Masoro EJ. Dietary restriction-induced life extension: a broadly based biological phenomenon. *Biogerontology* 2006;7:153–5.
- [54] Medawar PB. *An unsolved problem of biology*. London: HK Lewis; 1952.
- [55] Warby SC, Graham RK, Hayden MR. *Huntington disease*. GeneReviews. Seattle: University of Washington; 2010.
- [56] Haldane JBS. *New paths in genetics*. New York: Harper and Brothers; 1942.
- [57] Williams GC. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 1957;11:398–411.
- [58] Martin GM. Atherosclerosis is the leading cause of death in the developed societies. *Am J Pathol* 1998;153:1319–20.

- [59] Martin GM. APOE alleles and lipophilic pathogens. *Neurobiol Aging* 1999;20:441–3.
- [60] Campisi J. Aging, tumor suppression and cancer: high wire-act!. *Mech Ageing Dev* 2005;126:51–8.
- [61] Cicin-Sain L, Messaoudi I, Park B, Currier N, Planer S, Fischer M, Tackitt S, Nikolich-Zugich D, Legasse A, Axthelm MK, Picker LJ, Mori M, Nikolich-Zugich J. Dramatic increase in naive T cell turnover is linked to loss of naive T cells from old primates. *Proc Natl Acad Sci U S A* 2007;104:19960–5.
- [62] Basisty N, Dai DF, Gagnidze A, Gitari L, Fredrickson J, Maina Y, Beyer RP, Emond MJ, Hsieh EJ, MacCoss MJ, Martin GM, Rabinovitch PS. Mitochondrial-targeted catalase is good for the old mouse proteome, but not for the young: ‘reverse’ antagonistic pleiotropy? *Aging Cell* 2016;15:634–45.
- [63] Kirkwood TB, Rose MR. Evolution of senescence: late survival sacrificed for reproduction. *Philos Trans R Soc Lond B Biol Sci* 1991;332:15–24.
- [64] Partridge L, Barton NH. Optimality, mutation and the evolution of ageing. *Nature* 1993;362:305–11.
- [65] Partridge L, Gems D. Mechanisms of ageing: public or private? *Nat Rev Genet* 2002;3:165–75.
- [66] Crow JF. Spontaneous mutation in man. *Mutat Res* 1999;437:5–9.
- [67] Jung A, Schuppe HC, Schill WB. Are children of older fathers at risk for genetic disorders? *Andrologia* 2003;35:191–9.
- [68] Austad SN. *Why we age*. New York: John Wiley and Sons; 1997.
- [69] Austad SN. Retarded senescence in an insular population of Virginia opossums (*Didelphis virginiana*). *J Zool* 1993;229:695–708.
- [70] Reznick DN, Bryant MJ, Roff D, Ghalambor CK, Ghalambor DE. Effect of extrinsic mortality on the evolution of senescence in guppies. *Nature* 2004;431:1095–9.
- [71] Martin GM. Stochastic modulations of the pace and patterns of ageing: impacts on quasi-stochastic distributions of multiple geriatric pathologies. *Mech Ageing Dev* 2012;133:107–11.
- [72] Vanfleteren JR, De Vreese A, Braeckman BP. Two-parameter logistic and Weibull equations provide better fits to survival data from isogenic populations of *Caenorhabditis elegans* in axenic culture than does the Gompertz model. *J Gerontol A Biol Sci Med Sci* 1998;53:B393–403. discussion B404–398.
- [73] Luckinbill LS, Graves JL, Reed AH, Koetsawang S. Localizing genes that defer senescence in *Drosophila melanogaster*. *Heredity* 1988;60(Pt 3):367–74.
- [74] Martin GM. Genetic syndromes in man with potential relevance to the pathobiology of aging. *Birth Defects Orig Artic Ser* 1978;14:5–39.
- [75] Cutler RG. Evolution of human longevity and the genetic complexity governing aging rate. *Proc Natl Acad Sci U S A* 1975;72:4664–8.
- [76] Colman RJ, Anderson RM, Johnson SC, Kastman EK, Kosmatka KJ, Beasley TM, Allison DB, Cruzen C, Simmons HA, Kemnitz JW, Weindruch R. Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science* 2009;325:201–4.
- [77] Roth GS, Mattison JA, Ottinger MA, Chachich ME, Lane MA, Ingram DK. Aging in rhesus monkeys: relevance to human health interventions. *Science* 2004;305:1423–6.
- [78] Mattison JA, Colman RJ, Beasley TM, Allison DB, Kemnitz JW, Roth GS, Ingram DK, Weindruch R, de Cabo R, Anderson RM. Caloric restriction improves health and survival of rhesus monkeys. *Nat Commun* 2017;8:14063.
- [79] Muller FL, Lustgarten MS, Jang Y, Richardson A, Van Remmen H. Trends in oxidative aging theories. *Free Radic Biol Med* 2007;43:477–503.
- [80] Brand MD. Mitochondrial generation of superoxide and hydrogen peroxide as the source of mitochondrial redox signaling. *Free Radic Biol Med* 2016;100:14–31.
- [81] Marcinek DJ, Siegel MP. Targeting redox biology to reverse mitochondrial dysfunction. *Aging* 2013;5:588–9.
- [82] Orgel LE. The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc Natl Acad Sci U S A* 1963;49:517–21.
- [83] Orgel LE. The maintenance of the accuracy of protein synthesis and its relevance to ageing: a correction. *Proc Natl Acad Sci U S A* 1970;67:1476.
- [84] Rothstein M. An overview of age-related changes in proteins. *Prog Clin Biol Res* 1989;287:259–67.
- [85] Martin GM, Bressler SL. Transcriptional infidelity in aging cells and its relevance for the Orgel hypothesis. *Neurobiol Aging* 2000;21:897–900. discussion 903–894.
- [86] Rattan SI. Synthesis, modification and turnover of proteins during aging. *Adv Exp Med Biol* 2010;694:1–13.
- [87] Rattan SI, Derventzi A, Clark BF. Protein synthesis, posttranslational modifications, and aging. *Ann N Y Acad Sci* 1992;663:48–62.
- [88] Gershon H, Gershon D. Detection of inactive enzyme molecules in ageing organisms. *Nature* 1970;227:1214–7.
- [89] Stadtman ER. Protein oxidation and aging. *Free Radic Res* 2006;40:1250–8.
- [90] Kirshtein M, Aston C, Hintz R, Vlassara H. Receptor-specific induction of insulin-like growth factor I in human monocytes by advanced glycosylation end product-modified proteins. *J Clin Invest* 1992;90:439–46.

- [91] Asif M, Egan J, Vasan S, Jyothirmayi GN, Masurekar MR, Lopez S, Williams C, Torres RL, Wagle D, Ulrich P, Cerami A, Brines M, Regan TJ. An advanced glycation endproduct cross-link breaker can reverse age-related increases in myocardial stiffness. *Proc Natl Acad Sci U S A* 2000;97:2809–13.
- [92] Kristal BS, Yu BP. An emerging hypothesis: synergistic induction of aging by free radicals and Maillard reactions. *J Gerontol* 1992;47:B107–14.
- [93] Buxbaum JN. The systemic amyloidoses. *Curr Opin Rheumatol* 2004;16:67–75.
- [94] Edwards BJ, Morley JE. Amylin. *Life Sci* 1992;51:1899–912.
- [95] Higuchi K, Naiki H, Kitagawa K, Hosokawa M, Takeda T. Mouse senile amyloidosis. ASSAM amyloidosis in mice presents universally as a systemic age-associated amyloidosis. *Virchows Arch B Cell Pathol Incl Mol Pathol* 1991;60:231–8.
- [96] Tcw J, Goate AM. Genetics of beta-amyloid precursor protein in Alzheimer's disease. *Cold Spring Harb Perspect Med* 2017;7.
- [97] Clarke S. Aging as war between chemical and biochemical processes: protein methylation and the recognition of age-damaged proteins for repair. *Ageing Res Rev* 2003;2:263–85.
- [98] DeVry CG, Clarke S. Polymorphic forms of the protein L-isoaspartate (D-aspartate) O-methyltransferase involved in the repair of age-damaged proteins. *J Hum Genet* 1999;44:275–88.
- [99] Sala AJ, Bott LC, Morimoto RI. Shaping proteostasis at the cellular, tissue, and organismal level. *J Cell Biol* 2017;216:1231–41.
- [100] Szilard L. On the nature of the aging process. *Proc Natl Acad Sci U S A* 1959;45:30–45.
- [101] Bennett-Baker PE, Wilkowski J, Burke DT. Age-associated activation of epigenetically repressed genes in the mouse. *Genetics* 2003;165:2055–62.
- [102] Cattanaach BM. Position effect variegation in the mouse. *Genet Res* 1974;23:291–306.
- [103] Wareham KA, Lyon MF, Glenister PH, Williams ED. Age related reactivation of an X-linked gene. *Nature* 1987;327:725–7.
- [104] Migeon BR, Axelmann J, Beggs AH. Effect of ageing on reactivation of the human X-linked HPRT locus. *Nature* 1988;335:93–6.
- [105] Wilson VL, Jones PA. DNA methylation decreases in aging but not in immortal cells. *Science* 1983;220:1055–7.
- [106] Wilson VL, Smith RA, Ma S, Cutler RG. Genomic 5-methyldeoxycytidine decreases with age. *J Biol Chem* 1987;262:9948–51.
- [107] Issa JP, Ottaviano YL, Celano P, Hamilton SR, Davidson NE, Baylin SB. Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat Genet* 1994;7:536–40.
- [108] Imai SI, Guarente L. It takes two to tango: NAD⁺ and sirtuins in aging/longevity control. *NPJ Aging Mech Dis* 2016;2:16017.
- [109] Watroba M, Dudek I, Skoda M, Stangret A, Rzedkiewicz P, Szukiewicz D. Sirtuins, epigenetics and longevity. *Ageing Res Rev* 2017;40:11–9.
- [110] Linford NJ, Beyer RP, Gollahon K, Krajcik RA, Malloy VL, Demas V, Burmer GC, Rabinovitch PS. Transcriptional response to aging and caloric restriction in heart and adipose tissue. *Ageing Cell* 2007;6(5):673–88.
- [111] Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, Roetker NS, Just AC, Demerath EW, Guan W, Bressler J, Fornage M, Studenski S, Vandiver AR, Moore AZ, Tanaka T, Kiel DP, Liang L, Vokonas P, Schwartz J, Lunetta KL, Murabito JM, Bandinelli S, Hernandez DG, Melzer D, Nalls M, Pilling LC, Price TR, Singleton AB, Gieger C, Holle R, Kretschmer A, Kronenberg F, Kunze S, Linseisen J, Meisinger C, Rathmann W, Waldenberger M, Visscher PM, Shah S, Wray NR, McRae AF, Franco OH, Hofman A, Uitterlinden AG, Absher D, Assimes T, Levine ME, Lu AT, Tsao PS, Hou L, Manson JE, Carty CL, LaCroix AZ, Reiner AP, Spector TD, Feinberg AP, Levy D, Baccarelli A, van Meurs J, Bell JT, Peters A, Deary IJ, Panikow JS, Ferrucci L, Horvath S. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Ageing* 2016;8:1844–65.
- [112] Clark AM, Rubin MA. The modification by x-irradiation of the life span of haploids and diploids of the wasp, *Habrobracon* sp. *Radiat Res* 1961;15:244–53.
- [113] Martin GM, Smith AC, Ketterer DJ, Ogburn CE, Distechte CM. Increased chromosomal aberrations in first metaphases of cells isolated from the kidneys of aged mice. *Isr J Med Sci* 1985;21:296–301.
- [114] Horn PL, Turker MS, Ogburn CE, Distechte CM, Martin GM. A cloning assay for 6-thioguanine resistance provides evidence against certain somatic mutational theories of aging. *J Cell Physiol* 1984;121:309–15.
- [115] Martin GM, Ogburn CE, Colgin LM, Gown AM, Edland SD, Monnat Jr RJ. Somatic mutations are frequent and increase with age in human kidney epithelial cells. *Hum Mol Genet* 1996;5:215–21.
- [116] Trainor KJ, Wigmore DJ, Chrysostomou A, Dempsey JL, Seshadri R, Morley AA. Mutation frequency in human lymphocytes increases with age. *Mech Ageing Dev* 1984;27:83–6.
- [117] Turner DR, Morley AA, Haliandros M, Kutlaca R, Sanderson BJ. In vivo somatic mutations in human lymphocytes frequently result from major gene alterations. *Nature* 1985;315:343–5.

- [118] Dolle ME, Snyder WK, Dunson DB, Vijg J. Mutational fingerprints of aging. *Nucleic Acids Res* 2002;30:545–9.
- [119] Gerez L, de Haan A, Hol EM, Fischer DF, van Leeuwen FW, van Steeg H, Benne R. Molecular misreading: the frequency of dinucleotide deletions in neuronal mRNAs for beta-amyloid precursor protein and ubiquitin B. *Neurobiol Aging* 2005;26:145–55.
- [120] de Lange T. How telomeres solve the end-protection problem. *Science* 2009;326:948–52.
- [121] Shay JW, Wright WE. Hallmarks of telomeres in ageing research. *J Pathol* 2007;211:114–23.
- [122] Cawthon RM, Smith KR, O'Brien E, Sivatchenko A, Kerber RA. Association between telomere length in blood and mortality in people aged 60 years or older. *Lancet* 2003;361:393–5.
- [123] Epel ES, Blackburn EH, Lin J, Dhabhar FS, Adler NE, Morrow JD, Cawthon RM. Accelerated telomere shortening in response to life stress. *Proc Natl Acad Sci U S A* 2004;101:17312–5.
- [124] Fair B, Mellon SH, Epel ES, Lin J, Revesz D, Verhoeven JE, Penninx BW, Reus VI, Rosser R, Hough CM, Mahan L, Burke HM, Blackburn EH, Wolkowitz OM. Telomere length is inversely correlated with urinary stress hormone levels in healthy controls but not in un-medicated depressed individuals-preliminary findings. *J Psychosom Res* 2017;99:177–80.
- [125] Gerschman R, Nye SW, Gilbert DL, Dwyer P, Fenn WO. Studies on oxygen poisoning: protective effect of beta-mercaptoethylamine. *Proc Soc Exp Biol Med* 1954;85:75–7.
- [126] Harman D. Aging: a theory based on free radical and radiation chemistry. *J Gerontol* 1956;11:298–300.
- [127] Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, Bruder CE, Bohlooly YM, Gidlöf S, Oldfors A, Wibom R, Tornell J, Jacobs HT, Larsson NG. Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* 2004;429:417–23.
- [128] Scott L, Dawson VL, Dawson TM. Trumping neurodegeneration: targeting common pathways regulated by autosomal recessive Parkinson's disease genes. *Exp Neurol* 2017.
- [129] Gonzalez-Freire M, de Cabo R, Bernier M, Sollott SJ, Fabbri E, Navas P, Ferrucci L. Reconsidering the role of mitochondria in aging. *J Gerontol A Biol Sci Med Sci* 2015;70:1334–42.
- [130] Herati AS, Zhelyazkova BH, Butler PR, Lamb DJ. Age-related alterations in the genetics and genomics of the male germ line. *Fertil Steril* 2017;107:319–23.
- [131] Horie H, Kawasaki Y, Takenaka T. Lateral diffusion of membrane lipids changes with aging in C57BL mouse dorsal root ganglion neurons from a fetal stage to an aged stage. *Brain Res* 1986;377(2):246–50.
- [132] Medkour Y, Dakik P, McAuley M, Mohammad K, Mitrofanova D, Titorenko VI. Mechanisms underlying the essential role of mitochondrial membrane lipids in yeast chronological aging. *Oxid Med Cell Longev* 2017;2017:2916985.
- [133] Martin GM, Oshima J. Lessons from human progeroid syndromes. *Nature* 2000;408:263–6.
- [134] Oshima J, Martin GM, Hisama FM. Werner syndrome. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJH, Bird TD, Fong CT, Mefford HC, Smith RJH, Stephens K, editors. *GeneReviews(R)*, Seattle (WA). 2014.
- [135] Takemoto M, Mori S, Kuzuya M, Yoshimoto S, Shimamoto A, Igarashi M, Tanaka Y, Miki T, Yokote K. Diagnostic criteria for Werner syndrome based on Japanese nationwide epidemiological survey. *Geriatr Gerontol Int* 2013;13:475–81.
- [136] Goto M, Miller RW, Ishikawa Y, Sugano H. Excess of rare cancers in Werner syndrome (adult progeria). *Cancer Epidemiol Biomark Prev* 1996;5:239–46.
- [137] Huang S, Lee L, Hanson NB, Lenaerts C, Hoehn H, Poot M, Rubin CD, Chen DF, Yang CC, Juch H, Dorn T, Spiegel R, Oral EA, Abid M, Battisti C, Lucci-Cordisco E, Neri G, Steed EH, Kidd A, Isley W, Showalter D, Vittone JL, Konstantinow A, Ring J, Meyer P, Wenger SL, von Herbay A, Wollina U, Schuelke M, Huizenga CR, Leistriz DF, Martin GM, Mian IS, Oshima J. The spectrum of WRN mutations in Werner syndrome patients. *Hum Mutat* 2006;27:558–67.
- [138] Epstein CJ, Martin GM, Schultz AL, Motulsky AG. Werner's syndrome a review of its symptomatology, natural history, pathologic features, genetics and relationship to the natural aging process. *Medicine (Baltim)* 1966;45:177–221.
- [139] Goto M. Hierarchical deterioration of body systems in Werner's syndrome: implications for normal ageing. *Mech Ageing Dev* 1997;98:239–54.
- [140] Goto M, Ishikawa Y, Sugimoto M, Furuichi Y. Werner syndrome: a changing pattern of clinical manifestations in Japan (1917~2008). *Biosci Trends* 2013;7:13–22.
- [141] Postiglione A, Soricelli A, Covelli EM, Iazzetta N, Ruocco A, Milan G, Santoro L, Alfano B, Brunetti A. Premature aging in Werner's syndrome spares the central nervous system. *Neurobiol Aging* 1996;17:325–30.
- [142] Mori S, Zhou H, Yamaga M, Takemoto M, Yokote K. Femoral osteoporosis is more common than lumbar osteoporosis in patients with Werner syndrome. *Geriatr Gerontol Int* 2017;17:854–6.
- [143] Goto M, Kindynis P, Resnick D, Sartoris DJ. Osteosclerosis of the phalanges in Werner syndrome. *Radiology* 1989;172:841–3.

- [144] Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, Matthews S, Nakura J, Miki T, Ouais S, Martin GM, Mulligan J, Schellenberg GD. Positional cloning of the Werner's syndrome gene. *Science* 1996;272:258–62.
- [145] Gray MD, Shen JC, Kamath-Loeb AS, Blank A, Sopher BL, Martin GM, Oshima J, Loeb LA. The Werner syndrome protein is a DNA helicase. *Nat Genet* 1997;17:100–3.
- [146] Kitano K, Kim SY, Hakoshima T. Structural basis for DNA strand separation by the unconventional winged-helix domain of RecQ helicase WRN. *Structure* 2010;18:177–87.
- [147] Huang S, Li B, Gray MD, Oshima J, Mian IS, Campisi J. The premature ageing syndrome protein, WRN, is a 3'→5' exonuclease. *Nat Genet* 1998;20:114–6.
- [148] Muftuoglu M, Kulikowicz T, Beck G, Lee JW, Piotrowski J, Bohr VA. Intrinsic ssDNA annealing activity in the C-terminal region of WRN. *Biochemistry* 2008;47:10247–54.
- [149] Brosh Jr RM, Opresko PL, Bohr VA. Enzymatic mechanism of the WRN helicase/nuclease. *Methods Enzymol* 2006;409:52–85.
- [150] Croteau DL, Popuri V, Opresko PL, Bohr VA. Human RecQ helicases in DNA repair, recombination, and replication. *Annu Rev Biochem* 2014;83:519–52.
- [151] Crabbe L, Jauch A, Naeger CM, Holtgreve-Grez H, Karlseder J. Telomere dysfunction as a cause of genomic instability in Werner syndrome. *Proc Natl Acad Sci U S A* 2007;104:2205–10.
- [152] Opresko PL. Telomere ResQue and preservation—roles for the Werner syndrome protein and other RecQ helicases. *Mech Ageing Dev* 2008;129:79–90.
- [153] Yokote K, Chanprasert S, Lee L, Eirich K, Takemoto M, Watanabe A, Koizumi N, Lessel D, Mori T, Hisama FM, Ladd PD, Angle B, Baris H, Cefle K, Palanduz S, Ozturk S, Chateau A, Deguchi K, Easwar TK, Federico A, Fox A, Grebe TA, Hay B, Nampoothiri S, Seiter K, Streeten E, Pina-Aguilar RE, Poke G, Poot M, Posmyk R, Martin GM, Kubisch C, Schindler D, Oshima J. WRN mutation update: mutation spectrum, patient registries, and translational prospects. *Hum Mutat* 2017;38:7–15.
- [154] Suzuki T, Shiratori M, Furuichi Y, Matsumoto T. Diverged nuclear localization of Werner helicase in human and mouse cells. *Oncogene* 2001;20:2551–8.
- [155] Satoh M, Imai M, Sugimoto M, Goto M, Furuichi Y. Prevalence of Werner's syndrome heterozygotes in Japan. *Lancet* 1999;353:1766.
- [156] Masala MV, Scapaticci S, Olivieri C, Pirodda C, Montesu MA, Cuccuru MA, Pruneddu S, Danesino C, Cerimele D. Epidemiology and clinical aspects of Werner's syndrome in North Sardinia: description of a cluster. *Eur J Dermatol* 2007;17:213–6.
- [157] Hutchinson J. Congenital absence of hair and mammary glands with atrophic condition of the skin and its appendages, in a boy whose mother had been almost wholly bald from alopecia areata from the age of six. *Med Chir Trans* 1886;69:473–7.
- [158] Gilford H. Ateleiosis and progeria: continuous youth and premature old age. *Br Med J* 1904;2:914–8.
- [159] Brown WT, Kieras FJ, Houck Jr GE, Dutkowski R, Jenkins EC. A comparison of adult and childhood progerias: Werner syndrome and Hutchinson-Gilford progeria syndrome. *Adv Exp Med Biol* 1985;190:229–44.
- [160] Gordon LB, Brown WT, Collins FS. Hutchinson-gilford progeria syndrome. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJH, Bird TD, Fong CT, Mefford HC, Smith RJH, Stephens K, editors. *GeneReviews(R)*, 2015/01/08 ed, Seattle (WA). 2015.
- [161] Merideth MA, Gordon LB, Clauss S, Sachdev V, Smith AC, Perry MB, Brewer CC, Zalewski C, Kim HJ, Solomon B, Brooks BP, Gerber LH, Turner ML, Domingo DL, Hart TC, Graf J, Reynolds JC, Gropman A, Yanovski JA, Gerhard-Herman M, Collins FS, Nabel EG, Cannon 3rd RO, Gahl WA, Introne WJ. Phenotype and course of Hutchinson-Gilford progeria syndrome. *N Engl J Med* 2008;358:592–604.
- [162] Olive M, Harten I, Mitchell R, Beers JK, Djabali K, Cao K, Erdos MR, Blair C, Funke B, Smoot L, Gerhard-Herman M, Machan JT, Kutys R, Virmani R, Collins FS, Wight TN, Nabel EG, Gordon LB. Cardiovascular pathology in Hutchinson-Gilford progeria: correlation with the vascular pathology of aging. *Arterioscler Thromb Vasc Biol* 2010;30:2301–9.
- [163] De Sandre-Giovannoli A, Bernard R, Cau P, Navarro C, Amiel J, Boccaccio I, Lyonnet S, Stewart CL, Munich A, Le Merrer M, Levy N. Lamin A truncation in Hutchinson-Gilford progeria. *Science* 2003;300:2055.
- [164] Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robbins CM, Moses TY, Berglund P, Dutra A, Pak E, Durkin S, Csoka AB, Boehnke M, Glover TW, Collins FS. Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* 2003;423:293–8.
- [165] Broers JL, Ramaekers FC, Bonne G, Yaou RB, Hutchinson CJ. Nuclear lamins: laminopathies and their role in premature ageing. *Physiol Rev* 2006;86:967–1008.
- [166] Goldman RD, Shumaker DK, Erdos MR, Eriksson M, Goldman AE, Gordon LB, Gruenbaum Y, Khuon S, Mendez M, Varga R, Collins FS. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. *Proc Natl Acad Sci U S A* 2004;101:8963–8.

- [167] Dahl KN, Scaffidi P, Islam MF, Yodh AG, Wilson KL, Misteli T. Distinct structural and mechanical properties of the nuclear lamina in Hutchinson-Gilford progeria syndrome. *Proc Natl Acad Sci U S A* 2006;103:10271–6.
- [168] Shimi T, Pfliegerhaer K, Kojima S, Pack CG, Solovei I, Goldman AE, Adam SA, Shumaker DK, Kinjo M, Cremer T, Goldman RD. The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes Dev* 2008;22:3409–21.
- [169] Liu Y, Rusinol A, Sinensky M, Wang Y, Zou Y. DNA damage responses in progeroid syndromes arise from defective maturation of prelamin A. *J Cell Sci* 2006;119:4644–9.
- [170] Benson EK, Lee SW, Aaronson SA. Role of progerin-induced telomere dysfunction in HGPS premature cellular senescence. *J Cell Sci* 2010;123:2605–12.
- [171] Scaffidi P, Misteli T. Lamin A-dependent nuclear defects in human aging. *Science* 2006;312:1059–63.
- [172] Worman HJ, Bonne G. “Laminopathies”: a wide spectrum of human diseases. *Exp Cell Res* 2007;313:2121–33.
- [173] Hisama FM, Lessel D, Leistritz D, Friedrich K, McBride KL, Pastore MT, Gottesman GS, Saha B, Martin GM, Kubisch C, Oshima J. Coronary artery disease in a Werner syndrome-like form of progeria characterized by low levels of progerin, a splice variant of lamin A. *Am J Med Genet* 2011;155A:3002–6.
- [174] Capell BC, Olive M, Erdos MR, Cao K, Faddah DA, Tavarez UL, Conneely KN, Qu X, San H, Ganesh SK, Chen X, Avallone H, Kolodgie FD, Virmani R, Nabel EG, Collins FS. A farnesyltransferase inhibitor prevents both the onset and late progression of cardiovascular disease in a progeria mouse model. *Proc Natl Acad Sci U S A* 2008;105:15902–7.
- [175] Gabriel D, Gordon LB, Djabali K. Temsirolimus partially rescues the Hutchinson-Gilford progeria cellular phenotype. *PLoS One* 2016;11:e0168988.
- [176] Yang SH, Chang SY, Andres DA, Spielmann HP, Young SG, Fong LG. Assessing the efficacy of protein farnesyltransferase inhibitors in mouse models of progeria. *J Lipid Res* 2010;51:400–5.
- [177] Yang SH, Meta M, Qiao X, Frost D, Bauch J, Coffinier C, Majumdar S, Bergo MO, Young SG, Fong LG. A farnesyltransferase inhibitor improves disease phenotypes in mice with a Hutchinson-Gilford progeria syndrome mutation. *J Clin Invest* 2006;116:2115–21.
- [178] Gordon LB, Kleinman ME, Miller DT, Neuberg DS, Giobbie-Hurder A, Gerhard-Herman M, Smoot LB, Gordon CM, Cleveland R, Snyder BD, Fligor B, Bishop WR, Statkevich P, Regen A, Sonis A, Riley S, Ploski C, Correia A, Quinn N, Ullrich NJ, Nazarian A, Liang MG, Huh SY, Schwartzman A, Kieran MW. Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome. *Proc Natl Acad Sci U S A* 2012;109:16666–71.
- [179] Gordon LB, Shappell H, Massaro J, D'Agostino RB, Brazier Sr., J, Campbell SE, Kleinman ME, Kieran MW. Association of lonafarnib treatment vs no treatment with mortality rate in patients with Hutchinson-Gilford progeria syndrome. *JAMA* 2018;319(16):1687–95.
- [180] Yang SH, Chang SY, Ren S, Wang Y, Andres DA, Spielmann HP, Fong LG, Young SG. Absence of progeria-like disease phenotypes in knock-in mice expressing a non-farnesylated version of progerin. *Hum Mol Genet* 2011;20:436–44.
- [181] Weedon MN, Ellard S, Prindle MJ, Caswell R, Lango Allen H, Oram R, Godbole K, Yajnik CS, Sbraccia P, Novelli G, Turnpenny P, McCann E, Goh KJ, Wang Y, Fulford J, McCulloch LJ, Savage DB, O'Rahilly S, Kos K, Loeb LA, Semple RK, Hattersley AT. An in-frame deletion at the polymerase active site of POLD1 causes a multisystem disorder with lipodystrophy. *Nat Genet* 2013;45:947–50.
- [182] Kamath-Loeb AS, Shen JC, Schmitt MW, Loeb LA. The Werner syndrome exonuclease facilitates DNA degradation and high fidelity DNA polymerization by human DNA polymerase delta. *J Biol Chem* 2012;287:12480–90.
- [183] Lessel D, Hisama FM, Szakson K, Saha B, Sanjuanelo AB, Salbert BA, Steele PD, Baldwin J, Brown WT, Piussan C, Plauchu H, Szilvassy J, Horkay E, Hogel J, Martin GM, Herr AJ, Oshima J, Kubisch C. POLD1 germline mutations in patients initially diagnosed with Werner syndrome. *Hum Mutat* 2015;36:1070–9.
- [184] Pelosini C, Martinelli S, Ceccarini G, Magno S, Barone I, Basolo A, Fierabracci P, Vitti P, Maffei M, Santini F. Identification of a novel mutation in the polymerase delta 1 (POLD1) gene in a lipodystrophic patient affected by mandibular hypoplasia, deafness, progeroid features (MDPL) syndrome. *Metabolism* 2014;63:1385–9.
- [185] Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona LG, Ma Y, Kaur K, Dobbins S, Barclay E, Gorman M, Martin L, Kovac MB, Humphray S, Consortium C, Consortium WGS, Lucassen A, Holmes CC, Bentley D, Donnelly P, Taylor J, Petridis C, Roylance R, Sawyer EJ, Kerr DJ, Clark S, Grimes J, Kearsey SE, Thomas HJ, McVean G, Houlston RS, Tomlinson I. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 2013;45:136–44.

- [186] Laugel V. Cockayne syndrome: the expanding clinical and mutational spectrum. *Mech Ageing Dev* 2013;134:161–70.
- [187] Jaarsma D, van der Pluijm I, van der Horst GT, Hoeijmakers JH. Cockayne syndrome pathogenesis: lessons from mouse models. *Mech Ageing Dev* 2013;134:180–95.
- [188] Martein J, Lans H, Vermeulen W, Hoeijmakers JH. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol* 2014;15:465–81.
- [189] Chatre L, Biard DS, Sarasin A, Ricchetti M. Reversal of mitochondrial defects with CSB-dependent serine protease inhibitors in patient cells of the progeroid Cockayne syndrome. *Proc Natl Acad Sci U S A* 2015;112(22):E2910–9.
- [190] Niedernhofer LJ, Garinis GA, Raams A, Lalai AS, Robinson AR, Appeldoorn E, Odijk H, Oostendorp R, Ahmad A, van Leeuwen W, Theil AF, Vermeulen W, van der Horst GT, Meinecke P, Kleijer WJ, Vijg J, Jaspers NG, Hoeijmakers JH. A new progeroid syndrome reveals that genotoxic stress suppresses the somatotrophic axis. *Nature* 2006;444:1038–43.
- [191] Anderson BH, Kasher PR, Mayer J, Szykiewicz M, Jenkinson EM, Bhaskar SS, Urquhart JE, Daly SB, Dickerson JE, O'Sullivan J, Leibundgut EO, Muter J, Abdel-Salem GM, Babul-Hirji R, Baxter P, Berger A, Bonafe L, Brunstom-Hernandez JE, Buckard JA, Chitayat D, Chong WK, Cordelli DM, Ferreira P, Fluss J, Forrest EH, Franzoni E, Garone C, Hammans SR, Houge G, Hughes I, Jacquemont S, Jeannet PY, Jefferson RJ, Kumar R, Kutschke G, Lundberg S, Lourenco CM, Mehta R, Naidu S, Nischal KK, Nunes L, Ounap K, Philippart M, Prabhakar P, Risen SR, Schiffmann R, Soh C, Stephenson JB, Stewart H, Stone J, Tolmie JL, van der Knaap MS, Vieira JP, Vilain CN, Wakeling EL, Wermenbol V, Whitney A, Lovell SC, Meyer S, Livingston JH, Baerlocher GM, Black GC, Rice GI, Crow YJ. Mutations in CTC1, encoding conserved telomere maintenance component 1, cause Coats plus. *Nat Genet* 2012;44:338–42.
- [192] Gu P, Chang S. Functional characterization of human CTC1 mutations reveals novel mechanisms responsible for the pathogenesis of the telomere disease Coats plus. *Ageing Cell* 2013;12:1100–9.
- [193] Nolis T. Exploring the pathophysiology behind the more common genetic and acquired lipodystrophies. *J Hum Genet* 2014;59:16–23.
- [194] Lawson MA. Lipoatrophic diabetes: a case report with a brief review of the literature. *J Adolesc Health* 2009;44:94–5.
- [195] Nelson MD, Victor RG, Szczepaniak EW, Simha V, Garg A, Szczepaniak LS. Cardiac steatosis and left ventricular hypertrophy in patients with generalized lipodystrophy as determined by magnetic resonance spectroscopy and imaging. *Am J Cardiol* 2013;112:1019–24.
- [196] Machado PV, Daxbacher EL, Obadia DL, Cunha EF, Alves Mde F, Mann D. Do you know this syndrome? Berardinelli-Seip syndrome. *An Bras Dermatol* 2013;88:1011–3.
- [197] Wei S, Soh SL, Qiu W, Yang W, Seah CJ, Guo J, Ong WY, Pang ZP, Han W. Seipin regulates excitatory synaptic transmission in cortical neurons. *J Neurochem* 2013;124:478–89.
- [198] Jenning EH, de Vroede M, Hamers N, Breur JM, Verhoeven-Duif NM, Berger R, Kalkhoven E. A patient with congenital generalized lipodystrophy due to a novel mutation in BSCL2: indications for secondary mitochondrial dysfunction. *JIMD Rep* 2012;4:47–54.
- [199] Agrawala RK, Choudhury AK, Mohanty BK, Baliasinha AK. Berardinelli-Seip congenital lipodystrophy: an autosomal recessive disorder with rare association of duodenocolonic polyps. *J Pediatr Endocrinol Metab* 2014;27:989–91.
- [200] Ben-Avraham D, Govindaraju DR, Budagov T, Fradin D, Durda P, Liu B, Ott S, Gutman D, Sharvit L, Kaplan R, Bougneres P, Reiner A, Shuldiner AR, Cohen P, Barzilai N, Atzmon G. The GH receptor exon 3 deletion is a marker of male-specific exceptional longevity associated with increased GH sensitivity and taller stature. *Sci Adv* 2017;3:e1602025.
- [201] Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell Jr PC, Rimmeler JB, Locke PA, Conneally PM, Schmechel KE, et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat Genet* 1994;7:180–4.
- [202] Drenos F, Kirkwood TB. Selection on alleles affecting human longevity and late-life disease: the example of apolipoprotein E. *PLoS One* 2010;5:e10022.
- [202b] Suri S, Heise V, Trachtenberg AJ, Mackay CE. The forgotten APOE allele: a review of the evidence and suggested mechanisms for the protective effect of APOE ε2. *Neurosci Biobehav Rev* 2013 Dec;37(10 Pt 2):2878–86. <https://doi.org/10.1016/j.neubiorev.2013.10.010>. Epub 2013 Oct 29. Review. PMID: 24183852.
- [203] Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, Derre A, Villegier L, Farnier M, Beucier I, Bruckert E, Chambaz J, Chanu B, Lecerf JM, Luc G, Moulin P, Weissenbach J, Prat A, Krempf M, Junien C, Seidah NG, Boileau C. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet* 2003;34:154–6.
- [204] Horton JD, Cohen JC, Hobbs HH. PCSK9: a convertase that coordinates LDL catabolism. *J Lipid Res* 2009;50(Suppl.):S172–7.

- [205] Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ, Fennell T, Banks E, Ambrogio L, Cibulskis K, Kernysky A, Gonzalez E, Rudzicz N, Engert JC, DePristo MA, Daly MJ, Cohen JC, Hobbs HH, Altshuler D, Schonfeld G, Gabriel SB, Yue P, Kathiresan S. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med* 2010;363:2220–7.
- [206] Dewey FE, Gusarova V, Dunbar RL, O'Dushlaine C, Schurmann C, Gottesman O, McCarthy S, Van Hout CV, Bruse S, Dansky HM, Leader JB, Murray MF, Ritchie MD, Kirchner HL, Habegger L, Lopez A, Penn J, Zhao A, Shao W, Stahl N, Murphy AJ, Hamon S, Bouzelmat A, Zhang R, Shumel B, Pordy R, Gipe D, Herman GA, Sheu WHH, Lee IT, Liang KW, Guo X, Rotter JL, Chen YI, Kraus WE, Shah SH, Damrauer S, Small A, Rader DJ, Wulff AB, Nordestgaard BG, Tybjaerg-Hansen A, van den Hoek AM, Princen HMG, Ledbetter DH, Carey DJ, Overton JD, Reid JG, Sasiela WJ, Banerjee P, Shuldiner AR, Borecki IB, Teslovich TM, Yancopoulos GD, Mellis SJ, Gromada J, Baras A. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. *N Engl J Med* 2017;377:211–21.
- [207] Graham MJ, Lee RG, Brandt TA, Tai LJ, Fu W, Peralta R, Yu R, Hurh E, Paz E, McEvoy BW, Baker BF, Pham NC, Digenio A, Hughes SG, Geary RS, Witztum JL, Crooke RM, Tsimikas S. Cardiovascular and metabolic effects of ANGPTL3 antisense oligonucleotides. *N Engl J Med* 2017;377:222–32.
- [208] Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won HH, Karczewski KJ, O'Donnell-Luria AH, Samocha KE, Weisburd B, Gupta N, Zaidi M, Samuel M, Imran A, Abbas S, Majeed F, Ishaq M, Akhtar S, Trindade K, Mucksavage M, Qamar N, Zaman KS, Yaqoob Z, Saghir T, Rizvi SNH, Memon A, Hayyat Mallick N, Ishaq M, Rasheed SZ, Memon FU, Mahmood K, Ahmed N, Do R, Krauss RM, MacArthur DG, Gabriel S, Lander ES, Daly MJ, Frossard P, Danesh J, Rader DJ, Kathiresan S. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* 2017;544:235–9.
- [209] Wilhelmsen L, Dellborg M, Welin L, Svardsudd K. Men born in 1913 followed to age 100 years. *Scand Cardiovasc J* 2015;49:45–8.
- [210] Rajpathak SN, Liu Y, Ben-David O, Reddy S, Atzmon G, Crandall J, Barzilai N. Lifestyle factors of people with exceptional longevity. *J Am Geriatr Soc* 2011;59:1509–12.
- [211] Dragani TA, Canzian F, Pierotti MA. A polygenic model of inherited predisposition to cancer. *FASEB J* 1996;10:865–70.
- [212] Galvan A, Falvella FS, Spinola M, Frullanti E, Leoni VP, Noci S, Alonso MR, Zolin A, Spada E, Milani S, Pastorino U, Incarbone M, Santambrogio L, Gonzalez Neira A, Dragani TA. A polygenic model with common variants may predict lung adenocarcinoma risk in humans. *Int J Cancer* 2008;123:2327–30.
- [213] Esplin ED, Oei L, Snyder MP. Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics* 2014;15:1771–90.
- [214] Barzilai N, Atzmon G, Schechter C, Schaefer EJ, Cupples AL, Lipton R, Cheng S, Shuldiner AR. Unique lipoprotein phenotype and genotype associated with exceptional longevity. *J Am Med Assoc* 2003;290:2030–40.
- [215] Kleindorp R, Flachsbarf F, Puca AA, Malovini A, Schreiber S, Nebel A. Candidate gene study of FOXO1, FOXO4 and FOXO6 reveals no association with human longevity in Germans. *Aging Cell* 2011.
- [216] Guevara-Aguirre J, Balasubramanian P, Guevara-Aguirre M, Wei M, Madia F, Cheng CW, Hwang D, Martin-Montalvo A, Saavedra J, Ingles S, de Cabo R, Cohen P, Longo VD. Growth hormone receptor deficiency is associated with a major reduction in pro-aging signaling, cancer, and diabetes in humans. *Sci Transl Med* 2011;3:70ra13.
- [217] Krzisnik C, Grguric S, Cvijovic K, Laron Z. Longevity of the hypopituitary patients from the island Krk: a follow-up study. *Pediatr Endocrinol Rev* 2010;7:357–62.
- [218] Bartke A. Single-gene mutations and healthy ageing in mammals. *Philos Trans R Soc Lond B Biol Sci* 2011;366:28–34.
- [219] Martin GM. Syndromes of accelerated aging. *Natl Canc Inst Monogr* 1982;60:241–7.
- [220] Martin GM. Genetic modulation of senescent phenotypes in *Homo sapiens*. *Cell* 2005;120:523–32.
- [221] Xie Z, Jay KA, Smith DL, Zhang Y, Liu Z, Zheng J, Tian R, Li H, Blackburn EH. Early telomerase inactivation accelerates aging independently of telomere length. *Cell* 2015;160:928–39.
- [222] Nhan HS, Chiang K, Koo EH. The multifaceted nature of amyloid precursor protein and its proteolytic fragments: friends and foes. *Acta Neuropathol* 2015;129:1–19.
- [223] Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015;16:716–26.
- [224] Ambeskovic M, Roseboom TJ, Metz GAS. Transgenerational effects of early environmental insults on aging and disease incidence. *Neurosci Biobehav Rev* 2017.

Pharmacogenomics

Daniel W. Nebert^{1,2,3}, Ge Zhang^{2,3}

¹Department of Environmental Health and Center for Environmental Genetics, University of Cincinnati School of Medicine, Cincinnati, OH, United States

²Department of Pediatrics, University of Cincinnati School of Medicine, Cincinnati, OH, United States

³Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

16.1 INTRODUCTION

Individualized drug therapy represents a major portion of *personalized medicine*. The visionary human geneticist Arno G. Motulsky is credited with being first to propose in a publication that “*drug responses* will depend upon each patient’s genetic make-up” [146]. Another human geneticist, Friedrich O. Vogel, coined the term *pharmacogenetics* and defined it as “the study of heritable variability in drug response” [229], or, simply, “gene–drug interactions.”

In the 1990s, the term *pharmacogenomics* was introduced; this came about as a direct offshoot of the Human Genome Project. Pharmacogenomics is defined as “the study of how drugs interact with the *total genome*, to influence biological pathways and processes” [148], or, simply, “drug–genome interactions.” This field should help identify new drug targets and thus be instrumental in designing new drugs. Today, the terms pharmacogenetics and pharmacogenomics are used interchangeably; hence, in this chapter, we use the abbreviation “PGx” to incorporate them both.

In large part, our genetic make-up determines our *individual drug response*. However, each individual’s drug response is *holistic* in that it actually encompasses five contributing influences: (1) *genotype* (DNA single-nucleotide variants, insertions, deletions, duplications, and inversions); (2) *epigenetic effects* (DNA methylation, RNA interference, histone modifications,

and chromatin remodeling); (3) *endogenous influences* (age, gender, ethnicity, exercise, various disease states, and functional status of kidneys and other organs); (4) *environmental factors* (diet, cigarette smoking, lifestyle, drug–drug interactions, and significant exposure to occupational chemicals and other environmental pollutants); and (5) *microbiome differences* specific to each person.

Thus, each patient has his own unique “PGx profile”—just as each of us has his own distinct pattern of DNA microsatellite differences, profile of *single-nucleotide polymorphisms*, or *variants* (SNPs; SNVs), finger prints, and even the biometric properties of the eye’s iris. Besides the genetic component, however, the other four categories listed above are environment- and time-dependent, i.e., a constantly moving target. A patient’s response to a drug today might differ from that same patient’s response tomorrow, next month, or next year. Thus, it is best to keep in mind that any individual’s drug response is never entirely static.

16.1.1 Types of Drug Responses

Interindividual variability in drug response is defined as an “effect of varying intensity occurring in different individuals receiving a specified drug dose,” or “requirement of a range of doses (concentrations) in order to produce an effect of specified intensity in each patient” [11]. Classifications of drug response include: (1) the desired

beneficial effect (*efficacy*); (2) *adverse effect*; (3) no effect (*therapeutic failure*); and (4) *toxic effect*. The latter two effects are particularly dependent on drug dosage.

One goal of PGx intends to develop rational approaches to optimize drug therapy with respect to each patient's genotype. Another goal of PGx is to ensure maximum efficacy, combined with minimal adverse effects in each individual. If one considers, in addition, the above-mentioned issues of epigenetics, endogenous influences, environmental effects, and each patient's changing microbiome—it is easy to understand how complicated *genetic risk prediction of drug response* can be for each individual patient.

16.1.2 Adverse Drug Reactions

Two decades ago, it was reported that *adverse drug reactions* (ADRs) in the US rank as approximately the fifth leading cause of death [111]. ADRs have been divided into categories (Table 16.1): (1) dose-dependent; (2) dose-independent; (3) dose- and time-dependent (cumulative); and (4) time-related withdrawal reactions [39]. Dose-independent ADRs comprise idiosyncratic drug reactions and allergic reactions; the latter is not addressed in this chapter. A better understanding of PGx should help reduce morbidity and mortality caused by ADRs, but should especially help prevent ADRs caused by *idiosyncratic dose-independent drug reactions*.

In this chapter, we first introduce the topic of clinical pharmacology, followed by the history of genetics and its impact on PGx. Then we present several phenotype examples of early PGx studies, followed by more complex examples. Finally, given the complexity of the genome and interindividual differences, we attempt to clarify some of the challenges and show how difficult it will be—anytime in the near future—to be able to predict each patient's individual drug response, given each person's unique genome.

TABLE 16.1 A Possible Classification of Adverse Drug Reactions (ADRs)

Dose-dependent
Dose-independent
Idiosyncratic drug reactions
Allergic reactions
Dose- and time-dependent (cumulative)
Time-related withdrawal reactions

16.2 FUNDAMENTAL ASPECTS OF CLINICAL PHARMACOLOGY

The objective of clinical pharmacologists, in treating a patient with a drug, is to maintain optimal plasma levels of the *active principle* (drug) in the *therapeutic range* (Fig. 16.1). If the dose of drug is too small, the interval of administration inadequate, bioavailability of the active drug too low, or the active drug too rapidly metabolized and thus quickly cleared—then the active drug level in the blood might never reach its effective concentration. This would lead to absence of the expected response (*therapeutic failure*). On the contrary, if the dose of drug is too large, intervals of administration too short, or the active drug poorly metabolized and thus too slowly excreted, this accumulation can lead to *toxic concentrations*. Levels that cause toxicity will lead to ADRs; such a drug response might occur because of drug accumulation at toxic levels in blood, and accumulation might occur in one or more critical target organs, or in both blood and target organs. Dose-*in*dependent ADRs can happen at any dosage and are generally unpredictable.

The above scenarios exist if the *parent drug* is the component responsible for efficacy as well as toxicity. If a *metabolite* is the *active principle* (i.e., that which causes the efficacy) and it is also the toxicant, then one can replace the word “drug” (above) with

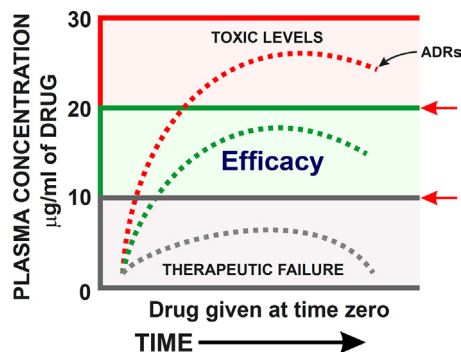


Figure 16.1 Theoretical plasma concentration curves for any drug, as a function of time after administration of the dose. In this hypothetical case, the horizontal line (at 10 µg/mL) is the minimum *effective concentration*; the horizontal line (at 20 µg/mL) is the point at which *toxicity* occurs. (From previous Emery and Rimoin edition; Nebert DW, Vesell ES. Chapter 19-“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

“metabolite” and describe the same events leading to *efficacy* versus *therapeutic failure* versus *toxic levels* [158]. A classic example in which the metabolite is more biologically active than the parent drug is the parent drug codeine, which becomes activated by the CYP2D6 enzyme into the more biologically active metabolite, morphine [54].

16.2.1 Pharmacokinetics (PK) and Pharmacodynamics (PD)

Pharmacokinetics (PK) represents “what the body does to the drug,” whereas pharmacodynamics (PD) can be defined as “what the drug does to the body,” or, more specifically, to target cell-types, tissues, or organs (Fig. 16.2). The end result of all collective PK and PD processes will lead to drug efficacy, therapeutic failure, or toxicity. Sometimes pharmacologists combine therapeutic failure and toxicity into the common category of “ADRs.”

Variations in the drug’s PK phase incorporate the processes of *absorption* (uptake), *distribution*, *metabolism*, and *excretion*. Pharmaceutical companies describe these four processes as “ADME”; these companies often have consolidated specific laboratories into “ADME sections, or divisions, of research.”

PK differences are generally detected by determinations of the active drug (or metabolite) in blood and less

often drug levels in urine (or saliva, sweat, feces, milk, semen). For most systemic drugs, concentration of the (unbound, or free) active principle in blood is almost always proportional to the concentration of *active principle* in the target tissue. An exception would include certain chemotherapeutic agents, designed to become concentrated in cancer cells while avoiding nearby non-cancer cells.

Traditionally, most PGx differences that were initially identified—from the 1940s into the 1990s—involved genes participating in drug metabolism (“PK genes”). Since the mid-1980s, PGx differences began to be described also in genes involved in distribution (binding, transport), absorption, and excretion. It seems reasonable to include cell-surface (both influx and outflux) transporter genes as PK genes, because cellular uptake and excretion are usually essential to the process of absorption in the ADME equation.

Variations in the PD phase occur “downstream” of the PK phase (Fig. 16.2); PD genes would include those that influence drug-action—such as binding to target receptors. There are examples of “recycling” of the active principle from the PD phase back to the PK phase. Thus, PD gene differences might be seen in: transporters and channels across the different intracellular membranes between organelles and in the cytosol and nucleus; tissue- and cell type-specific transcription

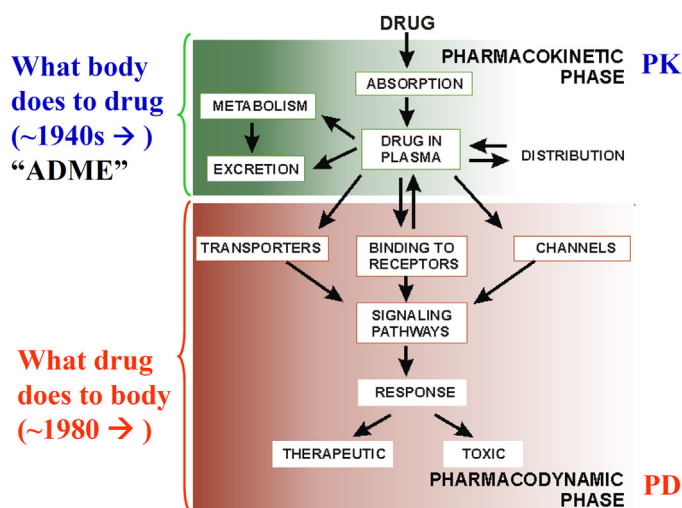


Figure 16.2 Fundamentals of clinical pharmacology. Processes involve the *pharmacokinetic* (PK) phase, and the *pharmacodynamic* (PD) phase—of any drug or over-the-counter preparation. (From previous Emery and Rimoim edition; Nebert DW, Vesell ES. Chapter 19-“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoim’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

factors; components of signal transduction pathways; nucleic acid and protein repair processes; molecular “chaperones”; and cell infrastructure (e.g., nuclear matrix, membranes, and subcellular organelles such as endoplasmic reticulum, Golgi bodies, melanosomes, or peroxisomes). Until the 1980s, what happened in the PD phase was largely a “black box.” Due to the rapid advances in molecular biology since the 1980s, however, downstream drug-signaling pathways have now become better understood, and the field of PD has exploded (reviewed in [158]).

Historically, drugs or chemicals were given to patients, human healthy volunteers, or laboratory animals—and then differences in urinary metabolite profiles were determined (e.g., [10,233]). Subsequently, PK variations in laboratory animals [226] and ultimately in humans [228]—reflected as differences in plasma drug levels or rate of clearance—led to the appreciation of *genetic variability*, presumed to be in drug-metabolizing enzymes (DMEs). These types of studies were carried out in human volunteers or patients, because blood samples are easier to obtain than tissue biopsies.

16.2.2 Plasma Clearance of a Drug

In the late 1960s, comparisons of drug clearances in monozygotic versus dizygotic twin-pairs were performed. The *heritability index* can be approximated as twice the difference in correlation between monozygotic (MZ) and dizygotic (DZ) twin-pairs. A heritability index of 1.0 would represent “purely genetic,” whereas a heritability index of less than 0.50 would indicate “predominantly environmental factors.” From twin studies—of plasma concentrations of dicoumarol, phenylbutazone, desipramine, halothane, nortryptiline, oxyphenbutazone, and antipyrine—the main conclusion was that large variations in drug clearance rates among healthy subjects reflect a strong genetic component [227].

Plasma clearance studies can sometimes be problematic if the drug exhibits “first-pass elimination” (i.e., metabolism or degradation before reaching the systemic circulation), as well as specific “hepatic first-pass kinetics,” i.e., involved in enterohepatic (intestine-to-liver-to-intestine) recycling. This process would be visualized in a graph as a series of rises and falls in the drug’s concentration in blood (plotted on the Y-axis) as a function of time (on the X-axis).

16.2.3 Extrahepatic PGx Differences and Endogenous Functions

Early on, a common misconception was that DMEs exist almost exclusively, or entirely, in liver. Another fallacy was that only drugs, and not endogenous compounds, are substrates for DMEs. Both of these myths are now realized to be wrong (reviewed in [147,157]). For example, CYP3A4, the most abundant cytochrome P450 monooxygenase in liver, is also present in large concentrations in the gastrointestinal (GI) tract. Numerous DMEs are located in lung and kidney. DMEs are even found in the ciliary body [191] and cornea [186] of the eye.

Many DMEs exist in blood cells, vascular endothelial cells, and virtually all cell-types in the body, participating also in the activation and deactivation of *lipid mediators* (LMs) of the *arachidonic acid*, *docosahexaenoic acid*, and *eicosapentaenoic acid* cascades. The end-result of these LM cascades leads to regulation of cell division, cell adhesion and migration, pre- and postinflammatory responses, cell migration, bronchoconstriction, vasodilation, and numerous other developmental and homeostatic mechanisms (reviewed in [154,157]). Because DMEs exist in essentially all cell-types of the brain, they participate in key roles in neuroendocrine functions. Emergence of the “brain-gut-microbiome” (reviewed in [150]) has become a very recent consideration in PGx studies, because of the capacity of gut bacterial enzymes to metabolize—both activate and detoxify—at least some drugs, similarly to that of the host’s DMEs.

DMEs metabolize steroids, fatty acids, bile acids, retinoic acid derivatives, vitamin D and metabolites, and sterols (reviewed in [157]). Because many endogenous ligands for intracellular receptors are metabolized by DMEs, DMEs can be considered as “occurring upstream” of ligand-receptor interactions and second-messenger pathways that are pivotal to virtually all critical-life functions. In fact, it has been suggested [147] that there might not be any DME that metabolizes only drugs—without having originated during evolution first to participate in an endogenous function.

There are hundreds of examples of DME activities occurring in cell-types at higher levels than those in liver, or of DME metabolism occurring exclusively in a tissue other than liver. For example, phenytoin hydroxylation is ~50-fold greater in human oral mucosa than

“Therapeutic Index”

- $$\frac{\text{Toxic dose (TD}_{50})}{\text{Effective dose (ED}_{50})} = 20 \text{ [large window] } \mathbf{A}$$
- $$\frac{\text{Toxic dose (TD}_{50})}{\text{Effective dose (ED}_{50})} = 2 \text{ [narrow window] } \mathbf{B}$$
- **If genetic differences in drug response are 10-fold,**
then drug A → no problem; drug B → ADRs

Figure 16.3 Simple equations illustrating a large versus small “therapeutic window” or “therapeutic index.” (From previous Emery and Rimoin edition; Nebert DW, Vesell ES. Chapter 19-“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

in liver [253]. Some DMEs exist at high concentrations in nasal mucosa [120]. In conclusion of this section, it should be appreciated that variability in drug response can occur in any tissue, as well as any cell-type—and not solely in liver. Furthermore, some specific DME activities can be much higher in a particular tissue than in liver, or exclusively in one or another cell-type instead of liver. Moreover, all that has been described herein for PK genes also holds true for PD genes.

16.2.4 Therapeutic Index (or “Window”)

Most early PGx studies owe their success to pharmacogeneticists’ selection of a drug having a narrow *therapeutic index*. If a drug shows a wide therapeutic window, it is unlikely to cause toxicity in a significant subset of any human population; therefore, this would be of little concern to public health, and the need to prevent ADRs would be small. For example, if the dose causing toxicity is 20 times greater than the dose needed to be effective (Fig. 16.3), and genetic differences in handling this drug are never greater than 10-fold across all human populations, this drug would be of little concern to pharmacogeneticists. On the other hand, if the dose causing toxicity is only twice the dose for efficacy—and PGx differences in handling this drug are 10-fold—then this drug would be an important candidate to study, in order to understand ADRs that might lead to morbidity and mortality. Therefore, the therapeutic index can be altered by both dosage and genotype.

16.2.5 Genetics of Drug Response

All the above-mentioned drug responses (*efficacy*, *therapeutic failure*, *adverse effect*, and *toxic effect*) can be regarded in genetic terms as *phenotypes* (or *traits*), which in this chapter collectively we call “PGx traits.” Any amount of success—in predicting outcome of a drug before treating the patient—will depend largely on the “genetic basis (*genotype*) of the PGx trait,” which will be influenced by the number of genes and genetic variants contributing to that *phenotype*, the allele frequency, the *effect-size* of each contributing genetic variant [166], and interactions between these genetic factors with the other environmental factors listed above.

The underlying genetic contribution to any phenotype represents the *genetic architecture*. One’s “genetic architecture” encompasses: the gene(s) and their *cis*-regulatory regions (introns, plus 5′- and 3′-flanking DNA segments near the gene), and *trans*-regulatory regions (DNA segments hundreds of kilobases away from the gene, or on other chromosomes); the number of alleles in any human population studied; distribution of allelic and mutational effects; and patterns of *pleiotropy*, *dominance*, and *epistasis* [70]. This definition of *genetic architecture* does not include *epigenetics* per se, but epigenetics can obviously influence the “mapping,” or association, of the genotype to a phenotype.

Therefore, epigenetics is not explicitly excluded from *genetic architecture*. The major feature should be whether

the “epigenetic effect” is *transgenerational* (examples are given later in the chapter) or not. If an epigenetic modification is not transgenerational, then its effect (e.g., developmental) on a phenotype pertains only to the patient himself, and it should not influence the *evolutionary trajectory* (i.e., heritable properties) of the trait. On the other hand, if the epigenetic modification has a transgenerational effect (e.g., imprinting, changes in risk of obesity, changes in risk of type II diabetes due to parent or grandparent influence), then it should be included as part of the genetic architecture. In summary, the overall outcome of a drug response will depend on the *genetic architecture* of the PGx trait—plus the *epigenetic*, *endogenous*, *environmental*, and *microbiome* factors mentioned earlier.

Recent genome-wide PGx studies have suggested (reviewed in [248]) that the genetic basis of variability in drug response can be grouped into three categories: (1) *monogenic* (Mendelian) *traits* that include many of the early examples of inherited disorders, as well as some severe idiosyncratic ADRs typically influenced by one or a few rare large-effect variants; (2) *predominantly oligogenic traits* that represent variability mainly elicited by a small number of major (PK or PD) genes; and (3) *complex PGx traits*—produced mostly by innumerable small-effect variants, together with epigenetic, endogenous, environmental, and microbiome influences. These three categories should not necessarily be considered as “distinct” from one another, but rather as an overlapping gradient. These categories will be detailed throughout the remainder of this chapter.

It is now realized that the vast majority of drug responses represent “group (3),” *PGx multifactorial traits*, similar in many ways to quantitative traits—such as, for example, height, weight, body mass index, blood pressure, and serum cholesterol levels. Multifactorial drug responses are also comparable to numerous complex diseases (e.g., type II diabetes, metabolic syndrome, coronary artery disease, bipolar disorder, schizophrenia, asthma, and cancer). Obviously, a major difference between PGx studies and complex-disease studies is that any patient who has never been challenged with a particular drug will not know his phenotype for that drug.

16.3 HISTORY OF GENETICS RELEVANT TO PGx

The complexity of understanding PGx traits, over decades of time, has paralleled our progressively better

understanding of the complexity of human genetics. Hence, most traits were regarded in Mendelian terms between 1860 and 1920, then predominantly oligogenic traits became more appreciated between 1920 and the late 1980s, and finally, consideration of the extreme complexity of multifactorial traits came to the forefront after 1990. This progression of conceptual thinking will be emphasized throughout the rest of this chapter.

At the laboratory bench, from the 1940s onward, enzyme assays of animal tissues and cell fractions became popular. In the 1940s and 1950s, tissue fractions that could be studied were simply tissue homogenates. After the invention and availability of high-speed centrifuges by the late 1950s, DME activities in microsomal versus mitochondrial versus cytosolic versus nuclear fractions became increasingly easy to study.

With advances in molecular-biology methodologies, DME genes began to be characterized in the 1980s. With the explosion in knowledge derived from the Human Genome Project (1990), elucidation of genetic differences in measurements other than DME gene studies became more commonplace. For example, receptor assays began in the late 1970s and rapidly advanced during the 1980s; transporter and ion-channel assays were initiated in the 1980s and became popular in the 1990s; analyses of signal transduction pathways, post-translational modifications, and many other subcellular processes have greatly expanded especially during the past three decades. A technical consideration in molecular biology studies during the 1980s and 1990s was that DME genes are in general smaller in length, spanning ~5–20 kb, whereas receptor and transporter genes are usually larger (~50 to >100 kb in length); cloning, sequencing, and characterization of DME genes were therefore technically easier than for the much larger genes.

16.3.1 Monogenic Traits

In the 1860s, Gregor Mendel introduced “dominant-versus-recessive” classical genetics—studying garden peas (e.g., red flower color *dominant*, white *recessive*). Whereas the F_1 cross yields all red flowers, the F_2 generation yields three red (one R/R homozygote and two R/r heterozygotes) and one white flower color (r/r). The F_2 population distribution thus illustrates the classical “Mendelian pattern of inheritance” (Fig. 16.4A).

In the early 1900s, Sir Archibald Garrod described four clinical “inborn-errors-of-metabolism”:

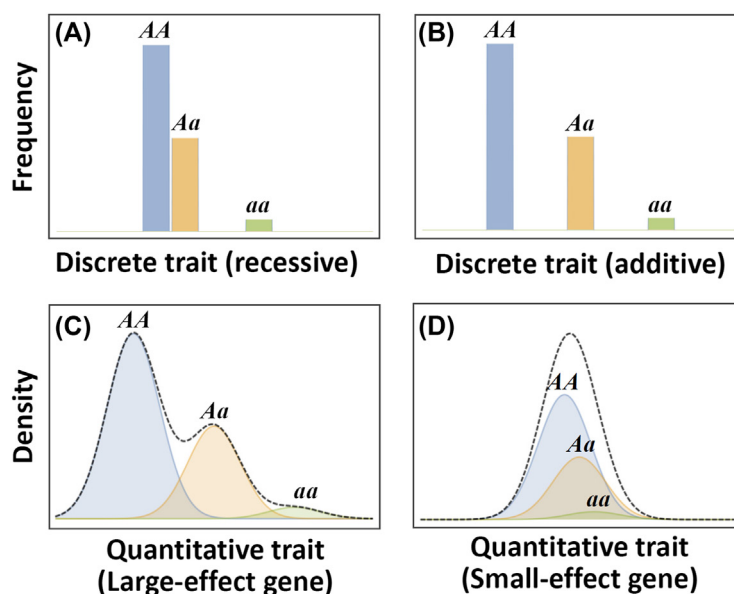


Figure 16.4 Phenotypic distribution of different traits. (A) Recessive Mendelian trait with two discrete phenotypes. (B) Distinct codominant Mendelian trait with three discrete phenotypes. (C) Quantitative trait—controlled predominantly by one large-effect gene, and unquestionably additional modifiers showing a continuous distribution with three distinct modes. (D) Quantitative trait influenced by genetic (innumerable small-effect genes)—plus epigenetic, endogenous, environmental, and perhaps microbiome effects; in other words, a polygenic trait that follows a normal distribution. (Modified from Zhang G, Nebert DW. Personalized medicine: genetic risk prediction of drug response. *Pharmacol Ther* 2017;175:75–90.)

albinism, alkaptonuria, cystinuria, and pentosuria. Each of these conditions was found to be inherited as an *autosomal recessive* trait, showing a pattern of inheritance similar to that of the white flower color of Mendel's garden pea. Garrod is recognized as spearheading the era of human genetics; the underlying tenet was “one gene, one disease,” or “one wild-type (healthy) allele, and one disease (mutant) allele.” For each pregnancy, two asymptomatic parents—who are both “carriers,” heterozygous for a disease allele—exhibit a one-in-four chance of producing a child having both disease alleles and, hence, inheriting the unwanted genetic disorder.

Each gene has two alleles in a chromosome pair—one from either parent. The Hardy–Weinberg equilibrium (HWE; $p^2 + 2pq + q^2 = 1$) was originally established to describe the expected genotype frequencies under random mating. Usually p was used to refer to the frequency of the “wild-type” allele and q for the frequency of the “mutant,” or variant, allele. For example, if $q = 0.10$, this means that the percentage of individuals homozygous for that recessive trait will be $q^2 = 0.01$, i.e., one in 100.

On the other hand, if $q = 0.01$, then the percentage of individuals who are homozygous for that recessive trait will be $q^2 = 0.0001$, i.e., 1 in 10,000.

Thus, due to early technical difficulties in recruiting and studying populations of 10,000 or greater, for most clinical studies the lowest frequency of variant alleles usually studied used to be 0.05, i.e., *common variants*. After it was realized that multiple alleles exist for every gene, q was then defined as the sum of all variant alleles. The term q has now been replaced by MAF (*minor allele frequency*), referring to “the frequency at which any particular variant allele—other than the major allele—occurs in a given population.”

Garrod's four disorders, described above, which normally skip a generation, were among the first *large-effect* single-gene mutations causing severe clinical disorders. During the next several decades, many additional examples of large-effect *autosomal recessive* clinical diseases were described (e.g., maple syrup urine disease, sickle-cell disease, Gaucher disease, phenylketonuria, cystic fibrosis, and congenital adrenal hyperplasia).

Autosomal dominant traits were also identified (e.g., Huntington disease, Marfan syndrome, neurofibromatosis, achondroplastic dwarfism, and hereditary spherocytosis); these traits typically appear in each generation, because the heterozygote manifests the disease.

X-linked recessive traits (e.g., hemophilia and red-green color blindness) were also discovered; these represent mutations on the X chromosome. Female carriers show a 50% chance of transmitting the defective allele to offspring; the disorder usually occurs in males but not females. Also, *X-linked dominant* traits were identified (e.g., incontinentia pigmenti and Coffin–Lowry syndrome)—also caused by mutations on the X chromosome; in this case, a single copy of the defective dominant allele is sufficient to cause the syndrome.

All the above-mentioned phenotypes follow bimodal distributions of Mendelian inheritance (Fig. 16.4A). Large-effect alleles can also lead to a trimodal distribution (Fig. 16.4B), in which additive traits from both parents result in an intermediate phenotype. An additional complexity to virtually all Mendelian diseases includes “modifier genes.” In other words, any number of additional genes can affect age-of-onset and/or degree-of-severity of the disorder, thereby causing overlapping of phenotypes when plotting a quantitative trait, as illustrated in Fig. 16.4C.

16.3.2 Resolution of Multifactorial Traits with Mendelian Inheritance

In addition to the relatively simple Mendelian traits having distinct patterns of inheritance, described to this point, many traits (e.g., height, body mass index, intellectual ability, and serum uric acid levels) exhibit a gradient variation in any population; moreover, one sees stronger similarity within any family than between families. The inheritance pattern of these continuous traits represented a dilemma, not readily explained by any simple Mendelian distribution.

Between 1880 and 1920, a group of biometric statisticians argued against Mendelian inheritance, because they saw that most phenotypic variation was continuous rather than bimodal or trimodal. Hence, because any phenotype of an offspring is approximately the average of that seen in his two parents—a “blending model” seemed more appropriate to explain the inheritance pattern of continuous traits.

This division between the *Mendelian inheritance school* and the *biometrics inheritance school* was most clearly resolved by Robert A. Fisher [47], a mathematician who had never obtained a graduate degree. In his breakthrough publication, Fisher presented evidence that a gradient of “continuous variation” could represent the collective result of many discrete genetic loci (Fig. 16.4D); thus, intrafamily resemblance of continuous traits could still be explained by Mendelian inheritance. It therefore became appreciated that most *human quantitative traits* (e.g., height, body mass index, serum lipid levels, and blood pressure) and *complex diseases* (e.g., type II diabetes, asthma, schizophrenia, and cancer) represent *multifactorial traits*—which reflect contributions from hundreds if not thousands of genes (polygenic), combined with additional modifying effects such as epigenetics and the other factors listed earlier.

16.3.3 Beginning of the Genomics Era

Following the advances in molecular biology, recombinant-DNA cloning, and DNA sequencing that began in the 1970s—the field of genomics quickly helped geneticists appreciate that “monogenic diseases” virtually always represent numerous “disease alleles.” Among the earliest breakthroughs was *phenylketonuria* (PKU), described as an *autosomal recessive* disorder caused by phenylalanine hydroxylase (PAH) deficiency. After cloning the *PAH* gene from one chromosome of a “carrier” parent of a PKU child [239], the Savio Woo lab reported that the *PAH* disease mutation was located at the 5′ splice-donor site of intron 12 [135]. This discovery was initially described as “*the* disease allele” for PKU. However, within months, a second mutation (this one changing an amino acid; i.e., *nonsynonymous*, or *missense*, mutation) was reported. Three years later [22], 18 distinct mutations had been identified. Soon the concept of *allelic heterogeneity* was widely accepted as “the norm” for virtually all genes in which single-nucleotide alterations, as well as insertions and deletions (indels), or duplications or inversions of DNA segments—can cause serious disease.

As of October 2017, distinct mutations (in and near the *PAH* gene) that cause variable symptoms of PKU, reported worldwide, total at least 1040 (<http://www.biopku.org/home/pah.asp>), with many reports of *ethnic differences* in allelic frequencies. Similar findings have been described for most other Mendelian disorders.

16.3.4 Single Nucleotide Polymorphisms/Variants (SNPs, SNVs)

Following initiation of the Human Genome Project in 1990, the field of genomics advanced exponentially. Since the early 1980s, yeast, fly, and worm geneticists had used the term “nucleotide substitution” for a mutation. However, in the mid-1990s several human genetics laboratories coined the term “*single nucleotide polymorphism*” (SNP), and “SNiPping through the DNA” sounded exciting. Consequently, “SNP fever” was launched. In retrospect, a better name for “SNP” would have been “single nucleotide variant,” and, in fact, in recent years, the use of “SNV” has increased in popularity.

In the mid-1990s, dozens of publications began demonstrating “statistically significant” associations (with P values < 0.05) between one or several SNPs and a complex disease—such as Alzheimer disease, type II diabetes, asthma, or autism spectrum disorder. Very soon thereafter, studies from other laboratories reported they were unable to corroborate those initial findings. It was quickly realized that the genetics of complex diseases would not be nearly as simple as that of Mendelian diseases.

Methods such as linkage studies, which had been successful in identifying major genes, were found to have limited power in detecting genes of modest effect or lower penetrance. Subsequently, a new method of “genotype–phenotype association studies” arose. Searching concurrently for all candidate genes associated with a trait would have greater power, even if this meant testing every gene in the genome. The landmark publication by Risch and Merikangas [179] described such a genome-wide approach—including the proposal to use permutation analysis and multilocus testing, with a P value of $< 5.0 \times 10^{-8}$ ($< 5.0e-08$) as the “statistically significant cut-off” for any genetic variants in the 3-billion-base-pair human haploid genome.

The vast majority of drug responses (efficacy, therapeutic failure, adverse effect, and toxic effect) involves not “Mendelian” or “predominantly oligogenic”—but rather *polygenic, multifactorial phenotypes*. What follows is therefore a brief overview of genome-wide association studies (GWAS), missing heritability, and rare versus common variants. It will be important to appreciate these concepts, in order to acquaint the reader later with the particularly problematic properties of multifactorial traits.

16.3.5 Genome-Wide Association Studies (GWAS)

Associations between five SNPs in the lymphotoxin- α gene (*LTA*) and myocardial infarction were reported, by using ~93,000 gene-based SNP markers [164]; this is purportedly the earliest published GWAS. Another early GWAS included >116,000 SNPs and demonstrated an association between the complement factor H gene (*CFH*) and age-related macular degeneration [103]. Currently, DNA-chip platforms containing 1 million to 5 million SNPs are available, easy to use, and relatively inexpensive. At the last count (<https://www.ebi.ac.uk/gwas/>), >61,000 SNP–trait associations have been reported in >3300 studies. These GWAS—with P values ranging from $< 10^{-8}$ to $< 10^{-600}$ —underscore the value of using stringent statistical significance levels when one is testing >1 million SNPs genome-wide of large cohorts comprising thousands, or even hundreds of thousands, of subjects.

GWAS quickly became much more reliable for genotype–phenotype association tests than commonly published studies involving one or several SNPs in small cohorts. These latter publications, using several dozen or even several hundred individuals, are highly prone to the statistical artifacts of *type I errors* (“false-positive”; erroneous rejection of a true null hypothesis) and *type II errors* (“false-negative”; incorrect acceptance of a false null hypothesis). Unfortunately, such published useless data continue to flood the literature, and have been variously called “the incidentalome” [105] and “ $P < .05$ false-positive/false-negative studies” [158].

Several parameters (e.g., effect-size, allele frequency, significance level, and sample size) will affect the statistical power for any genotype–phenotype association study. Clearly, the larger the numbers of cases and controls, the greater the statistical power. Moreover, as any MAF increases in frequency, fewer subjects are usually necessary in the study group, and the level of detectable contribution by an SNP to the trait will be lower. As any MAF decreases, greater numbers per group will be needed, and the level of detectable contribution by an SNP to the trait will need to be higher. GWAS will almost never have sufficient statistical power to detect epistasis (gene \times gene interactions; $G \times G$) [7,183] or gene–environment ($G \times E$) interactions [214]. GWAS studies of PGx traits are covered later in this chapter.

16.3.6 Variance Explained Versus “Missing Heritability”

Findings from GWAS eventually became unsatisfying to some investigators—who preferred clear-cut data that unequivocally quantified the total number of genes contributing to a multifactorial quantitative trait such as height or body mass index, or to explain the cause of a complex disease [62]. For most complex diseases or PGx traits, even DNA variants found together in a GWAS (e.g., using polygenic risk scores) typically explain only a small proportion of phenotypic variance (R^2) and therefore have limited clinical predictive value [132,248]. The absent proportion became known as “missing heritability” [110,133]. To make matters more unsatisfactory, although the “revealed heritability” continued to grow as the sizes of GWAS cohorts became increasingly larger, the *variance explained* rarely reached more than 20%–25% for various diseases as well as quantitative multifactorial traits [110].

Three overlapping theories to explain “missing heritability” were then proposed [58]: the “infinitesimal model” (large number of variants across the entire allele frequency spectrum of small-effects); the “rare variant model” (multiple large-effect rare variants that are poorly tagged by genotyping arrays); and the “broad-sense heritability model” (contributions from $G \times G$, $G \times E$, and/or epigenetic interactions). It is now clear that—as the GWAS cohort sizes continue to get larger—more and more *small-effect DNA variants*, in addition to credible candidate genes contributing to any multifactorial trait, will become statistically significant. However, even if the entire population on our planet could be studied, it now appears likely that *variance explained* will still not reach 100% for many of these complex diseases and quantitative multifactorial traits.

Yet, it is important to emphasize that some GWAS data have identified potential novel therapeutic targets for treating a complex disease. Similarly, some PGx GWAS data might uncover potential drug targets for improving efficacy or treating an ADR, by learning something about its mode-of-action, without necessarily understanding any precise mechanism-of-action.

16.4 EARLY PGX EXAMPLES

Due to space limitations, fewer than a dozen cases are described in this section. For additional examples, please see Table 16.2 and references therein. Histories of how

some of these PGx traits came to be discovered make for entertaining stories at cocktail hours and parties.

16.4.1 *N*-Acetylation Polymorphism (NAT2 Gene)

Originally called the “*isoniazid acetylation polymorphism*,” this PGx disorder was first noticed clinically in the 1940s when patients, who had converted from a negative to a positive tuberculin test, were routinely prescribed isoniazid. A high incidence of peripheral neuropathy was noted. This is an example of the *active principle* (parent drug) reaching toxic levels (Fig. 16.1)—when the major detoxification enzyme in the isoniazid metabolic pathway is defective.

Isoniazid was administered to volunteers, and their plasma isoniazid levels measured 6 h later (Fig. 16.5); the *bimodal distribution* found by the Victor McKusick lab [174] is most similar to that illustrated in Fig. 16.4A. The phenotypes were termed “slow acetylators” and “rapid acetylators” (i.e., slow vs. rapid plasma clearance of isoniazid). The true biological parents of slow-acetylator children were always slow acetylators—indicating that slow acetylators are homozygous for the “slow-acetylator” allele (r), whereas rapid acetylators are either heterozygous or homozygous for the “rapid” (R) allele. Hence, the *slow phenotype* is inherited as an *autosomal recessive* trait. The frequency of the r allele was ~ 0.72 in the US population that was studied [174]; i.e., if $q = 0.72$, then $q^2 = 0.52$. This means that about one in every two individuals (in this population) is homozygous for r/r , manifesting the slow-acetylator trait. This study also reflects the thinking at the time (Section 16.3.1): “one wild-type (healthy) allele, and one disease allele.”

Isoniazid *N*-acetyltransferase *variability* represents an example of a PK gene polymorphism. Three decades later, it was determined that there are two *N*-acetyltransferase genes (*NAT1*, *NAT2*), located in tandem on human chromosome (Chr) 8p22. The *NAT2* gene was responsible for the rapid- versus slow-acetylator phenotypes; when isoniazid and other arylamine substrates were studied, the *NAT2* enzyme was found to exhibit a 10-fold lower K_m than *NAT1*. Several *NAT2* slow-acetylator variant alleles were found to encode a stable protein having little or no enzymic activity [9].

A systematic allele nomenclature system for many human PK genes was initiated [30] and can now be found online (<https://www.pharmvar.org/>). A consensus nomenclature system for the *NAT1* and *NAT2* alleles

TABLE 16.2 History: Early Mendelian PGx Disorders^a

Disorder or Trait	Major Gene Known/Identified	Breakthrough Reference(s)
Phenylthiourea–nontaster	<i>TAS2R1</i>	[101,196]
Hypocatalasemia	<i>CAT</i>	[207]
Atypical serum cholinesterase	<i>BCHE</i>	[95]
Glucose-6-phosphate dehydrogenase deficiency	<i>G6PD</i>	[134]
Isoniazid slow <i>N</i> -acetylation	<i>NAT2</i>	[9,174]
Fish-odor syndrome trimethylaminuria	<i>FMO3</i>	[74,77]
Debrisoquine/sparteine oxidation poor metabolizer	<i>CYP2D6</i>	[40,67,128]
Serum paraoxonase low activity	<i>PON1</i>	[56,78]
Thiopurine methyltransferase deficiency	<i>TPMT</i>	[235]
Sensitivity to alcohol	<i>ALDH2</i>	[213]
<i>S</i> -mephenytoin oxidation deficiency	<i>CYP2C19</i>	[34,107]
Sulfotransferase deficiency	<i>SULT1A1, SULT1A2</i>	[234]
Nicotine oxidase deficiency	<i>CYP2B2</i>	[245]
P-glycoprotein transporter defect	<i>ABCB1</i>	[102]
Malignant hyperthermia	<i>RYR1</i>	[127]
Quinone oxidoreductase defect	<i>NQO1</i>	[218]
Peptide transporter defect	<i>TAP2</i>	[173]
Phenytoin, warfarin oxidation defect	<i>CYP2C9</i>	[33]
Debrisoquine ultrametabolizer	<i>CYP2D6*1XN</i>	[89]
Warfarin metabolism	<i>CYP2C9</i>	[63]
Epoxide hydrolase deficiency	<i>EPHX1</i>	[71]
Glutathione <i>S</i> -transferase null alleles	<i>GSTM1*0, GSTT1*0</i>	[98,236]
Long-QT syndrome	<i>KCNH2</i>	[23]
Dihydropyrimidine dehydrogenase deficiency	<i>DPYD</i>	[140]
Chlorzoxazone hydroxylation defect	<i>CYP2E1</i>	[76]
Peptide transporter defect	<i>TAP1</i>	[175]
Sulfonylurea receptor defect	<i>ABCC8</i>	[69]
Calcium channel defect	<i>CACNA1A</i>	[247]
Androstane glucuronosyl conjugation	<i>UGT2B4</i>	[116]
Congenital long-QT syndrome	<i>SCN5A</i>	[232]
<i>S</i> -oxazepam glucuronosyl conjugation	<i>UGT2B7</i>	[202]
Paclitaxel hydroxylase deficiency	<i>CYP2C8</i>	[26]
Chlorpyrifos oxidation deficiency	<i>CYP3A4</i>	[25]
Nicotine metabolism alterations	<i>CYP2A6</i>	[242a]
Acrodermatitis enteropathica	<i>SLC39A4</i>	[231]
Nifedipine oxidation deficiency	<i>CYP3A5</i>	[114]
Cyclophosphamide metabolism deficiency	<i>CYP2B6</i>	[109]
Hyperinsulinemic hypoglycemia	<i>SLC16A1</i>	[162]
Warfarin resistance	<i>VKORC1</i>	[181]
Hereditary folate malabsorption	<i>SLC46A1</i>	[250]
Warfarin metabolism	<i>CYP4F2</i>	[12]

^aThis list (not intended to be all-inclusive) in each case compares the consensus allele with one or more variant alleles that lead to a defective gene product. Other variants in PK and PD genes can be found at <https://www.pharmvar.org/>. The result is decreased metabolism or transporter (PK gene), or receptor or channel function (PD gene). The clinical consequence in most homozygous affected subjects is *toxicity*, due to drug accumulation with enhanced drug activity. Occasionally, decreased drug activity (*therapeutic failure*) ensues if the variant reflects ultrarapid drug metabolism or if, for activity, the drug requires metabolic conversion to an active form and this conversion is decreased in the variant. Some of the traits listed here might concern primarily environmental toxicants (e.g., *TAS2R1*, *CAT*, *FMO3*, and *PON1*) more so than prescribed drugs.

Modified from Nebert DW, Vesell ES. Chapter 19–“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.

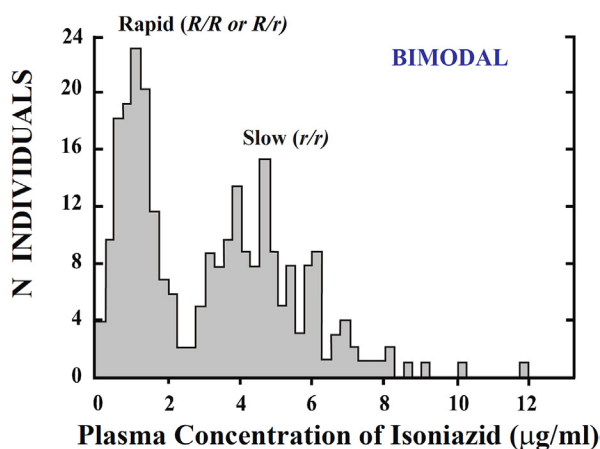


Figure 16.5 Plasma isoniazid concentrations 6h after the drug was given. Results were obtained in 267 members of 53 complete family units. All subjects received 9.8mg isoniazid per kg body weight [174]. (From previous Emery and Rimoin edition; Nebert DW, Vesell ES. Chapter 19: “Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

was developed in 1995 (http://nat.mbg.duth.gr/Human%20NAT2%20alleles_2013.htm); as of the time of this writing, it was last updated in April 2016, with more than 170 named alleles. The (rapid-acetylator) consensus allele is *NAT2*4*, with allele numbers designated (e.g., *NAT2*6A*, *NAT2*7D*, *NAT2*14J*, *NAT2*14K*, etc., with the highest allele number to date as *NAT2*27*).

16.4.2 Debrisoquine/Sparteine Oxidation Polymorphism (*CYP2D6* Gene)

In the 1970s, the debrisoquine/sparteine polymorphism (Fig. 16.6) was independently discovered by two groups. The Robert L. Smith laboratory in England [4] studied oxidative metabolism of the antihypertensive agent, debrisoquine. Soon after the drug was available in the UK, Smith noticed that debrisoquine caused a remarkably high incidence of ADRs; he correctly surmised that the combination of a narrow therapeutic index (described above; Fig. 16.3), with an underlying genetic variation in metabolism, might be responsible. Smith and three laboratory colleagues took the “recommended prescribed” dose of debrisoquine; Smith himself became hypotensive, and his urinary 4-hydroxy metabolite was ~20-fold lower than that of his three colleagues who appeared unaffected by that

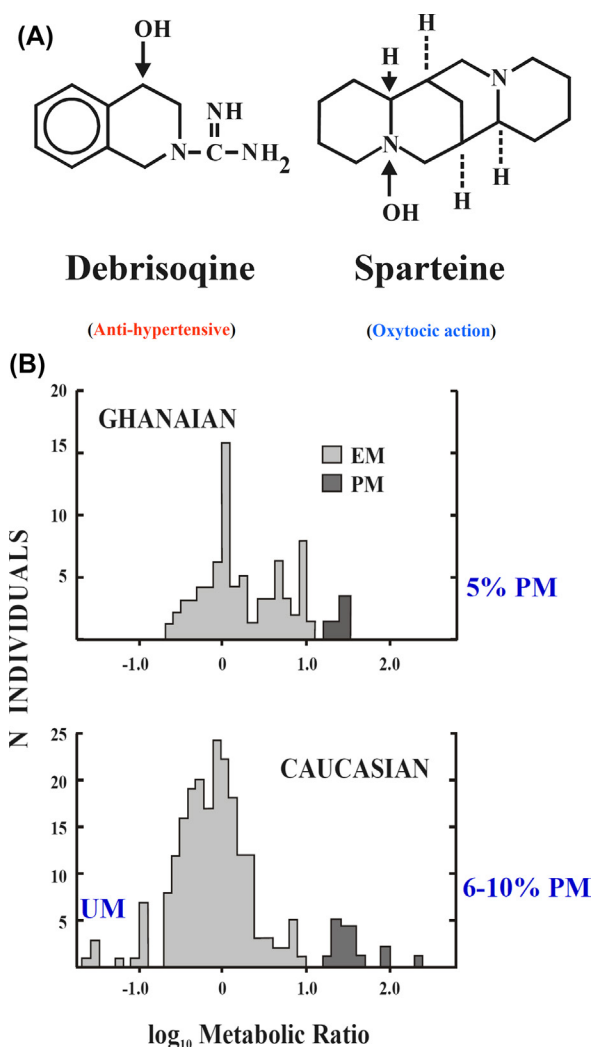


Figure 16.6 Illustration of the *CYP2D6* polymorphism. (A) Chemical structures and major metabolites of debrisoquine and sparteine, two substrates of *CYP2D6*. (B) Frequency of the efficient-metabolizer (EM) and poor-metabolizer (PM) phenotypes in a population from Ghana (top), and frequency of the EM, PM, and ultrametabolizer (UM) phenotypes in a population of Caucasians from the United Kingdom (bottom). Urinary “metabolic ratio” (MR) is defined as the “parent drug debrisoquine divided by hydroxylated debrisoquine metabolites.” Because PM individuals show less metabolism than EM and especially UM subjects, this higher ratio places PM subjects to the far right [240]. (From previous Emery and Rimoin edition; Nebert DW, Vesell ES. Chapter 19: “Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

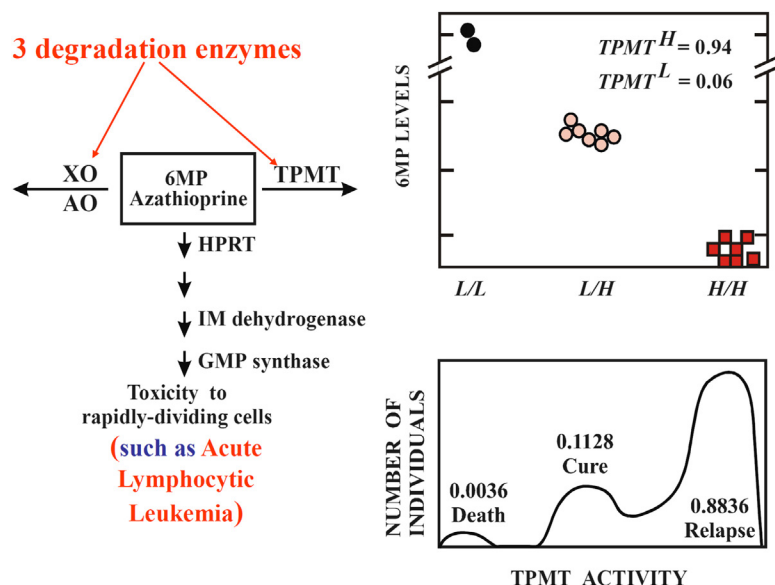


Figure 16.7 Diagram of the 6-mercaptopurine (6MP), azathioprine, or 6-thioguanine drug response phenotype. Toxicity by these chemotherapeutic agents occurs in all cells, but especially in rapidly dividing cells such as acute lymphocytic leukemia (ALL) white cells—due to disruption of purine biosynthesis. Given the “recommended” prescribed dose of 6MP, the patients’ drug response is plotted as a function of the three proposed genotypes (upper right); and the patients’ drug response is plotted as the number of individuals exhibiting toxicity as mortality, efficacy as being cured, and therapeutic failure as having a relapse of the disease (lower right). XO, xanthine dehydrogenase encoded by the *XDH* gene. AO, adenine oxidases-1 and -2 encoded by the *DUOX1* and *DUOX2* genes. TPMT, thiopurine methyltransferase. All three of these enzymes participate in detoxification of these purine analogs. About three in 1000 Caucasians are found to be homozygous for the *L/L* phenotype, 13% are *L/H* heterozygotes, and 88% are homozygous for the *H/H* phenotype [235]. (From previous Emery and Rimoin edition; Nebert DW, Vesell ES. Chapter 19-“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

dose. This is another example of a pharmacologically active drug reaching toxic levels due to insufficient detoxification to an inactive metabolite (Fig. 16.1). A larger population was screened and—similar to the isoniazid polymorphism—showed a *bimodal distribution* (much like Fig. 16.4A), separating “poor-metabolizer” (PM) from “extensive-metabolizer” (EM) subjects.

Michel Eichelbaum, for his 1975 thesis in the Hans J. Dengler lab in Germany, studied human metabolism of the oxytocic drug, sparteine. This drug was known to cause erratic and excessive uterine contractions in some, but not most, women; the urinary ratio of sparteine to the dehydrospartheines showed a *bimodal distribution* [41]. This variability in debrisoquine/sparteine metabolism is another example of a PK enzyme polymorphism.

Compared with EM-phenotype subjects that metabolize the drug 10–50 times more effectively, the PM phenotype for debrisoquine [80] occurs in 6%–10% of people of European descent (Fig. 16.6B). The incidence of the PM phenotype was found to be ~5% in an African population (Fig. 16.6B) and <1% in Asians. Subsequently, an “ultrarapid metabolizer” (UM) phenotype was also described (which actually would account for those few samples seen at the far left in Fig. 16.6B). This phenotype was found to be caused by multiple copies of the *CYP2D6* gene—from two, to as many as 13 copies [142]. The incidence of the UM phenotype is ~0.8% in northern Europeans, but 21% in Saudi Arabians, and 29% in Ethiopians [82]; the reason for this very high UM phenotype frequency in Saudi Arabia and Ethiopia is not known, but is most likely the result of either a genetic bottleneck or a selective environmental pressure such as diet.

CYP2D6 codes for the P450 enzyme responsible for the debrisoquine/sparteine polymorphism. Cloning the *CYP2D6* gene and characterization of several mutant alleles [67] represented the first time a genetic mechanism was demonstrated to explain a PGx phenotype. Different PM alleles—due to specific nucleotide changes—were shown to code for: an inactive enzyme, an unstable protein, incorrect splicing of the gene transcript, or complete deletion of the gene. All these alleles resulted in lowered, or completely absent, enzyme activity [67].

As mentioned above, a proposed unified system for naming human *CYP2D6* alleles [30] helped launch standardized allele nomenclature for many human PGx genes (<https://www.pharmvar.org/>). The *CYP2D6**1 allele is the consensus, or reference, sequence (wild-type, EM); currently, > 200 allelic variants or haplotypes have been reported—plus ~30 additional variants in which the haplotype has not yet been conclusively characterized.

The *CYP2D6* polymorphism is important in elimination of >20% of commonly prescribed drugs, as well as many over-the-counter drugs; St. John's wort and grapefruit juice are the two most popular examples. The debrisoquine "panel" now comprises >120 drugs, such as: tricyclics and other antidepressants including serotonin-reuptake inhibitors and monoamine-oxidase inhibitors; neuroleptics; antiarrhythmics and antihypertensives (including beta-blockers); the antiestrogen tamoxifen; and opiates (cf. the website designed by David Flockhart, <http://medicine.iupui.edu/clin-pharm/ddis/main-table/>).

The analgesics codeine, hydrocodone, and oxycodone are cleared via *CYP2D6*-mediated *O*-demethylation; the rate of clearance can occur ~200-fold more rapidly in EM than in PM patients. Formation of morphine, by way of *CYP2D6*-mediated *O*-demethylation of codeine, is central to codeine's analgesic PD effects. Patients who lack *CYP2D6*, or whose *CYP2D6* is inhibited, would not be expected to benefit from codeine, whereas *CYP2D6* UM-phenotype patients would be at greater risk of serious toxicity [54]. Thus, phenotyping for *CYP2D6*, and avoidance of *CYP2D6* inhibitors, has been proposed for chronic pain patients [13].

16.4.3 Thiopurine Methyltransferase Polymorphism (*TPMT* Gene)

TPMT plays a pivotal role in detoxification of 6-mercaptopurine (6MP), commonly used in chemotherapy for childhood acute lymphocytic leukemia (ALL). In the

original study of a Caucasian cohort of ~500 volunteers in Rochester, Minnesota [235], frequencies of high/high, high/low, and low/low metabolism phenotypes were reported as ~88%, ~11%, and ~0.4%, respectively. This meant that—in this population when the "commonly recommended prescribed dose" of 6MP is given—11% of patients would have high probability of being cured of their disease, 88% would have relapses in their leukemia due to undertreatment, and 1 out of ~300 patients was likely to die from 6MP toxicity (Fig. 16.7). In other words, if the metabolism of 6MP is too extensive in 88% of patients, therapeutic failure (illustrated in Fig. 16.1) would occur. It can be seen (Fig. 16.7) that distribution of this trait follows most closely that of Fig. 16.4B.

This PGx disorder is very dramatic because it can lead to life-or-death clinical situations. Accordingly, in 1994 the *TPMT* polymorphism was presented to the US Congress as "the quintessential pharmacogenetic disorder," and increased federal funding for PGx research was requested. Because the *TPMT* defect can lead to dire consequences, ALL patients are now routinely phenotyped for red-cell *TPMT* activity prior to initiation of 6MP chemotherapy. *TPMT*^{H/H} individuals generally show a favorable response with a four-times-larger dose, and *TPMT*^{L/L} patients with a 10- to 15-times-smaller dose. This regimen has resulted in substantially higher cure rates and longer survival rates for childhood ALL.

It should be emphasized that "red-cell *TPMT* activity" is a phenotyping test—not a genotyping test. The *TPMT* gene spans 26.8 kb. At least 50 allelic variants have now been identified, more than half of which alter *TPMT* activity (<https://databases.lovd.nl/shared/genes/TPMT>), resulting in very low or negligible catalytic activity. This is an example in which the phenotyping test is superior to any genotyping test, due to the ever-present possibility that a disease-causing variant lies beyond those variants that have been discovered (or any nongenetic factor, such as blood transfusion, affects the phenotype).

Azathioprine and 6-thioguanine are other *TPMT* substrates (Fig. 16.7). Azathioprine is widely used as an immunosuppressant in conditions as diverse as systemic lupus erythematosus and organ transplantation. Thioguanine is one of the agents used in treating chronic myelocytic leukemia. As with 6MP, azathioprine and 6-thioguanine can be lethal to the 1-in-300 homozygous *TPMT*^{L/L} patient—if that individual receives the "commonly recommended prescribed" dose.

Although >80% long-term survival rates in ALL have resulted from the TPMT-phenotyping test, morbidity due to drug-related myelotoxicity has continued to be problematic. Two additional relevant genes encode enzymes in the purine biosynthesis pathway: inosine triphosphatase (*ITPA*) and nudix hydrolase-15 (*NUDT15*), and it has been discovered that variants in both genes can lower the metabolism of purine analogs. One *NUDT15* variant was recently identified as a novel polymorphism linked to 6MP-induced leukopenia in inflammatory bowel disease and ALL patients [193]. There are 22 nudix hydrolase genes (*NUDT*) in the human genome. Genetic variants of *TPMT*, *ITPA*, and *NUDT15* have now been shown to affect 6MP (also, azathioprine and 6-thioguanine) metabolism, and ethnic differences of course exist in all three genes. What therefore began as a simplistic scenario—large-effect *TPMT* alleles altering an enzyme in the purine biosynthesis pathway—has now evolved into a more complex picture of PGx; this has led to further issues to prevent *drug toxicity* and the challenges of personalized medicine.

16.4.4 S-Mephenytoin Polymorphism (*CYP2C19* Gene)

CYP2C19 participates in metabolism of at least four dozen commonly prescribed drugs—including antiepileptics such as S-mephenytoin and diazepam, proton-pump inhibitors such as omeprazole, and other drugs such as amitriptyline, citalopram, and propranolol (<http://medicine.iupui.edu/clinpharm/ddis/>). The *CYP2C19* story is similar to the three above-described examples: the EM phenotype appears to be dominant over the PM phenotype, and almost always the parent drug in PM patients exhibits ADRs due to overdose (toxic levels; illustrated in Fig. 16.1).

At least 44 *CYP2C19* mutant alleles have been described, and another 40 haplotypes are not yet completely characterized (<https://www.pharmvar.org/>). The consensus, or wild-type, *CYP2C19**1 allele results in normal enzyme activity. Most of the mutant alleles (e.g., *CYP2C19**2 and *3 alleles) encode a protein having little or no activity. The *CYP2C19**17 allele is responsible for a *CYP2C19* that exhibits increased levels of enzymatic activity.

Voriconazole has become important for treatment of invasive fungal infections (e.g., aspergillosis and candidiasis). *CYP2C19* polymorphisms appear to account for the largest portion of variability in response to voriconazole. A role for *CYP2C19* genotyping to guide

the initial voriconazole dosing, followed by therapeutic-drug monitoring, has been proposed as a means of increasing the likelihood of achieving efficacy while avoiding toxicity [163].

CYP2C19 also catalyzes bioactivation of clopidogrel, an antiplatelet prodrug. Loss-of-function alleles (e.g., *CYP2C19**2) thus impair formation of the *active principle* (i.e., metabolite), resulting in decreased platelet inhibition and increased risk for adverse cardiovascular events, especially in PM patients undergoing percutaneous coronary intervention [91]. Therefore, alternative antiplatelet therapy (e.g., prasugrel, ticagrelor) is now recommended for patients who are *CYP2C19* PMs or intermediate metabolizers (IMs)—if there are no contraindications [188].

16.4.5 Glutathione S-Transferase Polymorphisms (*GST* Genes)

Encoded by PK genes, the glutathione S-transferases (GSTs) are conjugation enzymes that add glutathione to many drugs and chemicals. High GST activity can lead to rapid detoxification rates of antibiotics and chemotherapeutic agents [217]. Usually considered to be detoxification enzymes, it should be noted that GSTs can also be involved in bioactivation [144]. The *GST* gene family (<https://www.genenames.org>) comprises 17 genes in six subfamilies: *GSTA*, *GSTM*, *GSTO*, *GSTP*, *GSTT*, and *GSTZ* [155]. Human populations exhibit high frequencies for total deletion of the *GSTM1* or *GSTT1* genes (so-called “null alleles” *GSTM1**0, *GSTT1**0); the incidence of GST-null individuals ranges between 20% and 50% in East Asian populations, and varies among different ethnic populations.

During the past several decades, many dozens of genotype–phenotype associations have been reported—between cancer or toxicity and SNPs in the *NAT2* or *NAT1* genes, *CYP2D6* gene, *CYP2C19*, or in the *GSTM1**0 or *GSTT1**0 null alleles. Clearly, perhaps especially without glutathione conjugation, it seems reasonable to expect that genes encoding enzymes that detoxify drugs or environmental toxicants might be identifiable in genotype–phenotype association studies involving toxicity or cancer. However, as emphasized repeatedly in this chapter, virtually all studies involving one or a few SNPs associated with multifactorial traits such as cancer or drug toxicity in relatively small cohorts represent statistically underpowered false-positive data [105,152,158].

16.4.6 UDP Glucuronosyltransferase-1A1 Polymorphism (*UGT1A1* Gene)

UGT1A1 codes for the enzyme UGT1A1 that metabolizes irinotecan (commonly used for metastatic colorectal cancer), as well as many other drugs. Homozygotes having the poor-metabolizer *UGT1A1**28 allele were shown to be at high risk for irinotecan-induced neutropenia. In fact, on the irinotecan product label, it is recommended for *UGT1A1**28 homozygotes to decrease the starting dose of this drug. However, due to lack of sufficient prospective data, it remains uncertain whether this recommended dose-reduction will result in decreased toxicity. Combined toxicity analysis has indicated that most patients who experience grade 3 or 4 diarrhea and/or neutropenia are not homozygous for the *UGT1A1**28 allele [108].

16.4.7 Dihydropyrimidine Dehydrogenase Polymorphism (*DPYD* Gene)

The fluoropyrimidines are frequently prescribed anticancer drugs, and they are inactivated by hepatic dihydropyrimidine dehydrogenase (DPYD). As much as 5% of cancer populations exhibit DPYD deficiency. This information is considered to have practical value—and might even be cost-effective—for patients receiving these anticancer fluoropyrimidine substrates; the *DPYD**2A low-activity allele is highly associated with 5-FU-induced severe and life-threatening toxicity [35]. There is convincing evidence to implement prospective *DPYD* genotyping with an up-front dose adjustment in DPYD-deficient patients [124].

In each of these examples in which the active parent drug causes toxicity if not adequately metabolized, or if the PGx assay reveals diminished enzymatic activity or deficiency, then lower initial doses, or alternate drugs, are usually recommended.

16.4.8 Abacavir-Induced Hypersensitivity (*HLA* Loci)

Abacavir is an HIV-1 nucleoside-analog reverse-transcriptase inhibitor used to treat human immunodeficiency virus (HIV) infections. In an early example of identification of a PGx disorder by a “candidate-gene-region” study, 18 abacavir-treated patients exhibited a life-threatening hypersensitivity syndrome, out of 185 patients receiving abacavir [130]. Compared with 167 abacavir-resistant controls, SNP-typing of loci in the major histocompatibility complex (MHC)

region revealed a very strong association [odds ratio (OR)=117] with the *HLA-B**57:01 allele, and also in combination with the *HLA-DR7* and *HLA-DQ3* loci (OR=73).

These *HLA* loci represent immune-response genes. They encode specific cell-surface molecules responsible for presentation of endogenous peptides to cells of the immune system. This study of a highly significant association of abacavir-induced hypersensitivity with *HLA-B**57:01 [130] was later confirmed in a much larger double-blind prospective randomized study that involved ~2000 patients from 19 countries [131]. However, even though the OR for the abacavir-hypersensitivity syndrome is very large, penetration of this phenotype is very weak—and therefore PGx testing would never be economically feasible.

16.4.9 Warfarin Polymorphisms (*CYP2C9*, *VKORC1*, and *CYP4F2* Genes)

Up to this point, our examples have been predominantly single-gene large-effect responses of a specific subpopulation to a drug, or to different types of drugs that are substrates of the encoded enzyme. Optimizing warfarin, coumarin, or acenocoumarol dosage for anticoagulation therapy is clinically of extreme importance, because of the dangers of either too little drug (causing clotting) or too much drug (causing unwanted hemorrhaging). Warfarin metabolism is an early oligogenic example in which a substantial contribution of several genes was found—resulting in rapid versus slow PK and/or PD of coumarins. *CYP2C9* [63] and *VKORC1* [24] polymorphisms were independently discovered in candidate-gene studies. These two genes were subsequently confirmed in GWAS [20,208], along with the additional discovery of *CYP4F2* [12].

The *CYP2C9* and *CYP4F2* polymorphisms represent (relatively) large-effect PK genes coding for enzymes involved in metabolism. *VKORC1* is considered to be a large-effect PD gene. The *VKORC1* gene codes for vitamin K-epoxide reductase complex subunit-1, which is targeted directly by coumarins. Coumarins are considered as vitamin K antagonists [84], and, as such, are potent inhibitors of this reductase complex; inhibition of the complex results in depletion of reduced vitamin K, which is essential for normal coagulation. The epoxide reductase has therefore been considered as a drug target for coumarins—especially because, except for

the vitamin K substrate, the encoded epoxide reductase appears not to metabolize any other drug.

When variants of all three genes are combined, the combination provides ~45% of *variance explained*, i.e., the patient's total variability in drug response that can be accounted for [27]. Thus, the remaining ~55% of variability in coumarin response must originate from contributions by other genes and/or environmental factors. This oligogenic example, over a decade ago, was an excellent illustration of the growing complexity of the goal of predicting PGx phenotypes such as *drug efficacy* or *toxicity*—which has become increasingly appreciated during this past decade.

16.4.10 Ethnic Differences in Drug Metabolism

Numerous examples of *ethnic differences* in drug response are known [94] and several have been mentioned above. In several of these cases, the interethnic variability is sufficiently striking that PGx assays for a drug, or family of drugs, are recommended for one ethnic group, while being of considerably less importance to another ethnic group. Table 16.3 lists some examples of ethnic differences.

Among the earliest important ethnic differences discovered was the rapid-acetylator versus slow-acetylator phenotype (Table 16.3). Frequencies of the slow-acetylator allele range worldwide from less than 10% in Japanese populations to more than 90% in some Mediterranean peoples.

Mitochondrial aldehyde dehydrogenase-2 (ALDH2) deficiency is an interesting early example (Table 16.4). The incidence of the *ALDH2* Glu504Lys mutation ranges between 25% and 45% in most East Asian populations, yet is virtually never seen among Africans or Caucasians. A purportedly different *ALDH2* allele, also resulting in a lack of ALDH activity, was found in South American Amerindians [94], probably due to a founder effect or genetic bottleneck. These data led to speculation that ALDH2 deficiency might have arisen only in populations that traditionally have not been commonly exposed to ethanol. In other words, East Asians have boiled water for at least the past 30 centuries, compared with African and Caucasian populations that had used alcohol for enjoyment and preservation of foods for many earlier centuries.

For more than 100 years, nitroglycerin has been clinically used to treat angina and heart failure; it was recently

TABLE 16.3 Frequency of *N*-Acetylator *NAT2* PM Phenotypes in Different Ethnic Populations

Ethnic Population	No. of Studies	Frequency of PM Phenotypes
Japanese	7	0.09
Eskimo	4	0.23
South Pacific Islands	5	0.35
Korean/Chinese	14	0.37
North and South Amerindian	10	0.50
African ^a	19	0.71
Central and West Asian	22	0.74
European	50	0.75
Egyptian	2	0.96

^aExcluding the !Kung Bushmen of Southern Africa, in which the PM frequency is 0.18.

Data modified and condensed from Kalow W, Bertilsson L. Interethnic factors affecting drug response. *Adv Drug Res* 1994;25:1–53.

TABLE 16.4 Distribution of the *ALDH2* Deficiency Phenotype in Different Ethnic Populations

Ethnic Population	Percent Having <i>ALDH2</i> Deficiency ^a
Japanese	44
Central, East, and Southeast Asian	25–50
South Amerindian	40–45 ^b
North Amerindian	2–5
European, Mideast, and African	<0.1

^aThe mutation in Asians and North American Amerindians appears to be solely Glu504Lys, which causes a complete loss of ALDH2 activity in that subunit. Interestingly, the Lys504 allele contributes in large part to the lack of a clinically efficacious response to sublingual nitroglycerin [117]; this is particularly important among East Asian populations, 30%–50% of whom carry the *ALDH2**2 mutant allele. ALDH comprises four subunits; if one or more of the subunits are encoded by the *ALDH2**2 allele, then the entire tetramer is inactive. Thus, the *ALDH2**1/*2 heterozygote exhibits $(1/2)^5 = 1/16$, or 6.25%, of activity of the *ALDH2**1/*1 homozygous individual.

^bMutation in Amerindians from South America is purportedly different from that in Asians; to our knowledge, however, the DNA sequence of the Amerindian *ALDH2* allele from South America has never been published.

Data modified and condensed from Goedde HW, Agarwal DP. Aldehyde oxidation: ethnic variations in metabolism and response. *Prog Clin Biol Res* 1986;214:113–38.

discovered that mitochondrial ALDH2 is responsible for formation of nitric oxide [126], the metabolite required for nitroglycerin efficacy. Subsequently, it was shown that the catalytic efficiency of nitroglycerin metabolism of the consensus allele *ALDH2*1* (encoding the Glu504 protein) is ~10-fold higher than that of the mutant Lys504 enzyme. It has thus been recommended that *ALDH2*-genotyping be considered when administering nitroglycerin to patients—especially Asians, in whom 30%–50% carry the inactive *ALDH2*2* mutant allele [117].

Ethnic differences in *CYP2D6* alleles (<https://www.pharmvar.org/>) and many other P450 genes exist. *CYP2C19* is one of the most important PGx examples of ethnic variability. *CYP2D6* and *CYP2C19* were discussed previously, and their wide range of drug substrates is what makes any substantial ethnic differences more relevant to clinical pharmacology. Ethnic differences for the *CYP2C19* enzyme encoded by this PK gene are very striking; e.g., the East Asian PM subset is ~33% and the Oceanian PM subset >50%, whereas the Caucasian PM subset represents <6% [143].

The incidence of the *CYP2C19*2A* and **2B* alleles (responsible for splicing defects) is 2%–5% in Caucasians, yet 20%–30% in Asians. This is an excellent example in which the physician must be more careful in prescribing “any drug in the *CYP2C19* repertoire” for patients of Asian ancestry than for those having primarily Caucasian ancestry. Similarly, East Asian drug companies have recognized the importance of the *CYP2C19* polymorphism in their populations more so than companies in predominantly non-Asian countries.

16.4.11 Why Might Ethnic Differences in Drug Metabolism Exist?

It had been proposed [66] that PGx genes in animals might have originated because animals eat plants, and plant metabolites are similar in their intramolecular composition and molecular size to drugs. The Great Human Diaspora “Out of Africa” occurred over many tens of thousands of years of evolution. Ethnic groups originated as populations living in geographic isolation for >10,000 years and subsisting on distinct foods and diets relevant to that geographic region and their culture (lifestyle). It seems feasible that the Great Human Diaspora would explain today’s observations of interethnic differences in drug metabolism and drug response [153].

Examination of *NAT2* variants in ~15,000 subjects from 128 populations [182], revealed a higher prevalence

of the slow-acetylator *NAT2* phenotype in populations practicing agriculture-and-herding, when compared with those relying mostly on hunting-and-gathering. Perpetuation of mutant alleles, resulting in monogenic disorders and in some PGx genes, can also result from enhanced resistance in the heterozygote to certain infections [153,238].

How long might it take for a genome to adapt to dietary selective pressures? By means of mutations (nucleotide substitutions, indels, inversions, duplications, crossing-over events, etc.), genetic drift, and natural selection, new gene alleles will become fixed and passed on to the next generation—if the new allele confers reproductive and ecological advantages, or is neutral (i.e., no detrimental effect), to the species. The response of a genome to environmental pressures, over a minimal number of generations, has been variously described as *molecular drive* [38], *meiotic drive* [211], cryptic genetic variation [59], and *decanalization* [57]. This gene–environment response most likely involves both genetics and epigenetics and probably also plays a role in the processes of drug- and/or plant-induced efficacy and toxicity.

Whereas mutational changes happen slowly over many dozens of generations, epigenetic changes (i.e., no alterations in DNA sequence) occur rapidly—in response to severe environmental challenges [104,178]. If the population of a species decreases dramatically, allelic frequencies in that population will change even more radically. Within a population, the emergence of individuals resistant to environmental changes has been shown experimentally, in various organisms from prokaryotes to insects and vertebrates, but why this happens remains basically obscure.

For example, Atlantic tomcod fish (*Microgadus tomcod*), living in the polluted Hudson River for 50–100 years (50–100 generations), have developed resistance to polychlorinated biphenyls (PCBs); a 6-nucleotide deletion in the *AHR2* gene—encoding an aryl hydrocarbon receptor-2 protein having poor-affinity for planar PCBs—was the basis for PCB resistance [237]. Depending on the organism studied, between nine and 45 generations appear to be required [151]. For humans, nine to 45 generations would extrapolate to times between ~200 and ~1000 years.

The five major *Homo sapiens* geographically isolated subgroups [251] are estimated to have diverged from one another between 20,000 and 45,000 years ago.

Based on studies in various animal species, 10,000 years of geographic isolation are believed to be sufficient for striking differences in PGx genes to have arisen from selective pressures—such as tribal differences in diet, or exposure to other environmental signals (e.g., altered climate and altitude). For example, it would seem likely that a tribe subsisting for ~10,000 years on a diet principally of goat meat and milk products on a high desert, might have different selective pressures than that of a tribe eating tropical fruit and fish at the seashore. Among the best examples would be the dramatic ethnic differences seen in lactase deficiency and response to milk products (reviewed in [86]). During a time period of ~10,000 years, it therefore seems feasible that striking differences in allelic frequencies of PGx genes would likely have arisen.

16.5 PHARMACOGENOMICS

As stated early in this chapter, *pharmacogenomics* began as a field distinct from pharmacogenetics and was defined as “the study of how drugs interact with the *total genome*, to influence biological pathways and processes” [148]; in other words, “drug–genome interactions.” This field is expected to help identify new druggable targets and thus be instrumental in designing new drugs.

Advances in genome technologies, along with the development of statistical packages for large-scale data analysis, have enabled huge collaborative groups of investigators to carry out GWAS involving drug responses. Since 2007, dozens of PGx GWAS—having sufficiently large numbers in their cohorts—have led to the identification of numerous genes having SNPs associated with various responses of drug efficacy, ADRs, and toxicity. GWAS are similar to “fishing expeditions,” i.e., they are hypothesis-free and thus do not require a priori assumptions about chromosomal locations of functional variants [137,201]. Accordingly, GWAS have provided an unbiased, powerful tool for the systematic discovery of genetic variants associated with a growing number of PGx traits [145].

16.5.1 ADRs can Be Indistinguishable From Complex Diseases

As we introduced early in this chapter, recent genome-wide PGx studies have refined our thinking [248]. We believe it is reasonable to classify the genetic basis of variability in drug response—together with epigenetic, endogenous,

environmental, and microbiome effects—into three groups. Moreover, these groups should not be considered separate from one another, but rather a gradient. (1) We first presented more than a half-dozen *monogenic* (Mendelian) *traits*, and described early classical examples of severe toxicity or idiosyncratic dose-independent ADRs, typically reflecting one or a few rare coding large-effect variants, usually in PK genes or immune-response genes. (2) Among the *predominantly oligogenic traits*, we presented warfarin metabolism as an excellent example of variability caused in moderate amounts mostly by a relatively small number of large-effect (PK or PD) genes. (3) *Complex PGx traits*, which comprise the remainder of this chapter, represent typically the contribution of large numbers of small-effect variants; discovery of such genes contributing to complex PGx traits will almost always require GWAS to enable identification.

The clinician should be aware that idiosyncratic dose-independent ADRs and complex diseases are traits that can often be difficult to distinguish from one another (Fig. 16.8). In the case of a complex disease, after a stimulus (or stimuli) that triggers the disorder, there is a cascade of downstream effects leading to the phenotype (the complex disease), which can be rapid, but also might develop slowly over many years. In the case of an ADR, the drug elicits the stimulus (or stimuli) that sets into motion the downstream cascade of effects causing the phenotype (the ADR); this usually occurs in a matter of a few hours, days, or weeks. Table 16.5 is a partial list of common ADRs known to occur in subpopulations of patients—after they have received the recommended dosage of a commonly prescribed drug.

How does this happen? This is among the most intriguing mysteries in clinical pharmacology. How does a small-molecular-weight drug—given to some patients, but not the majority of patients in any cohort—cause an ADR that is often indistinguishable from the appearance of a complex disease? For example, sitagliptin, a reversible inhibitor of dipeptidyl-peptidase-4, is approved by the FDA to treat type II diabetes; yet, a small subset taking the recommended prescribed dose develops acute pancreatitis (reviewed in [187]). In a meta-analysis study of >50,000 patients with type II diabetes and >270,000 controls [122], it was shown that treatment with LDL-cholesterol-lowering drugs was correlated with higher risk of type II diabetes—and this trait was significantly associated with SNPs in or near the *NPC1L1* gene and at least four other genes.

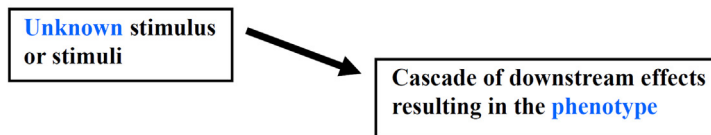
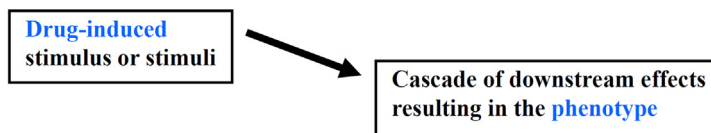
Human complex diseases:**Drug-induced ADRs:**

Figure 16.8 Similarities between human complex diseases and drug-induced ADRs, both in their genetic origins and their phenotypic manifestations. (From previous Emery and Rimoin edition; Nebert DW, Vesell ES. Chapter 19-“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.)

Treatment of malaria, lupus erythematosus, or rheumatoid arthritis with hydroxychloroquine can lead to acute pancreatitis. In a subpopulation of patients receiving many psychotropic drugs (e.g., valproic acid), undesirable weight gain occurs as a dose-independent ADR; in a smaller subset, hepatic steatosis as a dose-independent ADR has also been found (reviewed in [2]). In a small subpopulation of patients taking bisphosphonates for osteoporosis, increased risk of esophageal and gastric cancer has been repeatedly reported; however, a thorough review and meta-analysis of this association [242] has not found any significantly increased risk.

Complex PGx traits include not only dose-independent idiosyncratic ADRs and many instances of drug toxicity, but also drug efficacy. Hence, it seems reasonable to conclude that attempts to dissect differences in drug response (*efficacy, therapeutic failure, ADRs, toxicity*) will be very similar to attempts to dissect differences in complex diseases; in other words, both represent polygenic multifactorial traits.

16.5.2 Genome-Wide Association Studies of ADRs

It goes without saying that any recruitment of a sufficiently large cohort to study a PGx trait (i.e., patients receiving the same drug and preferably a similar dose of that drug) will be far more difficult than recruiting

sufficient numbers of patients or volunteers to study a complex disease, such as type II diabetes—or a quantitative trait such as height or body mass index. Nonetheless, during the past decade, a growing number of GWAS (Table 16.6) have identified novel genetic loci associated with severe ADRs [28].

16.5.2.1 Statin-Induced Myopathy

Among the earliest PGx GWAS was simvastatin-treated patients that had developed myopathy [121]. Initially, the authors had screened ~300,000 markers in 85 cases and 90 controls selected from a clinical trial of ~12,000 participants; their findings were then replicated in a second cohort of ~20,000 subjects. (For a growing number of journals, “replication in a second cohort” has become a standard requirement for acceptance of a GWAS publication.) Simvastatin-induced myopathy was found to be associated with a mutation that changed an amino acid (i.e., a *nonsynonymous variant*) in the *SLCO1B1* gene. Being homozygous for this variant was found to confer an odds ratio (OR) of 16.9 for patients that developed simvastatin-caused myopathy, compared with patients not having this variant and receiving simvastatin without developing myopathy.

The *SLCO1B1* gene encodes an organic-anion SLC transporter, presumed to function in cellular uptake of statins [106], and therefore it appears

TABLE 16.5 Archetypal ADRs That can Occur in Patients Receiving Commonly Prescribed Drugs

Organ or System	Possible ADRs (Multifactorial Traits)
Central nervous system	Headache; fainting; hallucinations; stroke; mental or mood changes (e.g., new or worsening anxiety; nervousness; agitation; suicidal thoughts; confusion depression; restlessness; sleeplessness; inability to concentrate); memory loss; new or worsening nightmares; tremor; seizures; irreversible brain damage; transient psychotic episodes; toxic psychosis; ataxia; cogwheel rigidity; speech disorder (dysphasia); irritability; panic attacks; blacking out due to hypotension; progressive multifocal leukoencephalopathy
Eye	Acute-angle closure glaucoma; changes in vision; blurred vision; loss of vision; photosensitivity; phototoxicity; dry eye; periorbital edema; increased tearing; corneal keratitis; cataracts
Gastrointestinal tract	Constipation; severe or persistent diarrhea; gas; nausea; vomiting; severe or persistent stomach pain/cramps; difficulty in swallowing; abdominal cramps; bloody or tarry stools; heartburn (dyspepsia); indigestion; ulcers of mouth, esophagus or colon; gingival overgrowth
Heart	Hypotension; hypertension; chest pain; shortness of breath; heart attack; angina; heart failure; edema; swelling below the knees; heart block; cardiac arrhythmias; sinus tachycardia; palpitations; atrial fibrillation; ventricular fibrillation; postural hypotension; cardiac tamponade
Hematological system	Anemia; unusual bruising or bleeding; excessive bleeding (sometimes can be fatal); clotting disorders; bone marrow suppression
Immune system	Severe allergic reactions (e.g., rash; hives; pruritis; difficulty breathing; tightness in chest; swelling of the mouth, face, lips, or tongue); unusual hoarseness; immunosuppression; autoimmune diseases
Inner ear	Tinnitus; dizziness; light-headedness; hearing loss
Kidney	Decreased or painful urination; changes in glomerular filtration rate; changes in creatinine levels; renal insufficiency or failure; hypertension; hypotension; hyponatremia; hyperkalemia; nephrogenic systemic fibrosis
Liver	Symptoms of liver problems (e.g., dark urine; loss of appetite; pale stools; jaundice); chemical hepatitis; liver failure
Musculoskeletal system	Pain, soreness, redness, swelling, weakness, or bruising of a tendon or joint area; muscle pain or weakness; osteonecrosis of the jaw; ankle swelling; inability to move or bear weight on a joint or tendon area; irreversible tendon damage; spontaneous tendon ruptures; tremor; ankylosing spondylitis; rhabdomyolysis; osteopenia; osteoporosis; leg cramps
Pancreas	Symptoms similar to diabetes (e.g., high blood sugar; dizziness; fainting; rapid breathing; flushing; increased thirst, hunger, or urination; increased sweating; vision changes); hypoglycemia; acute pancreatitis
Peripheral nervous system	Symptoms of nerve problems (e.g., changes in perception of heat or cold; decreased sensation of touch; unusual burning, numbness, tingling, pain, or weakness of the arms, hands, legs, or feet); tremors; irreversible peripheral neuropathy
Reproductive system	Vaginal discharge, irritation, or odor; hypomenorrhea; hypermenorrhea; dysmenorrhea; erectile dysfunction; loss of libido; increased libido (males or females); priapism
Respiratory tract, lower	Shortness of breath; asthmatic-like wheezing; pulmonary edema; pulmonary thrombosis; lower respiratory infection; cough
Respiratory tract, upper	Fever, chills, sore throat, or unusual cough; runny nose; dry nose and throat; stuffy nose and congestion; upper respiratory infection; malignant hyperpyrexia
Skin	Moderate or severe sunburn; red, swollen, blistered, or peeling skin; irreversible skin damage; Stephen-Johnson syndrome; toxic epidermal necrolysis; hives; eczema; edema
“Systemic” (?)	Esophageal cancer, urinary bladder cancer, and other types of cancer; signs and symptoms of cardiac failure; facial flushing; obesity; anorexia; weight loss; weight gain; bigorexia; drug–drug interactions; food–drug interactions; death
DNA changes	DNA damage; mutations; increased risk of cancer

Modified from Nebert DW, Vesell ES. Chapter 19: “Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and Practice of medical genetics. 6th ed. Oxford: Academic Press; 2013. pp. 1–27.

TABLE 16.6 Selected PGx GWAS With Genome-Wide Statistically Significant Findings of $P < 5.0 \times 10^{-8}$ ($P < 5e-8$)^a

Drug	Response	Gene(s)	Effect	P-Value	Year	Reference
ADR (Resembles Mendelian Trait With Incomplete Penetrance)						
Statin	Myopathy	<i>SLCO1B1</i>	OR = 4.5	4.00e-09	2008	[121]
Flucloxacillin	DILI ^b	<i>HLA-B*57:01</i>	OR = 80.6	8.70e-33	2009	[31]
Lumiracoxib	DILI	<i>HLA-DRB1*15:01</i>	OR = 5.0	6.80e-25	2010	[192]
Carbamazepine	Hypersensitivity	<i>HLA-DRB*15:02</i>	OR = 12.4	3.50e-08	2011	[138]
Heparin	Thrombocytopenia	<i>TDAG8</i>	OR = 18.6 ^c	3.18e-09	2015	[97]
Anthracycline	Cardiotoxicity	<i>RARG</i>	OR = 4.7	5.90e-08	2015	[3]
Lapatinib	DILI	<i>HLA-DRB1*07:01</i>	OR = 14 ^d	7.8e-11	2016	[165]
Asparaginase	Hypersensitivity	<i>NFATC2</i>	OR = 3.11	4.10e-08	2015	[46]
Cisplatin	Hearing loss	<i>ACYP2</i>	HR = 4.5	3.90e-08	2015	[243]
Dosage/Efficacy (Having Major Gene Effect)						
Warfarin	Dosage	<i>VKORC1</i> and <i>CYP2C9</i>	34%	6.20e-13	2008	[20]
Warfarin	Dosage	<i>VKORC1</i> , <i>CYP2C9</i> and <i>CYP4F2</i>	43.1%	<1e-78	2009	[208]
Acenocoumarol	Dosage	<i>VKORC1</i> , <i>CYP2C9</i> , <i>CYP4F2</i> and <i>CYP2C18</i>	48.80%	<5e-8	2009	[212]
PegIFN-alpha	Viral clearance	<i>IL28B</i>	OR = 7.3	1.00e-38	2009	[55]
		<i>IL28B</i>	OR = 1.98	9.30e-09	2009	[204]
		<i>IL28B</i>	OR = 27.4	2.70e-32	2009	[210]
Citalopram and escitalopram	Concentration	<i>CYP2C19</i> and <i>CYP2D6</i>	NA	<5e-8	2014	[88]
	TPMT activity	<i>TPMT</i>		1.2e-72	2017	[209]
Clopidogrel	Platelet reactivity	<i>CYP2C19</i>		5.1e-40	2018	[6]
Clopidogrel	Clopidogrel active metabolite levels	<i>CYP2C19</i> and two other loci		9.5e-15, 3.3e-11 1.3e-8	2017	[5]
Complex PGx Traits (Influenced by Multiple Genetic Variants Each With Small Effect)						
Statins	LDL response	<i>SORT1/CELSR2/PSRC1</i> , <i>SLCO1B1</i> , <i>APOE</i> , <i>LPA</i>	1.3%–5.2% ^d	<5e-8	2014	[172]
Metformin	HbA1c <7%	<i>ATM</i>	OR = 1.35	2.90e-09	2011	[60]
		<i>SLC2A2</i>	0.17% greater	6.60e-14	2016	[252]
Antidepressant	Questionnaire physician's opinion	<i>HPRT4</i>	OR = 1.36	5.03 e-08	2015	[8]
Methadone	Dosage	<i>OPRM1</i>		2.8 e-8	2017	[195]
Corticosteroids	Adrenal suppression	<i>PDGFD</i>	OR = 4.05	3.5 e-10	2018	[73]

^aCondensed from [248] and updated.^bDILI, drug-induced liver injury.^cOR of homozygous.^dPercentage of extra LDL-C lowering in carriers versus noncarriers of the SNP.

to be a credible candidate gene. Association of this *SLCO1B1* variant with simvastatin-triggered myopathy (Table 16.6) was subsequently reproduced in further clinical studies [230]. Consequently, a simvastatin-dosing regimen, based on this *SLCO1B1* genotype, has been recommended [176] by the Clinical Pharmacogenomics Implementation Consortium (CPIC). Interestingly, for whatever reason, association of this *SLCO1B1* variant with myopathy has been less conclusive for other statins [161]. Later PGx GWAS also showed that the same gene variant is a primary determinant of methotrexate pharmacokinetics and clinical effects [177,219].

16.5.2.2 Drug-Induced Liver Injury (DILI) and/or Hypersensitivity

Another early PGx GWAS identified the *HLA-B*57:01* allele as a major risk factor for flucloxacillin-induced liver injury [31]; this ADR is quite rare (~8.5 in every 100,000). The authors found a strong correlation with a variant SNP in complete *linkage disequilibrium* (LD) with the *HLA-B*57:01* allele, having an OR of 80.6 (Table 16.6). Despite the strong association, however, only 1 in every 500–1000 flucloxacillin-treated patients having this allele will develop liver injury; this infrequent occurrence of DILI limits the clinical utility, because the test will have a very high rate of false positives. This genotype–phenotype association study is an excellent example in which *incomplete penetrance* makes it difficult for the physician who must decide, on a patient-by-patient basis, who should receive this drug.

Other DILI traits strongly correlated with different *HLA* haplotypes (Table 16.6) were reported for patients taking lumiracoxib [192] and lapatinib [165]. Another dose-independent ADR phenotype associated with an *HLA* locus was carbamazepine-induced hypersensitivity [138]. Yet, the clinically predictive values, for typing each of these *HLA* alleles associated with these ADRs, are regrettably very low (reviewed in [29]).

In addition to relationships with the *HLA* region, PGx GWAS have identified a number of additional associations with various ADRs [46,97,243]; remarkably, none of these loci (Table 16.6) would have been suspected—based on established data about known functions of each drug's PK and/or PD. Although these GWAS probably do not have immediate clinical utility, they do identify new biological pathways that might underlie the mode-of-action or mechanism of toxicity for each of these drugs.

16.5.2.3 Drug-Induced QT-Interval Prolongation and Osteonecrosis of the Jaw

It should be emphasized that not all PGx GWAS have produced robust associations. For example, a number of GWAS of drug-induced QT-interval prolongation did not produce consistent results (reviewed in [160]). On the other hand, bisphosphonate-induced osteonecrosis of the jaw was reported to be associated with *CYP2C8* variants [184]; *CYP2C8* is a plausible candidate because it metabolizes several anticancer drugs. However, additional studies [42,99,203] have not been able to reproduce those data. Osteonecrosis of the jaw is well known to display an extremely heterogeneous range of phenotypes—a phenomenon seen in numerous other complex diseases; thus, unless one has a very large sample size, as well as an unequivocal diagnosis of the phenotype being studied [225], identification of statistically significant loci that can be replicated will continue to be problematic.

16.5.3 Genome-Wide Association Studies of Drug Efficacy

16.5.3.1 Efficacy of Anticoagulants

As discussed earlier, the coumarins including warfarin are widely used anticoagulants; however, their regimens are complicated—because of the wide-ranging interindividual variability in dosage and narrow therapeutic “window” (as illustrated in Fig. 16.3). Trying to perfect the dosage of coumarins including warfarin is therefore a very active area of PGx research. Before the GWAS era, candidate–gene associations and molecular studies had already identified the *CYP2C9* and *VKORC1* genes as being important, as detailed in Section 16.4.9.

Several GWAS (Table 16.6) have now validated those associations of *CYP2C9* and *VKORC1* with warfarin efficacy. Moreover, these GWAS have provided more accurate estimates of the amounts of contribution of these two genes in defining the *variance explained*: the contribution of *VKORC1* is ~30% and that of *CYP2C9* is ~12% [20,208]. In addition, the latter GWAS, having a larger sample-size [208], provided sufficient statistical power to uncover a third locus with smaller effect—i.e., the contribution of 1.1% by the *CYP4F2* gene.

Nonetheless, due to its relatively small contribution, the *CYP4F2* polymorphism has not been included in most dosing-recommendation algorithms [90,171]. For other coumarin anticoagulants, GWAS have also identified similar genetic determinants [212]. Furthermore,

GWAS of anticoagulants have been reported to be different in several different ethnic groups [14,167,169].

16.5.3.2 Efficacy of Hepatitis C Virus Infection

The study of variability in response to interferon- α treatment of hepatitis C virus (HCV) infection is another GWAS example of drug efficacy. Three independent studies [55,204,210] described DNA variants near the interleukin-28B gene (*IL28B*) associated with response to interferon- α treatment. The degree of efficacious response was measured by the *sustained virological response* (SVR), i.e., the absence of detectable virus at the end of clinical follow-up. The same variant was also reported to be associated with spontaneous clearance of HCV [215]; these data suggest that the variant allele might directly influence the host immune response—with or without drug intervention.

The advantageous *IL28B* allele (a cytidine at nucleotide position rs12979860) is more frequent in Caucasians and Asians than in Africans; this finding would explain an earlier report [17] that, in trials of interferon- α treatment, HCV-infected African-Americans show relatively worse outcomes than Caucasians or Asians. Despite an estimated large effect size (OR=7.3), the imperfect predictive power—combined with the lack of alternative treatment regimens—has resulted in limited immediate clinical utility for determining the *IL28B* genotype, with regard to decisions involving personalized treatment [79].

An additional caveat is that the *direct antiviral agents* (DAAs), in combination with PEGylated interferon/ribavirin, represent the new standard of HCV treatment; this makes it important to re-evaluate the clinical utility of the *IL28B* genotype [216]. Whereas *IL28B* genotyping will continue to provide useful clinical information with regard to the predicted treatment response, it is likely that the *IL28B* genotype assay will eventually lose its usefulness—primarily because HCV therapy has begun to distance itself from interferon-based regimens [87,119].

16.5.3.3 Mutational Landscape of G-Protein-Coupled Receptor (GPCR) Drug Targets

In an avante gard data-mining in silico approach—rather than GWAS—the first study of its kind searched for DNA variants in or near each of the 108 G-protein-coupled receptor genes (*GPCRs*); these 108 genes are known targets of 475 Food and Drug Administration (FDA)-approved drugs, which account for a global sales volume of > US\$180 billion annually [72]. Each

of the genomes of almost 68,500 individuals was then separately investigated for missense variants in and near each of the *GPCR* genes and the clinical associations with altered drug response were gleaned from the literature. To estimate the de novo missense mutation rate within these *GPCR* genes, authors also identified de novo mutations from >1700 control trios (having no reported pathological conditions), which were compiled from 10 different studies registered in the “de novo-database,” a collection of germline de novo variants (<http://denovo-db.gs.washington.edu/denovo-db/>).

In proof-of-principle, the authors then experimentally showed that certain variants of the mu-opioid and cholecystokinin receptors resulted in altered drug responses and/or idiosyncratic dose-independent ADRs. These data—on just two of the 108 *GPCR* genes—underscore the need to characterize SNPs among all 108 of the *GPCR* genes. The authors suggest that the ultimate results of this kind of in silico study “might enhance prescription precision, improve patients’ quality-of-life, and remove some of the economic and societal burden caused by variability in drug response.” It is anticipated that such “dry-lab” data-mining studies, such as this landmark publication [72], are likely to become a new major approach to PGx research in the near future.

16.5.4 Genome-Wide Association Studies of Complex PGx Traits

PGx GWAS have been attempted in studies of the treatment of common complex diseases such as type II diabetes, dyslipidemia, hypertension, and psychiatric disorders. Although some statistically significant findings have been reported, none of the identified PGx associations has clinical predictive value.

16.5.4.1 GWAS of Type II Diabetes Treatment

In GWAS examining the *glycemic-index response* to metformin in >1000 individuals with type II diabetes, for example, the GoDarts Group [60] identified several variants in the ataxia-telangiectasia-mutated gene (*ATM*) that are associated with treatment success; in an independent cohort with almost 1800 samples, they were able to replicate this association. For each minor allele reflecting the most strongly associated variant, the combined effect was reported to be 1.35 times higher than that of the major allele—for reaching the treatment goal of lowering glucose-bound hemoglobin (HbA1c) levels by at least 7%.

It was exciting to see the correlation of metformin efficacy with this *ATM* variant replicated in multiple additional cohorts [224]. However, the variant of the *ATM* gene appears to have an insufficient impact on the predictive value, with regard to metformin action on the glycemic index control [48].

Nevertheless—for reasons not clear—several additional research findings do indicate a substantial role of *ATM* variants in glucose homeostasis. For example, ataxia telangiectasia patients sometimes manifest a severe form of diabetes. Moreover, *Atm*(−/−) knockout mice exhibit insulin resistance and abnormal adipose distribution [206]. In a GWAS comprising ~13,000 participants [252], the Metformin Genetics (MetGen) Consortium identified an additional association between a variant in a facilitated glucose-transporter gene (*SLC2A2*) and the response by type II diabetes patients to metformin treatment (Table 16.6); the *SLC2A2* transporter is considered a credible candidate.

16.5.4.2 GWAS of Cardiovascular Disease Treatment

The cardiovascular response to efficacy of statins has been extensively tested by multiple GWAS [15,37,75,172] (reviewed in [115]). The GWAS sample sizes ranged between ~2000 and >40,000 subjects. These studies have identified common variants in the *LPA*, *APOE*, *SLCO1B1*, *SORT1*, and *ABCG2* genes most robustly associated with a favorable statin-induced response to low-density-lipoprotein-cholesterol (LDL-C) levels. However, none of the variants was statistically significantly correlated with modifications in risk reduction for cardiovascular events. These data suggest “limited clinical utility.”

16.5.5 Genome-Wide Association Studies of Unsuccessful PGx Examples

16.5.5.1 GWAS Dilemma of Hypertension Treatment

Probably the most prevalent modifiable risk factor for worldwide disease burden is elevated blood pressure. Accordingly, many PGx GWAS have attempted to identify genetic variants that might cause the largely inconsistent responses to various antihypertensive medications. Despite extreme efforts, PGx GWAS attempting to find links between genotypes and efficacy of antihypertensive agents have not generated robust independently replicable findings [125,141]; such failures likely reflect the extreme heterogeneity and complexity of hypertensive disease.

Elevated blood pressure has multiple etiologies and involves numerous physiological changes throughout one's lifetime. Such confounders, of course, create limitations in any genotype–phenotype association study. Additional caveats include the multiple classes of drugs that are used to treat hypertension (diuretics, ACE inhibitors, angiotensin II receptor blockers, calcium channel blockers, beta-blockers)—alone, and in combinations. These limitations further complicate “clean” PGx studies—not to mention individual variability caused by fluctuations of blood pressure (e.g., changes in diet, stress, exercise, time-of-day, combinations of prescribed drugs, usage of over-the-counter drugs, amount of cigarette-smoking, alcohol intake, etc.).

Several PGx GWAS have reported variants in plausible genes associated with thiazide diuretics [16,220,222], angiotensin II receptor blockers [51,221], and beta-blockers [64,65]. More recently, the PEAR (Pharmacogenomic Evaluation of Antihypertensive Responses) group reported genetic variants associated with chlorthalidone-induced increases in blood glucose levels [194] and with heart rate response to beta-blockers [189]. However, most of these GWAS involve relatively small sample sizes, and thus are underpowered to establish robust associations having genome-wide significance; furthermore, the findings will need to be replicated in future independent studies.

Huge international consortia, aimed at increasing opportunities for discovery and replication by assembling large samples, have been established [19]. Even though large consortium studies might be able to identify genotype–phenotype associations for hypertensive drug efficacy in the future, however, it is difficult to envision that—even in combination—any of these newly identified small-effect variants will ever achieve important clinical predictive power. (However, new drug targets might be identified.)

16.5.5.2 GWAS Dilemma of Psychotropic Drugs

Just as is true for hypertensive drug treatment, GWAS of antidepressant treatment response have shown little promise of success [8,43,53,83,223]. Besides patient compliance, one specific challenge in PGx studies of psychotropic medications is the evaluation of disease conditions and quantitative measurements of treatment responses—efficacies of antidepressant drugs are all based on multidimensional and (the physician's and the patient's) subjective psychiatric diagnostic criteria.

For many years, these “soft” measurements have been criticized for their lack of reliability. One must trust physicians’ consensus of the diagnosis and treatment response, determined by different clinicians, as well as by how honestly the patient answers the questionnaire (reviewed in [11]). This problem—i.e., dealing with an “equivocal phenotype” instead of an “unequivocal phenotype”—has previously been emphasized [158].

An excellent example of questionable diagnostic reliability exists with the major depressive disorder (MDD) and how to quantify treatment efficacy [52]. Clinically based psychiatric criteria clearly represent a gradient, rather than any biologically homogeneous condition; one solution offered to help in classifying an unequivocal trait is the “extreme discordant phenotype” (EDP) method of analysis [149,249].

Consequently, difficulties in phenotypic definitions of psychiatric conditions are highly heterogeneous, and measurements of treatment responses are decidedly inaccurate. A recent large GWAS meta-analysis [241] identified 44 loci associated with MDD; the revealed *genetic architecture* suggests MDD is not a distinct pathological condition, but rather an anthropocentric clinical construct associated with a wide range of diverse outcomes—the end result of which is a complex process of intertwined genetic and environmental effects.

On the other hand, a GWAS evaluating *plasma concentrations* of the parent drugs citalopram and escitalopram, and their *metabolites*, was successful; in 435 MDD patients, investigators succeeded in identifying highly significant associations with *CYP2C19* and the *CYP2D6* gene variants [88]. This study underscores an important distinction between *subjective* psychiatric clinical phenotypes and *quantitative measurements* of blood or urine drug- or metabolite-level phenotypes; as mentioned earlier, this concept of an equivocal phenotype versus unequivocal phenotype has previously been emphasized [158]. One therefore can recognize the difficulties in attempting to use PGx GWAS—as well as pharmacometabolomic association studies, such as those described for MDD [88]—to predict drug efficacy, or risk of ADRs, or genetic risk of many complex clinical responses.

16.6 RESPONSE TO DRUGS OTHER THAN GENOTYPE OF THE PATIENT

Early in this chapter, we described four additional factors other than the *genetic architecture* of each patient that can influence drug response. These include *epigenetic*

effects, *endogenous influences*, *environmental factors*, and *interindividual differences in the microbiome*. These are briefly covered below.

16.6.1 Epigenetics

Epigenetic effects on drug response are likely to be important, but, currently, data showing the extent to which this happens are very much limited. Examples of “epigenetic effects” include: DNA-methylation patterns, RNA-interference regulatory processes, histone modifications, and chromatin remodeling. Ready-to-use kits are now available—and becoming increasingly less expensive—to study genome-wide DNA methylation, as well as microRNA assays. Modifications of histones and remodeling of chromatin, on the other hand, are two fields of active research that continue to be advanced at the present time; currently, there are not yet any simple screening procedures. Epigenetics might provide a new framework in our search for etiological factors contributing to complex diseases, drug efficacy, and ADRs.

Genetic differences in epigenetic effects are expected to be found. For example, a comparison of DNA-methylation in buccal mucosal cells [96] found highly significantly ($P = 1.2 \times 10^{-294}$) less epigenetic variability between monozygotic twin pairs than between dizygotic twin pairs.

Some epigenetic variants can be inherited by offspring (and the offspring’s offspring)—which apparently represents a *transgenerational* mechanism [32] for “biological heredity apparently not based on DNA sequence.” For example, during the Dutch Famine of 1945, risk of neurodevelopmental disorders was increased in grandchildren whose grandparents were exposed prenatally [205]. Famine in a small Swedish village—at the time that males were entering puberty—appears to send an unknown message to their grandsons, curiously resulting in less risk of type II diabetes; famine in the same population, at the time that females’ oocytes are forming during the third trimester in utero, increases risk of obesity and type II diabetes in those babies’ granddaughters [168].

Grandmothers who smoke cigarettes during pregnancy are associated with a fourfold increased risk of asthma in their granddaughters [118]. Germ cells can also carry epigenetic effects from the grandmother’s diet [18]. Individuals whose grandparents suffered malnutrition in utero during the 1959–61 Chinese Famine were reported to have increased risk of schizophrenia [244]; however, another analysis of these same data suggests

this result might have reflected an epidemiological artifact called *population stratification* [198]. Chronic stress in pregnant mothers [92], as well as dietary effects in fathers' sperm [159], were reported to be sufficient to induce changes in the *epigenetic landscape* of the developing embryo and fetus in utero.

Development of the *chromosome-conformation capture* method [36] has demonstrated that chromatin is partitioned into active and inactive compartments (reviewed in [200]). This type of chromatin remodeling would fall under the category of “epigenetic effects on phenotype.” Understanding the impact of “three-dimensional genomics” on multifactorial traits—such as complex diseases (e.g., cancer, innate immunity), quantitative traits (e.g., height, body mass index), and drug efficacy, ADRs, and toxicity—is predicted to become relevant to PGx in the future.

Epigenetic changes are well known to be involved in various complex diseases, including cancer and asthma [45]. However, the extent to which epigenetic factors—both developmental and transgenerational—might affect PGx phenotypes (efficacy, therapeutic failure, ADRs, and/or toxicity) is unknown.

16.6.2 Endogenous Influences

Any drug's PK and PD phenotypes can be expected to change as a function of *age*. Changes in the neonate and infant during the first few months of life occur much more rapidly than after 1 year of age. Age-related effects are then less striking throughout childhood, adolescence, and most of adulthood. In *geriatric populations*, physicians need to pay more attention to increased inter-individual variability; this is in part due to decreased *renal excretion* (as much as 50%) in about two-thirds of elderly patients over the age of 75, but also due to confounding factors such as *coronary heart disease* and *hypertension* (reviewed in [139,190]).

There are *hormonal differences* in PK and PD traits in men, women, and pregnant women [197]. *Ethnic differences* in drug response have been repeatedly mentioned throughout this chapter, as well as especially addressed in Section 16.4.10. Lastly, vigorous *exercise* diminishes blood flow in the liver—which will mostly affect the degradation of drugs having *hepatic high-clearance rates* (e.g., lidocaine, nitrates); on the other hand, metabolic breakdown of hepatic low-clearance drugs (e.g., amobarbital, antipyrine, and diazepam) is not significantly affected by exercise [246]. Strenuous exercise can also lower renal plasma flow, urinary excretion rates, and

urine pH—which would explain why serum levels of drugs that are eliminated through the kidneys increase during physical stress.

16.6.3 Environmental Factors

Environmental agents, when present in substantial amounts, can affect drug response (*efficacy*, *therapeutic failure*, *adverse effect*, and *toxic effect*); clearly, these effects might influence, or be “confounders” in, genetic association studies of PGx traits. Examples include over-the-counter medications, dietary factors, drug-drug interactions, cigarette smoking, heavy alcohol usage, occupation-related chemicals in the workplace, and living in regions where a hazardous substance, or mixture of toxic substances, is/are present in considerable amounts. For over-the-counter medications and dietary factors, St. John's wort and other botanical preparations [199], as well as grapefruit juice [113], are among the most-often mentioned examples. Miniscule amounts of environmental *bis*-phenol A, manganese, and polyfluoroalkyl pollution have been extensively studied, but any significant effects on PGx variability would seem highly unlikely.

In the case of drug-drug interactions [100], obviously if a drug—not being analyzed in the GWAS cohort—is being taken by the patient, and is a substrate for a (PK or PD) gene target that competes with the drug being studied, this would be a caveat. This other drug will interfere with the “purity” of the GWAS. In other words, the phenotype will be less certain, which, in turn, could impact the power of the overall genotype-phenotype association study.

Cigarette smoking or hazardous occupational chemicals might sometimes affect drug response. For cigarette smoking, an accurate “smoking history” (i.e., cigarette-pack-years) is usually a reasonably quantifiable effect on drug response that can be studied. For occupation-related chemicals in the workplace, precise quantification (i.e., the amount of exposure to any individual work as a function of time) is more problematic than cigarette-smoking history. For patients living in regions where considerable amounts of a hazardous substance are present (in the air, ground, or water), any influence of drug response becomes even less quantifiable, i.e., more uncertain.

In PGx studies, the dose of a drug, and how long it has been taken by the patient (presuming that the patient has been *compliant*), is more quantitative than exposure to any of these above-mentioned groups of

environmental agents. Nonetheless, it is worth mentioning that there are GWAS of environmental stimuli that have identified specific genetic loci. For example, a meta-analysis representing five population-based GWAS of ~47,000 “habitual coffee drinkers” of European descent—adjusted for age, sex, and smoking history—found two statistically significant genes [21]: *AHR* ($P=2.4\times10^{-19}$) and the *CYP1A2_1A1* locus ($P=5.2\times10^{-14}$). Both of these loci are credible biological candidates because CYP1A2 metabolizes caffeine, and AHR regulates cigarette-smoke-inducible CYP1A1 and CYP1A2 enzymatic activities that participate in the metabolism of chemicals in cigarette smoke. A second example is a GWAS of ~1300 arsenic-exposed Bangladeshi individuals [170], in which five highly significant SNPs in and near the arsenite methyltransferase gene (*AS3MT*) showed independent associations with arsenic-caused toxic effects.

16.6.4 Microbiome Differences

Benign bacteria (and a few viruses and fungi)—which normally inhabit our bodies synergistically—have been termed “the microbiome” [112]. Although the microbiome has largely been overlooked (except for attempts to suppress or eradicate microorganisms), these microorganisms constitute ~90% of the total number of cells associated with our bodies, i.e., human cells comprise merely the remaining 10% [185]. Because of recent advances in genomics technology, we have only begun to appreciate contributions of the microbiome to modifications of health and disease; this includes conditions once believed to be genetically encoded purely by the host’s chromosomes [68].

Although many drugs have GI side-effects and the gut microbiome itself is pivotal for human health [93], the role of the microbiome in these processes has rarely been considered. Recent studies have demonstrated that drugs designed to target human cells, rather than microbes—such as the antidiabetic metformin [50], proton-pump inhibitors [81,85], nonsteroidal antiinflammatory drugs [180], and atypical antipsychotics [49]—have been associated with alterations in composition of the microbiome. A larger cohort study [44] suggests that medications can also substantially alter the GI microbiome composition in more general ways.

Recently, in a screen of >1000 marketed drugs against 40 representative gut bacterial strains, 24% of drugs designed for targets in the human host, including

members of all therapeutic classes mentioned above, were found to inhibit in vitro the growth of at least one bacterial strain [129]. These recent advances indicate a growing trend of appreciation for future research on *drug–microbiome interactions*.

Low-molecular-weight chemicals, produced and enzymatically altered in the gut flora, are chemically similar in structure to drugs, as well as to ligands that activate the host’s endogenous receptors, and lipid mediators (LMs) that participate in second-messenger pathways mediated by the arachidonic acid, eicosapentaenoic acid, and docosahexaenoic acid cascades (reviewed in [150]). Thus, the degree to which various drugs and drug metabolites administered to the host might be affected by gut flora chemicals is largely unappreciated at the present time. It therefore is reasonable to expect that the microbiome will influence drug response (*efficacy, therapeutic failure, adverse effect, and toxic effect*); in turn, such effects are clearly likely to have an effect on GWAS of PGx.

A specialized PGx aspect of *neuropsychopharmacogenomics* involves the recent realization of the importance of the “brain–gut–microbiome” (reviewed in [123,136]). It is now clear that there is bidirectional communication between the GI tract and the central nervous system, and that the CNS provides communication pathways between intestinal microbiota and the patient’s neural circuitry. Variations in GI tract and CNS function via the microbiome are now proposed to include such ill-defined traits as “mood,” “behavior,” “suicidal tendencies,” “obsessive-compulsive disorder,” cognitive functions, appetite, and autism spectrum disorder. Production of bioactive compounds by microbiota, and their potential probiotic activities includes neuroactive molecules such as histamine, serotonin, catecholamines, and trace amines [123,136].

In the near future, we therefore expect specific studies on effects of the brain–gut–microbiome on drug-response phenotypes. Perhaps more importantly, the outcome of response to psychotropic drugs might be particularly affected. Hence, problems of GWAS involving psychotropic drugs (as detailed in Section 16.5.5.2) might be due, in part, to the impact of the brain–gut–microbiome. We predict this topic will become a major thrust of PGx research in the near future.

To summarize Section 16.6, these nongenetic factors—*epigenetic effects, endogenous influences, environmental factors, and interindividual differences in the microbiome*—contribute to the overall phenotype in PGx studies. The extent to which each of these contributes,

to each individual volunteer or patient in the cohort, remains to be determined but will become an important area of research in the future.

16.7 FDA RECOMMENDATIONS FOR PGx GENOTYPING

The US Food and Drug Administration (FDA) recommends that, before prescribing, physicians should genotype their patients for specific biomarkers. Currently, there is FDA-approved information on the labels of more than 260 drugs and various medications <https://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm> (last updated February 8, 2018). The FDA suggests that these genetic biomarkers might help physicians identify patients in whom commonly prescribed drugs might be “less efficacious, insufficiently metabolized, or more likely to be toxic.” In recent surveys, most physicians agree that genetic profiles of patients can affect drug therapy. However, only about one in 10 physicians feel they have been adequately educated about using such genetic biomarkers in patients; therefore, few physicians actually order these tests.

Furthermore, the usual range, on average (~1.8–2.0-fold), of increased risk associated with positive results of these tests—might not seem large enough to cause sufficient concern. There are a number of reasons why physicians avoid PGx genomics-testing usage and implementation, and these include: the need to demonstrate clinical utility unequivocally and more clearly; the continuous introduction of new FDA-approved drugs on the market almost monthly; the almost daily flood of bewildering, indigestible volumes of new genomics information to which physicians are exposed; and the fact that not many drugs exhibit a therapeutic index (Fig. 16.3) that is substantially higher than ~1.8–2-fold (e.g., compared to the safety of aspirin). Moreover, genomics-testing results are often irreproducible, or they vary between different ethnic populations; this often makes insurance reimbursements difficult to process. Hence, “individualized drug therapy” and “personalized medicine” based entirely on DNA-sequence testing still seem far from becoming a clinical reality.

16.8 CONCLUSIONS

1. A major portion of *personalized medicine* is *individualized drug therapy*; however, because we continue
2. to see how incredibly complex the human genome is—and how unique each individual is—confident prediction of most drug responses is not possible in the foreseeable future.
3. *Pharmacogenetics* was originally defined as “gene–drug interactions” and *pharmacogenomics* defined as “drug–genome interactions,” but now both fields have become combined into the latter term, *pharmacogenomics* (PGx).
4. Each individual’s drug response is *holistic*, encompassing five contributing factors: *genotype*, *epigenetic effects*, *endogenous influences*, *environmental factors*, and *microbiome differences*. Whereas genotype is virtually always constant, the remaining four factors are dynamic and nongenetic.
5. Drug responses can include: *efficacy*, *adverse reaction*, *therapeutic failure*, and *toxic effect*. Any drug response can be regarded as a *phenotypic trait*.
6. Adverse drug reactions (ADRs) in the US have been reported to rank as high as the fifth leading cause of death.
7. Drug metabolism and drug target responses exist in every cell type of the body, and are often strikingly different from one cell type to another.
8. Variability in interindividual drug response, while it can be seen as a gradient, can be classified as: *monogenic* (Mendelian) *traits*, typically influenced by one or a few rare coding variants; *predominantly oligogenic traits* that usually represent variability largely elicited by a small number of major pharmacogenes; and *complex PGx traits*—produced mostly by innumerable small-effect variants. This last category is by far the most common.
9. Since the mid-2000s, PGx genome-wide association studies (GWAS) involving large cohorts were found to be able to detect genes not only in the *monogenic* and *oligogenic* categories, but also some genetic variants in the *complex PGx* category. The larger the cohort, the more genetic variants are revealed; however, each of these variants usually contributes a small effect to the trait and has no clinical predictive value—even when combined.
10. Recruiting a large number of individuals to study a quantitative trait (e.g., height, blood pressure, or body mass index) or a complex disease (e.g., type II diabetes, schizophrenia, or ulcerative colitis) is less difficult than identifying large numbers of patients being treated with a specific drug.

10. One major difference between pharmacogenomics and complex-disease GWAS is that any patient who has not been challenged with a particular drug will never know his phenotype for that drug.
11. Statistical predictions are best applied at the population level. For the individual, we need to focus on gene–gene, gene–environment, and drug–drug interactions in that particular patient. Thus, overall effects from DNA variants that show statistical significance in a large cohort are generally too small to predict in the individual patient.
12. Epigenetics includes DNA-methylation effects, RNA-interference processes, histone modifications, and chromatin remodeling. There are assays now available for the first two categories; the latter two categories are more complicated and will require many more years of study before assays become readily available.
13. Each patient's microbiota, as well as the brain–gut–microbiome specifically, might contribute more to drug-response phenotypes than is currently appreciated.
14. As detailed in this chapter, given all the complexity of each individual's *genetic architecture*—as well as *epigenetic effects*, *endogenous influences*, *environmental factors*, and *differences in the microbiome*—it can be seen how difficult *individualized drug therapy* will be in the foreseeable future.

ACKNOWLEDGMENTS

We thank our colleagues for valuable discussions and a careful reading of this manuscript. This work was funded, in part, by NIH Grant P30 ES006096.

REFERENCES

- [1] Aboraya A, Rankin E, France C, El-Missiry A, John C. Reliability of psychiatric diagnosis revisited: the clinician's guide to improve reliability of psychiatric diagnosis. *Psychiatry* 2006;34:41–50.
- [2] Amacher DE, Chalasani N. Drug-induced hepatic steatosis. *Semin Liver Dis* 2014;34:205–14.
- [3] Aminkeng F, Bhavsar AP, Visscher H, Rassekh SR, Li Y, Lee JW, Brunham LR, Caron HN, van Dalen EC, Kremer LC, van der Pal HJ, Amstutz U, Rieder MJ, Bernstein D, Carleton BC, Hayden MR, Ross CJ. A coding variant in *RARG* confers susceptibility to anthracycline-induced cardiotoxicity in childhood cancer. *Nat Genet* 2015;47:1079–84.
- [4] Angelo M, Dring LG, Lancaster R, Latham A, Smith RL. Proceedings: a correlation between the response to debrisoquine and the amount of unchanged drug excreted in the urine. *Br J Pharmacol* 1975;55:264P.
- [5] Backman JD, O'Connell JR, Tanner K, Peer CJ, Figg WD, Spencer SD, Mitchell BD, Shuldiner AR, Yerges-Armstrong LM, Horenstein RB, Lewis JP. Genome-wide analysis of clopidogrel active metabolite levels identifies novel variants that influence antiplatelet response. *Pharmacogenet Genom* 2017;27:159–63.
- [6] Bergmeijer TO, Reny JL, Pakyz RE, Gong L, Lewis JP, Kim EY, Aradi D, Fernandez-Cadenas I, Horenstein RB, Lee MTM, Whaley RM, Montaner J, Gensini GF, Cleator JH, Chang K, Holmvang L, Hochholzer W, Roden DM, Winter S, Altman RB, Alexopoulos D, Kim HS, Dery JP, Gawaz M, Bliden K, Valgimigli M, Marcucci R, Campo G, Schaeffeler E, Dridi NP, Wen MS, Shin JG, Simon T, Fontana P, Giusti B, Geisler T, Kubo M, Trenk D, Siller-Matula JM, Ten Berg JM, Gurbel PA, Hulot JS, Mitchell BD, Schwab M, Ritchie MD, Klein TE, Shuldiner AR, Investigators I. Genome-wide and candidate gene approaches of clopidogrel efficacy using pharmacodynamic and clinical end points — rationale and design of the International Clopidogrel Pharmacogenomics Consortium (ICPC). *Am Heart J* 2018;198:152–9.
- [7] Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, Chatterjee N. Using principal components of genetic variation for robust and powerful detection of gene-gene interactions in case-control and case-only studies. *Am J Hum Genet* 2010;86:331–42.
- [8] Biernacka JM, Sangkuhl K, Jenkins G, Whaley RM, Barman P, Batzler A, Altman RB, Arolt V, Brockmoller J, Chen CH, Domschke K, Hall-Flavin DK, Hong CJ, Illi A, Ji Y, Kampman O, Kinoshita T, Leinonen E, Liou YJ, Mushiroda T, Nonen S, Skime MK, Wang L, Baune BT, Kato M, Liu YL, Praphanphoj V, Stingl JC, Tsai SJ, Kubo M, Klein TE, Weinshilboum R. The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response. *Transl Psychiatry* 2015;5:e553.
- [9] Blum M, Grant DM, McBride W, Heim M, Meyer UA. Human arylamine *N*-acetyltransferase genes: isolation, chromosomal localization, and functional expression. *DNA Cell Biol* 1990;9:193–203.
- [10] Brodie BB, Aronow L, Axelrod J. The fate of benzazoline (prisco-line) in dog and man and a method for its estimation in biological material. *J Pharmacol Exp Therapeut* 1952;106:200–7.
- [11] Brunton L, Chabner B, Knollman B. Goodman & Gilman's the pharmacological basis of therapeutics. McGraw-Hill Companies, Inc; 2011. [Printed in China].

- [12] Caldwell MD, Awad T, Johnson JA, Gage BF, Falkowski M, Gardina P, Hubbard J, Turpaz Y, Langa TY, Eby C, King CR, Brower A, Schmelzer JR, Glurich I, Vidaillet HJ, Yale SH, Qi ZK, Berg RL, Burmester JK. *CYP4F2* genetic variant alters required warfarin dose. *Blood* 2008;111:4106–12.
- [13] Caraco Y, Sheller J, Wood AJ. Pharmacogenetic determination of the effects of codeine and prediction of drug interactions. *J Pharmacol Exp Therapeut* 1996;278:1165–74.
- [14] Cha PC, Mushiroda T, Takahashi A, Kubo M, Minami S, Kamatani N, Nakamura Y. Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. *Hum Mol Genet* 2010;19:4735–44.
- [15] Chasman DI, Giulianini F, MacFadyen J, Barratt BJ, Nyberg F, Ridker PM. Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction. Justification for Use of statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) trial. *Circ Cardiovasc Genet* 2012;5:257–64.
- [16] Chittani M, Zaninello R, Lanzani C, Frau F, Ortu MF, Salvi E, Fresu G, Citterio L, Braga D, Piras DA, Carpinì SD, Velayutham D, Simonini M, Argiolas G, Pozzoli S, Troffa C, Glorioso V, Kontula KK, Hiltunen TP, Donner KM, Turner ST, Boerwinkle E, Chapman AB, Padmanabhan S, Dominiczak AF, Melander O, Johnson JA, Cooper-DeHoff RM, Gong Y, Rivera NV, Condorelli G, Trimarco B, Manunta P, Cusi D, Glorioso N, Barlassina C. *TET2* and *CSMD1* genes affect SBP response to hydrochlorothiazide in never-treated essential hypertensives. *J Hypertens* 2015;33:1301–9.
- [17] Conjeevaram HS, Fried MW, Jeffers LJ, Terrault NA, Wiley-Lucas TE, Afdhal N, Brown RS, Belle SH, Hoofnagle JH, Kleiner DE, Howell CD. Peginterferon and ribavirin treatment in African-American and Caucasian-American patients with hepatitis-C genotype 1. *Gastroenterology* 2006;131:470–7.
- [18] Cooney CA. Germ cells carry the epigenetic benefits of grandmother's diet. *Proc Natl Acad Sci U S A* 2006;103:17071–2.
- [19] Cooper-DeHoff RM, Johnson JA. Hypertension pharmacogenomics: in search of personalized treatment approaches. *Nat Rev Nephrol* 2016;12:110–22.
- [20] Cooper GM, Johnson JA, Langa TY, Feng H, Stanaway IB, Schwarz UI, Ritchie MD, Stein CM, Roden DM, Smith JD, Veenstra DL, Rettie AE, Rieder MJ. Genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 2008;112:1022–7.
- [21] Cornelis MC, Monda KL, Yu K, Paynter N, Azzato EM, Bennett SN, Berndt SI, Boerwinkle E, Chanock S, Chatterjee N, Couper D, Curhan G, Heiss G, Hu FB, Hunter DJ, Jacobs K, Jensen MK, Kraft P, Landi MT, Nettleton JA, Purdue MP, Rajaraman P, Rimm EB, Rose LM, Rothman N, Silverman D, Stolzenberg-Solomon R, Subar A, Yeager M, Chasman DI, van Dam RM, Caporaso NE. Genome-wide meta-analysis identifies regions on 7p21 (*AHR*) and 15q24 (*CYP1A2*) as determinants of habitual caffeine consumption. *PLoS Genet* 2011;7:e1002033.
- [22] Cotton RG. Heterogeneity of phenylketonuria at the clinical, protein and DNA levels. *J Inher Metab Dis* 1990;13:739–50.
- [23] Curran ME, Splawski I, Timothy KW, Vincent GM, Green ED, Keating MT. A molecular basis for cardiac arrhythmia: *HERG* mutations cause long-QT syndrome. *Cell* 1995;80:795–803.
- [24] D'Andrea G, D'Ambrosio RL, Di PP, Chetta M, Santacroce R, Brancaccio V, Grandone E, Margaglione M. Polymorphism in the *VKORC1* gene associated with an interindividual variability in the dose-anticoagulant effect of warfarin. *Blood* 2005;105:645–9.
- [25] Dai D, Tang J, Rose R, Hodgson E, Bienstock RJ, Mohrenweiser HW, Goldstein JA. Identification of variants of *CYP3A4* and characterization of their abilities to metabolize testosterone and chlorpyrifos. *J Pharmacol Exp Therapeut* 2001;299:825–31.
- [26] Dai D, Zeldin DC, Blaisdell JA, Chanas B, Coulter SJ, Ghanayem BI, Goldstein JA. Polymorphisms in human *CYP2C8* decrease metabolism of the anticancer drug paclitaxel and arachidonic acid. *Pharmacogenetics* 2001;11:597–607.
- [27] Daly AK. Pharmacogenomics of anticoagulants: steps toward personal dosage. *Genome Med* 2009;1:10.
- [28] Daly AK. Using genome-wide association studies to identify genes important in serious adverse drug reactions. *Annu Rev Pharmacol Toxicol* 2012;52:21–35.
- [29] Daly AK. Human leukocyte antigen (*HLA*) pharmacogenomic tests: potential and pitfalls. *Curr Drug Metabol* 2014;15:196–201.
- [30] Daly AK, Brockmoller J, Broly F, Eichelbaum M, Evans WE, Gonzalez FJ, Huang JD, Idle JR, Ingelman-Sundberg M, Ishizaki T, Jacqz-Aigrain E, Meyer UA, Nebert DW, Steen VM, Wolf CR, Zanger UM. Nomenclature for human *CYP2D6* alleles. *Pharmacogenetics* 1996;6:193–201.
- [31] Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, Graham J, Park BK, Dillon JF, Bernal W, Cordell HJ, Pirmohamed M, Aithal GP, Day CP. *HLA-B*5701* genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat Genet* 2009;41:816–9.

- [32] Daxinger L, Whitelaw E. Transgenerational epigenetic inheritance: more questions than answers. *Genome Res* 2010;20:1623–8.
- [33] de Morais SM, Schweikl H, Blaisdell J, Goldstein JA. Gene structure and upstream regulatory regions of human *CYP2C9* and *CYP2C18*. *Biochem Biophys Res Commun* 1993;194:194–201.
- [34] de Morais SM, Wilkinson GR, Blaisdell J, Nakamura K, Meyer UA, Goldstein JA. The major genetic defect responsible for the polymorphism of *S*-mephenytoin metabolism in humans. *J Biol Chem* 1994;269:15419–22.
- [35] Deenen MJ, Meulendijks D, Cats A, Sechterberger MK, Severens JL, Boot H, Smits PH, Rosing H, Mandigers CM, Soesan M, Beijnen JH, Schellens JH. Upfront genotyping of *DPYD*2A* to individualize fluoropyrimidine therapy: a safety and cost analysis. *J Clin Oncol* 2016;34:227–34.
- [36] Denker A, de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev* 2016;30:1357–82.
- [37] Deshmukh HA, Colhoun HM, Johnson T, McK-eigue PM, Betteridge DJ, Durrington PN, Fuller JH, Livingstone S, Charlton-Menys V, Neil A, Poulter N, Sever P, Shields DC, Stanton AV, Chatterjee A, Hyde C, Calle RA, Demicco DA, Trompet S, Postmus I, Ford I, Jukema JW, Caulfield M, Hitman GA. Genome-wide association study of genetic determinants of LDL-Chol response to atorvastatin therapy: importance of *LPA*. *J Lipid Res* 2012;53:1000–11.
- [38] Dover GA. Molecular drive in multigene families: how biological novelties arise, spread, and are assimilated. *Trends Genet* 1986;2:159–65.
- [39] Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *Lancet* 2000;356:1255–9.
- [40] Eichelbaum M. Ein neu entdeckte defect im Arznei-Mittelstoffwechsel des Menschen: die fehlende *N*-Oxydation des Spartein [thesis]. University of Bonn; 1975.
- [41] Eichelbaum M, Spannbrucker N, Steincke B, Dengler HJ. Defective *N*-oxidation of sparteine in man: a new pharmacogenetic defect. *Eur J Clin Pharmacol* 1979;16:183–7.
- [42] English BC, Baum CE, Adelberg DE, Sissung TM, Kluetz PG, Dahut WL, Price DK, Figg WD. A SNP in *CYP2C8* is not associated with development of bisphosphonate-related osteonecrosis of the jaw in men with castrate-resistant prostate cancer. *Therapeut Clin Risk Manag* 2010;6:579–83.
- [43] Fabbri C, Corponi F, Souery D, Kasper S, Montgomery S, Zohar J, Rujescu D, Mendlewicz J, Serretti A. The genetics of treatment-resistant depression: a critical review and future perspectives. *Int J Neuropsychopharmacol* 2018. [Epub ahead of print].
- [44] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D'Hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. Population-level analysis of gut microbiome variation. *Science* 2016;352:560–4.
- [45] Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;447:433–40.
- [46] Fernandez CA, Smith C, Yang W, Mullighan CG, Qu C, Larsen E, Bowman WP, Liu C, Ramsey LB, Chang T, Karol SE, Loh ML, Raetz EA, Winick NJ, Hunger SP, Carroll WL, Jeha S, Pui CH, Evans WE, Devidas M, Relling MV. Genome-wide analysis links *NFATC2* with asparaginase hypersensitivity. *Blood* 2015;126:69–75.
- [47] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinb* 1919;52:399–433.
- [48] Florez JC, Jablonski KA, Taylor A, Mather K, Horton E, White NH, Barrett-Connor E, Knowler WC, Shuldiner AR, Pollin TI, Diabetes Prevention Program Research G. The C allele of *ATM* rs11212617 does not associate with metformin response in the Diabetes Prevention Program. *Diabetes Care* 2012;35:1864–7.
- [49] Flowers SA, Evans SJ, Ward KM, McLinnis MG, El-lingrod VL. Interaction between atypical antipsychotics and the gut microbiome in a bipolar disease cohort. *Pharmacotherapy* 2017;37:261–7.
- [50] Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Pedersen HK, Arumugam M, Kristiansen K, Voigt AY, Vestergaard H, Hercog R, Costea PI, Kultima JR, Li J, Jorgensen T, Levenez F, Dore J, Meta HITc, Nielsen HB, Brunak S, Raes J, Hansen T, Wang J, Ehrlich SD, Bork P, Pedersen O. Disentangling type-2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 2015;528:262–6.
- [51] Frau F, Zaninello R, Salvi E, Ortu MF, Braga D, Velayutham D, Argiolas G, Fresu G, Troffa C, Bulla E, Bulla P, Pitzoi S, Piras DA, Glorioso V, Chittani M, Bernini G, Bordini M, Fallo F, Malatino L, Stancanelli B, Regolisti G, Ferri C, Desideri G, Scioli GA, Galletti F, Sciacqua A, Perticone F, Degli EE, Sturani A, Semplicini A, Veglio F, Mulatero P, Williams TA, Lanzani C, Hiltunen TP, Kontula K, Boerwinkle E, Turner ST, Manunta P, Barlassina C, Cusi D, Glorioso N. Genome-wide association study identifies *CAMKID* variants involved in blood pressure response to losartan: the SOPHIA study. *Pharmacogenomics* 2014;15:1643–52.

- [52] Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, Gabbard GO, Gau SS, Javitt DC, Oquendo MA, Shrout PE, Vieta E, Yager J. The initial field trials of DSM-5: new blooms and old thorns. *Am J Psychiatr* 2013;170:1–5.
- [53] Garriock HA, Kraft JB, Shyn SI, Peters EJ, Yokoyama JS, Jenkins GD, Reinalda MS, Slager SL, McGrath PJ, Hamilton SP. A genome-wide association study of citalopram response in major depressive disorder. *Biol Psychiatr* 2010;67:133–8.
- [54] Gasche Y, Daali Y, Fathi M, Chiappe A, Cottini S, Dayer P, Desmeules J. Codeine intoxication associated with ultrarapid CYP2D6 metabolism. *N Engl J Med* 2004;351:2827–31.
- [55] Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M, McHutchison JG, Goldstein DB. Genetic variation in *IL28B* predicts hepatitis C treatment-induced viral clearance. *Nature* 2009;461:399–401.
- [56] Geldmacher-v Mallinckrodt M, Hommel G, Dumbach J. On the genetics of the human serum paraoxonase (EC 3.1.1.2). *Hum Genet* 1979;50:313–26.
- [57] Gibson G. Decanalization and the origin of complex disease. *Nat Rev Genet* 2009;10:134–40.
- [58] Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet* 2011;13:135–45.
- [59] Gibson G, Dworkin I. Uncovering cryptic genetic variation. *Nat Rev Genet* 2004;5:681–90.
- [60] GoDarts, Group UDPS, Wellcome Trust Case Control C, Zhou K, Bellenguez C, Spencer CC, Bennett AJ, Coleman RL, Tavendale R, Hawley SA, Donnelly LA, Schofield C, Groves CJ, Burch L, Carr F, Strange A, Freeman C, Blackwell JM, Bramon E, Brown MA, Casas JP, Corvin A, Craddock N, Deloukas P, Dronov S, Duncanson A, Edkins S, Gray E, Hunt S, Jankowski J, Langford C, Markus HS, Mathew CG, Plomin R, Rautanen A, Sawcer SJ, Samani NJ, Trembath R, Viswanathan AC, Wood NW, investigators M, Harries LW, Hattersley AT, Doney AS, Colhoun H, Morris AD, Sutherland C, Hardie DG, Peltonen L, McCarthy MI, Holman RR, Palmer CN, Donnelly P, Pearson ER. Common variants near *ATM* are associated with glycemic response to metformin in type-2 diabetes. *Nat Genet* 2011;43:117–20.
- [61] Goedde HW, Agarwal DP. Aldehyde oxidation: ethnic variations in metabolism and response. *Prog Clin Biol Res* 1986;214:113–38.
- [62] Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009;360:1696–8.
- [63] Goldstein JA, de Morais SM. Biochemistry and molecular biology of the human *CYP2C* subfamily. *Pharmacogenetics* 1994;4:285–99.
- [64] Gong Y, McDonough CW, Beitelshes AL, El RN, Hiltunen TP, O'Connell JR, Padmanabhan S, Langaee TY, Hall K, Schmidt SO, Curry Jr RW, Gums JG, Donner KM, Kontula KK, Bailey KR, Boerwinkle E, Takahashi A, Tanaka T, Kubo M, Chapman AB, Turner ST, Pepine CJ, Cooper-DeHoff RM, Johnson JA. *PTPRD* gene associated with blood pressure response to atenolol and resistant hypertension. *J Hypertens* 2015;33:2278–85.
- [65] Gong Y, Wang Z, Beitelshes AL, McDonough CW, Langaee TY, Hall K, Schmidt SO, Curry Jr RW, Gums JG, Bailey KR, Boerwinkle E, Chapman AB, Turner ST, Cooper-DeHoff RM, Johnson JA. Pharmacogenomic genome-wide meta-analysis of blood pressure response to β -blockers in hypertensive African Americans. *Hypertension* 2016;67:556–63.
- [66] Gonzalez FJ, Nebert DW. Evolution of the P450 gene superfamily: animal-plant 'warfare', molecular drive, and human genetic differences in drug oxidation. *Trends Genet* 1990;6:182–6.
- [67] Gonzalez FJ, Skoda RC, Kimura S, Umeno M, Zanger UM, Nebert DW, Gelboin HV, Hardwick JP, Meyer UA. Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* 1988;331:442–6.
- [68] Grice EA, Segre JA. The human microbiome: our second genome. *Annu Rev Genom Hum Genet* 2012;13:151–70.
- [69] Hansen T, Echwald SM, Hansen L, Moller AM, Almind K, Clausen JO, Urhammer SA, Inoue H, Ferrer J, Bryan J, Aguilar-Bryan L, Permutt MA, Pedersen O. Decreased tolbutamide-stimulated insulin secretion in healthy subjects with sequence variants in the high-affinity sulfonylurea receptor gene. *Diabetes* 1998;47:598–605.
- [70] Hansen TF. The evolution of genetic architecture. *Annu Rev Ecol Evol Systemat* 2006;37:123–57.
- [71] Hassett C, Aicher L, Sidhu JS, Omiecinski CJ. Human microsomal epoxide hydrolase: genetic polymorphism and functional expression in vitro of amino acid variants. *Hum Mol Genet* 1994;3:421–8.
- [72] Hauser AS, Chavali S, Masuho I, Jahn LJ, Martemyanov KA, Gloriam DE, Babu MM. Pharmacogenomics of GPCR drug targets. *Cell* 2018;172:41–54. e19.
- [73] Hawcutt DB, Francis B, Carr DF, Jorgensen AL, Yin P, Wallin N, O'Hara N, Zhang EJ, Bloch KM, Ganguli A, Thompson B, McEvoy L, Peak M, Crawford AA, Walker BR, Blair JC, Couriel J, Smyth RL, Pirmohamed M. Susceptibility to corticosteroid-induced adrenal suppression: a genome-wide association study. *Lancet Respir Med* 2018;6:442–50.
- [74] Hernandez D, Addou S, Lee D, Orengo C, Shephard EA, Phillips IR. Trimethylaminuria and a human *FMO3* mutation database. *Hum Mutat* 2003;22:209–13.

- [75] Hopewell JC, Parish S, Offer A, Link E, Clarke R, Lathrop M, Armitage J, Collins R. Impact of common genetic variation in response to simvastatin therapy among 18,705 participants in the Heart Protection Study. *Eur Heart J* 2013;34:982–92.
- [76] Hu Y, Oscarson M, Johansson I, Yue QY, Dahl ML, Tabone M, Arinco S, Albano E, Ingelman-Sundberg M. Genetic polymorphism of human *CYP2E1*: characterization of two variant alleles. *Mol Pharmacol* 1997;51:370–6.
- [77] Humbert JA, Hammond KB, Hathaway WE. Trimethylaminuria: the fish-odour syndrome. *Lancet* 1970;2:770–1.
- [78] Humbert R, Adler DA, Disteché CM, Hassett C, Omiecinski CJ, Furlong CE. The molecular basis of the human serum paraoxonase activity polymorphism. *Nat Genet* 1993;3:73–6.
- [79] Iadonato SP, Katze MG. Genomics: hepatitis C virus gets personal. *Nature* 2009;461:357–8.
- [80] Idle JR, Smith RL. Polymorphisms of oxidation at carbon centers of drugs and their clinical significance. *Drug Metabol Rev* 1979;9:301–17.
- [81] Imhann F, Bonder MJ, Vich Vila A, Fu J, Mujagic Z, Vork L, Tigchelaar EF, Jankipersadsing SA, Cenit MC, Harmsen HJ, Dijkstra G, Franke L, Xavier RJ, Jonkers D, Wijmenga C, Weersma RK, Zhernakova A. Proton-pump inhibitors affect the gut microbiome. *Gut* 2016;65:740–8.
- [82] Ingelman-Sundberg M, Oscarson M, McLellan RA. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *Trends Pharmacol Sci* 1999;20:342–9.
- [83] Ising M, Lucae S, Binder EB, Bettecken T, Uhr M, Ripke S, Kohli MA, Hennings JM, Horstmann S, Kloiber S, Menke A, Bondy B, Rupprecht R, Domschke K, Baune BT, Arolt V, Rush AJ, Holsboer F, Muller-Myhsok B. A genomewide association study points to multiple loci that predict anti-depressant drug treatment outcome in depression. *Arch Gen Psychiatr* 2009;66:966–75.
- [84] Jackson CM, Suttie JW. Recent developments in understanding the mechanism of vitamin K and vitamin K-antagonist drug action and consequences of vitamin K action in blood coagulation. *Prog Hematol* 1977;10:333–59.
- [85] Jackson MA, Goodrich JK, Maxan ME, Freedberg DE, Abrams JA, Poole AC, Sutter JL, Welter D, Ley RE, Bell JT, Spector TD, Steves CJ. Proton pump inhibitors alter the composition of the gut microbiota. *Gut* 2016;65:749–56.
- [86] Jarvela I, Torniaainen S, Kolho KL. Molecular genetics of human lactase deficiencies. *Ann Med* 2009;41:568–75.
- [87] Jensen DM, Pol S. *IL28B* genetic polymorphism testing in the era of direct acting antivirals therapy for chronic hepatitis C: ten years too late? *Liver Int* 2012;32(Suppl. 1):74–8.
- [88] Ji Y, Schaid DJ, Desta Z, Kubo M, Batzler AJ, Snyder K, Mushiroda T, Kamatani N, Ogburn E, Hall-Flavin D, Flockhart D, Nakamura Y, Mrazek DA, Weinshilboum RM. Citalopram and escitalopram plasma drug and metabolite concentrations: genome-wide associations. *Br J Clin Pharmacol* 2014;78:373–83.
- [89] Johansson I, Lundqvist E, Bertilsson L, Dahl ML, Sjoqvist F, Ingelman-Sundberg M. Inherited amplification of an active gene in the cytochrome P450 *CYP2D* locus as a cause of ultrarapid metabolism of debrisoquine. *Proc Natl Acad Sci U S A* 1993;90:11825–9.
- [90] Johnson JA, Gong L, Whirl-Carrillo M, Gage BF, Scott SA, Stein CM, Anderson JL, Kimmel SE, Lee MT, Pirmohamed M, Wadelius M, Klein TE, Altman RB, Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clin Pharmacol Ther* 2011;90:625–9.
- [91] Johnson JA, Roden DM, Lesko LJ, Ashley E, Klein TE, Shuldiner AR. Clopidogrel: a case for indication-specific pharmacogenetics. *Clin Pharmacol Ther* 2012;91:774–6.
- [92] Johnstone SE, Baylin SB. Stress and the epigenetic landscape: a link to the pathobiology of human diseases? *Nat Rev Genet* 2010;11:806–12.
- [93] Kahrstrom CT, Pariente N, Weiss U. Intestinal microbiota in health and disease. *Nature* 2016;535:47.
- [94] Kalow W, Bertilsson L. Interethnic factors affecting drug response. *Adv Drug Res* 1994;25:1–53.
- [95] Kalow W, Genest K. A method for the detection of atypical forms of human serum cholinesterase: determination of dibucaine numbers. *Can J Biochem Physiol* 1957;35:339–46.
- [96] Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A. DNA-methylation profiles in monozygotic and dizygotic twins. *Nat Genet* 2009;41:240–5.
- [97] Karnes JH, Cronin RM, Rollin J, Teumer A, Pouplard C, Shaffer CM, Blanquicett C, Bowton EA, Cowan JD, Mosley JD, Van Driest SL, Weeke PE, Wells QS, Bakchoul T, Denny JC, Greinacher A, Gruel Y, Roden DM. A genome-wide association study of heparin-induced thrombocytopenia using an electronic medical record. *Thromb Haemostasis* 2015;113:772–81.

- [98] Katoh T. The frequency of glutathione-S-transferase M1 (*GSTM1*) gene deletion in patients with lung and oral cancer. *Sangyo Igaku* 1994;36:435–9.
- [99] Katz J, Gong Y, Salmasinia D, Hou W, Burkley B, Ferreira P, Casanova O, Langaee TY, Moreb JS. Genetic polymorphisms and other risk factors associated with bisphosphonate-induced osteonecrosis of the jaw. *Int J Oral Maxillofac Surg* 2011;40:605–11.
- [100] Kim BY, Sharafoddini A, Tran N, Wen EY, Lee J. Consumer mobile apps for potential drug-drug interaction check: systematic review and content analysis using the Mobile App Rating Scale (MARS). *JMIR Mhealth Uhealth* 2018;6:e74.
- [101] Kim UK, Jorgenson E, Coon H, Leppert M, Risch N, Drayna D. Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. *Science* 2003;299:1221–5.
- [102] Kioka N, Tsubota J, Kakehi Y, Komano T, Gottesman MM, Pastan I, Ueda K. P-glycoprotein gene (*MDR1*) cDNA from human adrenal: normal P-glycoprotein carries Gly185 with an altered pattern of multidrug resistance. *Biochem Biophys Res Commun* 1989;162:224–31.
- [103] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005;308:385–9.
- [104] Klironomos FD, Berg J, Collins S. How epigenetic mutations can affect genetic evolution: model and mechanism. *Bioessays* 2013;35:571–8.
- [105] Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. *J Am Med Ass* 2006;296:212–5.
- [106] König J, Seithel A, Gradhand U, Fromm MF, ö. Pharmacogenomics of human (*SLCO* gene) OATP transporters. *Naunyn-Schmiedeberg's Arch Pharmacol* 2006;372:432–43.
- [107] Kupfer A, Preisig R. Pharmacogenetics of mephenytoin: a new drug hydroxylation polymorphism in man. *Eur J Clin Pharmacol* 1984;26:753–9.
- [108] Kweekel D, Guchelaar HJ, Gelderblom H. Clinical and pharmacogenetic factors associated with irinotecan toxicity. *Canc Treat Rev* 2008;34:656–69.
- [109] Lamba V, Lamba J, Yasuda K, Strom S, Davila J, Hancock ML, Fackenthal JD, Rogan PK, Ring B, Wrighton SA, Schuetz EG. Hepatic CYP2B6 expression: gender and ethnic differences and relationship to *CYP2B6* genotype and CAR (constitutive androstane receptor) expression. *J Pharmacol Exp Therapeut* 2003;307:906–22.
- [110] Lander ES. Initial impact of the sequencing of the human genome. *Nature* 2011;470:187–97.
- [111] Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *J Am Med Ass* 1998;279:1200–5.
- [112] Lederberg J, McCray AT. 'Ome Sweet 'Omics — a genealogical treasury of words. *Scientist* 2001;15:8.
- [113] Lee JW, Morris JK, Wald NJ. Grapefruit juice and statins. *Am J Med* 2016;129:26–9.
- [114] Lee SJ, Usmani KA, Chanas B, Ghanayem B, Xi T, Hodgson E, Mohrenweiser HW, Goldstein JA. Genetic findings and functional studies of human *CYP3A5* single-nucleotide polymorphisms in different ethnic groups. *Pharmacogenetics* 2003;13:461–72.
- [115] Leusink M, Onland-Moret NC, de Bakker PI, de Boer A, Maitland-van der Zee AH. Seventeen years of statin pharmacogenetics: a systematic review. *Pharmacogenomics* 2016;17:163–80.
- [116] Levesque E, Beaulieu M, Hum DW, Belanger A. Characterization and substrate specificity of UGT2B4 (E458): a UDP-glucuronosyltransferase encoded by a polymorphic gene. *Pharmacogenetics* 1999;9:207–16.
- [117] Li Y, Zhang D, Jin W, Shao C, Yan P, Xu C, Sheng H, Liu Y, Yu J, Xie Y, Zhao Y, Lu D, Nebert DW, Harrison DC, Huang W, Jin L. Mitochondrial aldehyde dehydrogenase-2 (ALDH2) Glu504Lys polymorphism contributes to the variation in efficacy of sublingual nitroglycerin. *J Clin Invest* 2006;116:506–11.
- [118] Li YF, Langholz B, Salam MT, Gilliland FD. Maternal and grandmaternal smoking patterns are associated with early childhood asthma. *Chest* 2005;127:1232–41.
- [119] Liapakis A, Jesudian AB. Is there clinical utility to *IL28B* genotype testing in the treatment of chronic hepatitis C virus infection? *Pharmacogenomics* 2012;13:1317–9.
- [120] Ling G, Gu J, Genter MB, Zhuo X, Ding X. Regulation of cytochrome P450 gene expression in the olfactory mucosa. *Chem Biol Interact* 2004;147:247–58.
- [121] Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R. *SLCO1B1* variants and statin-induced myopathy — a genomewide study. *N Engl J Med* 2008;359:789–99.
- [122] Lotta LA, Sharp SJ, Burgess S, Perry JRB, Stewart ID, Willems SM, Luan J, Ardanaz E, Arriola L, Balkau B, Boeing H, Deloukas P, Forouhi NG, Franks PW, Grioni S, Kaaks R, Key TJ, Navarro C, Nilsson PM, Overvad K, Palli D, Panico S, Quiros JR, Riboli E, Rolandsson O, Sacerdote C, Salamanca EC, Slimani N, Spijkerman AM, Tjonneland A, Tumino R, van der AD, van der Schouw YT, McCarthy MI, Barroso I, O'Rahilly S, Savage DB, Sattar N, Langenberg C, Scott RA, Wareham NJ. Association between low-density

- lipoprotein cholesterol-lowering genetic variants and risk of type-2 diabetes: a meta-analysis. *J Am Med Ass* 2016;316:1383–91.
- [123] Luna RA, Savidge TC, Williams KC. The brain-gut-microbiome axis: what role does it play in autism spectrum disorder? *Curr Dev Disord Rep* 2016;3:75–81.
- [124] Lunenburg CA, Henricks LM, Guchelaar HJ, Swen JJ, Deenen MJ, Schellens JH, Gelderblom H. Prospective *DPYD* genotyping to reduce the risk of fluoropyrimidine-induced severe toxicity: ready for prime-time. *Eur J Canc* 2016;54:40–8.
- [125] Lupoli S, Salvi E, Barcella M, Barlassina C. Pharmacogenomics considerations in the control of hypertension. *Pharmacogenomics* 2015;16:1951–64.
- [126] Mackenzie IS, Maki-Petaja KM, McEniery CM, Bao YP, Wallace SM, Cheriyan J, Monteith S, Brown MJ, Wilkinson IB. Aldehyde dehydrogenase-2 plays a role in the bioactivation of nitroglycerin in humans. *Arterioscler Thromb Vasc Biol* 2005;25:1891–5.
- [127] MacLennan DH, Duff C, Zorzato F, Fujii J, Phillips M, Korneluk RG, Frodis W, Britt BA, Worton RG. Ryanodine receptor gene is a candidate for predisposition to malignant hyperthermia. *Nature* 1990;343:559–61.
- [128] Mahgoub A, Idle JR, Dring LG, Lancaster R, Smith RL. Polymorphic hydroxylation of debrisoquine in man. *Lancet* 1977;2:584–6.
- [129] Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, Brochado AR, Fernandez KC, Dose H, Mori H, Patil KR, Bork P, Typas A. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* 2018;555:623–8.
- [130] Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, Sayer D, Castley A, Mamotte C, Maxwell D, James I, Christiansen FT. Association between presence of *HLA-B*5701*, *HLA-DR7*, and *HLA-DQ3* and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *N Engl J Med* 2008;358:568–79.
- [131] Mallal S, Phillips E, Carosi G, Molina J-M, Workman C., Tomažič J, Jägel-Guedes E, Rugina S, Kozyrev O, Cid JF, Hay P, Nolan D, Hughes S, Hughes A, Ryan S, Fitch N, Thorborn D, Benbow A, Team P-S. *HLA-B*5701* screening for hypersensitivity to abacavir. *Virology* 2008;5:88.
- [132] Manolio TA. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* 2013;14:549–58.
- [133] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarrroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53.
- [134] Marks PA, Gross RT. Erythrocyte glucose-6-phosphate dehydrogenase deficiency: evidence of differences between Negroes and Caucasians with respect to this genetically-determined trait. *J Clin Invest* 1959;38:2253–62.
- [135] Marvit J, DiLella AG, Brayton K, Ledley FD, Robson KJ, Woo SL. GT to AT transition at a splice-donor site causes skipping of the preceding exon in phenylketonuria. *Nucleic Acids Res* 1987;15:5613–28.
- [136] Mazzoli R, Pessione E. The neuro-endocrinological role of microbial glutamate and GABA signaling. *Front Microbiol* 2016;7:1934.
- [137] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty, and challenges. *Nat Rev Genet* 2008;9:356–69.
- [138] McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperavičiute D, Carrington M, Sills GJ, Marson T, Jia X, de Bakker PI, Chinthapalli K, Molokhia M, Johnson MR, O'Connor GD, Chaila E, Alhusaini S, Shianna KV, Radtke RA, Heinzen EL, Walley N, Pandolfo M, Pichler W, Park BK, Depondt C, Sisodiya SM, Goldstein DB, Deloukas P, Delanty N, Cavalleri GL, Pirmohamed M. *HLA-A*3101* and carbamazepine-induced hypersensitivity reactions in Europeans. *N Engl J Med* 2011;364:1134–43.
- [139] McLachlan AJ, Pont LG. Drug metabolism in older people — a key consideration in achieving optimal outcomes with medicines. *J Gerontol A Biol Sci Med Sci* 2012;67:175–80.
- [140] Meinsma R, Fernandez-Salguero P, Van Kuilenburg AB, Van Gennip AH, Gonzalez FJ. Human polymorphism in drug metabolism: mutation in the dihydropyrimidine dehydrogenase gene results in exon skipping and thymine uraciluria. *DNA Cell Biol* 1995;14:1–6.
- [141] Menni C. Blood pressure pharmacogenomics: gazing into a misty crystal ball. *J Hypertens* 2015;33:1142–3.
- [142] Meyer UA. Pharmacogenetics: the slow, the rapid, and the ultrarapid. *Proc Natl Acad Sci U S A* 1994;91:1983–4.
- [143] Mizutani T. PM frequencies of major CYPs in Asians and Caucasians. *Drug Metab Rev* 2003;35:99–106.
- [144] Monks TJ, Anders MW, Dekant W, Stevens JL, Lau SS, van Bladeren PJ. Glutathione conjugate mediated toxicities. *Toxicol Appl Pharmacol* 1990;106:1–19.
- [145] Motsinger-Reif AA, Jorgenson E, Relling MV, Krotetz DL, Weinshilboum R, Cox NJ, Roden DM. Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenet Genom* 2013;23:383–94.
- [146] Motulsky AG. Drug reactions, enzymes, and biochemical genetics. *J Am Med Ass* 1957;165:835–7.

- [147] Nebert DW. Proposed role of drug-metabolizing enzymes: regulation of steady state levels of the ligands that effect growth, homeostasis, differentiation, and neuroendocrine functions. *Mol Endocrinol* 1991;5:1203–14.
- [148] Nebert DW. Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist? *Clin Genet* 1999;56:247–58.
- [149] Nebert DW. Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics. *Eur J Pharmacol* 2000;410:107–20.
- [150] Nebert DW. Aryl hydrocarbon receptor (AHR): “pioneer member” of the basic-helix/loop/helix per-Arnt-sim (bHLH/PAS) family of “sensors” of foreign and endogenous signals. *Prog Lipid Res* 2017;67:38–57.
- [151] Nebert DW, Carvan 3rd MJ. Ecogenetics: from ecology to health. *Toxicol Ind Health* 1997;13:163–92.
- [152] Nebert DW, Dalton TP. The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. *Nat Rev Canc* 2006;6:947–60.
- [153] Nebert DW, Dieter MZ. The evolution of drug metabolism. *Pharmacology* 2000;61:124–35.
- [154] Nebert DW, Karp CL. Endogenous functions of aryl hydrocarbon receptor (AHR): intersection of cytochrome P450 1 (CYP1)-metabolized eicosanoids and AHR biology. *J Biol Chem* 2008;283:36061–5.
- [155] Nebert DW, Vasiliou V. Analysis of the glutathione S-transferase (*GST*) gene family. *Hum Genom* 2004;1:460–4.
- [156] Nebert DW, Vesell ES. Chapter 19-“Pharmacogenetics and pharmacogenomics”. In: Emery and Rimoin’s principles and practice of medical genetics. Oxford: Academic Press; 2013. p. 1–27.
- [157] Nebert DW, Wikvall K, Miller WL. Human cytochromes P450 in health and disease. *Philos Trans R Soc Lond B Biol Sci* 2013;368:20120431.
- [158] Nebert DW, Zhang G, Vesell ES. From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug Metab Rev* 2008;40:187–224.
- [159] Ng SF, Lin RC, Laybutt DR, Barres R, Owens JA, Morris MJ. Chronic high-fat diet in fathers programs β -cell dysfunction in female rat offspring. *Nature* 2010;467:963–6.
- [160] Niemeijer MN, van den Berg ME, Eijgelsheim M, Rijnbeek PR, Stricker BH. Pharmacogenetics of drug-induced QT-interval prolongation: an update. *Drug Saf* 2015;38:855–67.
- [161] Niemi M. Transporter pharmacogenetics and statin toxicity. *Clin Pharmacol Ther* 2010;87:130–3.
- [162] Otonkoski T, Kaminen N, Ustinov J, Lapatto R, Meissner T, Mayatepek E, Kere J, Sipila I. Physical exercise-induced hyperinsulinemic hypoglycemia is an autosomal-dominant trait characterized by abnormal pyruvate-induced insulin release. *Diabetes* 2003;52:199–204.
- [163] Owusu Obeng A, Egelund EF, Alsultan A, Peloquin CA, Johnson JA. *CYP2C19* polymorphisms and therapeutic drug monitoring of voriconazole: are we ready for clinical implementation of pharmacogenomics? *Pharmacotherapy* 2014;34:703–18.
- [164] Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T. Functional SNPs in the lymphotoxin- α gene (*LTA*) that are associated with susceptibility to myocardial infarction. *Nat Genet* 2002;32:650–4.
- [165] Parham LR, Briley LP, Li L, Shen J, Newcombe PJ, King KS, Slater AJ, Dilthey A, Iqbal Z, McVean G, Cox CJ, Nelson MR, Spraggs CF. Comprehensive genome-wide evaluation of lapatinib-induced liver injury yields single genetic signal centered on known risk allele *HLA-DRB1*07:01*. *Pharmacogenomics J* 2016;16:180–5.
- [166] Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni Jr JF, Chatterjee N. Distribution of allele frequencies and effect-sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A* 2011;108:18026–31.
- [167] Parra EJ, Botton MR, Perini JA, Krithika S, Bourgeois S, Johnson TA, Tsunoda T, Pirmohamed M, Wadelius M, Limdi NA, Cavallari LH, Burmester JK, Rettie AE, Klein TE, Johnson JA, Hutz MH, Suarez-Kurtz G. Genome-wide association study of warfarin maintenance dose in a Brazilian sample. *Pharmacogenomics* 2015;16:1253–63.
- [168] Pembrey ME. Time to take epigenetic inheritance seriously. *Eur J Hum Genet* 2002;10:669–71.
- [169] Perera MA, Cavallari LH, Limdi NA, Gamazon ER, Konkashbaev A, Daneshjou R, Pluzhnikov A, Crawford DC, Wang J, Liu N, Tatonetti N, Bourgeois S, Takahashi H, Bradford Y, Burkley BM, Desnick RJ, Halperin JL, Khalifa SI, Langae TY, Lubitz SA, Nutescu EA, Oetjens M, Shahin MH, Patel SR, Sagreiya H, Tector M, Weck KE, Rieder MJ, Scott SA, Wu AH, Burmester JK, Wadelius M, Deloukas P, Wagner MJ, Mushiroda T, Kubo M, Roden DM, Cox NJ, Altman RB, Klein TE, Nakamura Y, Johnson JA. Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. *Lancet* 2013;382:790–6.
- [170] Pierce BL, Kibriya MG, Tong L, Jasmine F, Argos M, Roy S, Paul-Brutus R, Rahman R, Rakibuz-Zaman M, Parvez F, Ahmed A, Quasem I, Hore SK, Alam S, Islam T, Slavkovich V, Gamble MV, Yunus M, Rahman M, Baron JA, Graziano JH, Ahsan H. Genome-wide

- association study identifies chromosome 10q24.32 variants associated with arsenic metabolism and toxicity phenotypes in Bangladesh. *PLoS Genet* 2012;8:e1002522.
- [171] Pirmohamed M, Kamali F, Daly AK, Wadelius M. Oral anticoagulation: a critique of recent advances and controversies. *Trends Pharmacol Sci* 2015;36:153–63.
- [172] Postmus I, Trompet S, Deshmukh HA, Barnes MR, Li X, Warren HR, Chasman DI, Zhou K, Arsenault BJ, Donnelly LA, Wiggins KL, Avery CL, Griffin P, Feng Q, Taylor KD, Li G, Evans DS, Smith AV, de Keyser CE, Johnson AD, de Craen AJ, Stott DJ, Buckley BM, Ford I, Westendorp RG, Slagboom PE, Sattar N, Munroe PB, Sever P, Poulter N, Stanton A, Shields DC, O'Brien E, Shaw-Hawkins S, Chen YD, Nickerson DA, Smith JD, Dube MP, Boekholdt SM, Hovingh GK, Kastelein JJ, McKeigue PM, Betteridge J, Neil A, Durrington PN, Doney A, Carr F, Morris A, McCarthy MI, Groop L, Ahlqvist E, Bis JC, Rice K, Smith NL, Lumley T, Whitsel EA, Sturmer T, Boerwinkle E, Ngwa JS, O'Donnell CJ, Vasan RS, Wei WQ, Wilke RA, Liu CT, Sun F, Guo X, Heckbert SR, Post W, Sotoodehnia N, Arnold AM, Stafford JM, Ding J, Herrington DM, Kritchevsky SB, Eiriksdottir G, Launer LJ, Harris TB, Chu AY, Giulianini F, MacFadyen JG, Barratt BJ, Nyberg F, Stricker BH, Uitterlinden AG, Hofman A, Rivadeneira F, Emilsson V, Franco OH, Ridker PM, Gudnason V, Liu Y, Denny JC, Ballantyne CM, Rotter JJ, Adrienne CL, Psaty BM, Palmer CN, Tardif JC, Colhoun HM, Hitman G, Krauss RM, Wouter JJ, Caulfield MJ. Pharmacogenetic meta-analysis of genome-wide association studies of LDL-cholesterol response to statins. *Nat Commun* 2014;5:5068.
- [173] Powis SH, Mockridge I, Kelly A, Kerr LA, Glynne R, Gileadi U, Beck S, Trowsdale J. Polymorphism in a second *ABC* transporter gene located within the class II region of the human major histocompatibility complex. *Proc Natl Acad Sci U S A* 1992;89:1463–7.
- [174] Price Evans DA, Manley KA, McKusick VA. Genetic control of isoniazid metabolism in man. *Br Med J* 1960;2:485–91.
- [175] Quadri SA, Singal DP. Peptide transport in human lymphoblastoid and tumor cells: effect of transporter associated with antigen presentation (TAP) polymorphism. *Immunol Lett* 1998;61:25–31.
- [176] Ramsey LB, Johnson SG, Caudle KE, Haidar CE, Voora D, Wilke RA, Maxwell WD, McLeod HL, Krauss RM, Roden DM, Feng Q, Cooper-DeHoff RM, Gong L, Klein TE, Wadelius M, Niemi M. Clinical pharmacogenetics implementation consortium guideline for *SLCO1B1* and simvastatin-induced myopathy: 2014 update. *Clin Pharmacol Ther* 2014;96:423–8.
- [177] Ramsey LB, Panetta JC, Smith C, Yang W, Fan Y, Winick NJ, Martin PL, Cheng C, Devidas M, Pui CH, Evans WE, Hunger SP, Loh M, Relling MV. Genome-wide study of methotrexate clearance replicates *SLCO1B1*. *Blood* 2013;121:898–904.
- [178] Richards EJ. Inherited epigenetic variation -- revisiting soft inheritance. *Nat Rev Genet* 2006;7:395–401.
- [179] Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516–7.
- [180] Rogers MAM, Aronoff DM. The influence of non-steroidal anti-inflammatory drugs on the gut microbiome. *Clin Microbiol Infect* 2016;22:178. e171–178 e179.
- [181] Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Horta-nagel K, Pelz HJ, Lappegard K, Seifried E, Scharrer I, Tuddenham EG, Muller CR, Strom TM, Oldenburg J. Mutations in *VKORC1* cause warfarin resistance and multiple coagulation factor deficiency type-2. *Nature* 2004;427:537–41.
- [182] Sabbagh A, Darlu P, Crouau-Roy B, Poloni ES. Arylamine *N*-acetyltransferase 2 (*NAT2*) genetic diversity and traditional subsistence: a worldwide population survey. *PLoS One* 2011;6:e18507.
- [183] Sackton TB, Hartl DL. Genotypic context and epistasis in individuals and populations. *Cell* 2016;166:279–87.
- [184] Sarasquete ME, Garcia-Sanz R, Marin L, Alcoceba M, Chillón MC, Balanzategui A, Santamaria C, Rosinol L, de la Rubia J, Hernandez MT, Garcia-Navarro I, Lahuerta JJ, Gonzalez M, San Miguel JF. Bisphosphonate-related osteonecrosis of the jaw is associated with polymorphisms of cytochrome P450 *CYP2C8* in multiple myeloma: a genome-wide single-nucleotide polymorphism analysis. *Blood* 2008;112:2709–12.
- [185] Savage DC. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* 1977;31:107–33.
- [186] Schwartzman ML, Davis KL, Nishimura M, Abraham NG, Murphy RC. The cytochrome P450 metabolic pathway of arachidonic acid in the cornea. *Adv Prostag Thromb Leukot Res* 1991;21A:185–92.
- [187] Scott LJ. Sitagliptin: a review in type-2 diabetes. *Drugs* 2017;77:209–24.
- [188] Scott SA, Sangkuhl K, Stein CM, Hulot JS, Mega JL, Roden DM, Klein TE, Sabatine MS, Johnson JA, Shuldiner AR, Clinical Pharmacogenetics Implementation Consortium guidelines for *CYP2C19* genotype and clopidogrel therapy: 2013 update. *Clin Pharmacol Ther* 2013;94:317–23.
- [189] Shahin MH, Conrado DJ, Gonzalez D, Gong Y, Lobmeyer MT, Beitelshes AL, Boerwinkle E, Gums JG, Chapman A, Turner ST, Cooper-DeHoff RM, Johnson JA. Genome-wide association approach identified novel genetic predictors of heart rate response to β -blockers. *J Am Heart Ass* 2018;7:e006463.
- [190] Shi S, Klotz U. Age-related changes in pharmacokinetics. *Curr Drug Metabol* 2011;12:601–10.

- [191] Shichi H, Nebert DW. Genetic differences in drug metabolism associated with ocular toxicity. *Environ Health Perspect* 1982;44:107–17.
- [192] Singer JB, Lewitzky S, Leroy E, Yang F, Zhao X, Klickstein L, Wright TM, Meyer J, Paulding CA. A genome-wide study identifies *HLA* alleles associated with lumiracoxib-related liver injury. *Nat Genet* 2010;42:711–4.
- [193] Singh M, Bhatia P, Khera S, Trehan A. Emerging role of *NUDT15* polymorphisms in 6-mercaptopurine metabolism and dose related toxicity in acute lymphoblastic leukaemia. *Leuk Res* 2017;62:17–22.
- [194] Singh S, McDonough CW, Gong Y, Alghamdi WA, Arwood MJ, Bargal SA, Dumeny L, Li WY, Mehanna M, Stockard B, Yang G, de Oliveira FA, Fredette NC, Shahin MH, Bailey KR, Beitelshes AL, Boerwinkle E, Chapman AB, Gums JG, Turner ST, Cooper-DeHoff RM, Johnson JA. Genome-wide association study identifies the *HMGCS2* locus to be associated with chlorthalidone-induced glucose increase in hypertensive patients. *J Am Heart Ass* 2018;7:e007339.
- [195] Smith AH, Jensen KP, Li J, Nunez Y, Farrer LA, Hakonarson H, Cook-Sather SD, Kranzler HR, Gelernter J. Genome-wide association study of therapeutic opioid dosing identifies a novel locus upstream of *OPRM1*. *Mol Psychiatr* 2017;22:346–52.
- [196] Snyder LH. Studies in human inheritance. IX, the inheritance of taste deficiency in man. *Ohio J Sci* 1932;32:436–68.
- [197] Soldin OP, Mattison DR. Sex differences in pharmacokinetics and pharmacodynamics. *Clin Pharmacokinet* 2009;48:143–57.
- [198] Song S, Wang W, Hu P. Famine, death, and madness: schizophrenia in early adulthood after prenatal exposure to the Chinese Great Leap Forward Famine. *Soc Sci Med* 2009;68:1315–21.
- [199] Sprouse AA, van Breemen RB. Pharmacokinetic interactions between drugs and botanical dietary supplements. *Drug Metab Dispos* 2016;44:162–71.
- [200] Stadhouders R. Expanding the toolbox for 3D genomics. *Nat Genet* 2018;50:634–5.
- [201] Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 2011;187:367–83.
- [202] Strassburg CP, Kneip S, Topp J, Obermayer-Straub P, Barut A, Tukey RH, Manns MP. Polymorphic gene regulation and interindividual variation of UDP-glucuronosyltransferase activity in human small intestine. *J Biol Chem* 2000;275:36164–71.
- [203] Such E, Cervera J, Terpos E, Bagan JV, Avaria A, Gomez I, Margaix M, Ibanez M, Luna I, Cordon L, Roig M, Sanz MA, Dimopoulos MA, de la Rubia J. *CYP2C8* gene polymorphism and bisphosphonate-related osteonecrosis of the jaw in patients with multiple myeloma. *Haematologica* 2011;96:1557–9.
- [204] Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, Bassendine M, Spengler U, Dore GJ, Powell E, Riordan S, Sheridan D, Smedile A, Fragomeli V, Muller T, Bahlo M, Stewart GJ, Booth DR, George J. *IL28B* is associated with response to chronic hepatitis-C interferon- α and ribavirin therapy. *Nat Genet* 2009;41:1100–4.
- [205] Susser E, Hoek HW, Brown A. Neurodevelopmental disorders after prenatal famine: the story of the Dutch Famine Study. *Am J Epidemiol* 1998;147:213–6.
- [206] Takagi M, Uno H, Nishi R, Sugimoto M, Hasegawa S, Piao J, Ihara N, Kanai S, Kakei S, Tamura Y, Suganami T, Kamei Y, Shimizu T, Yasuda A, Ogawa Y, Mizutani S. *ATM* regulates adipocyte differentiation and contributes to glucose homeostasis. *Cell Rep* 2015;10:957–67.
- [207] Takahara S. Progressive oral gangrene probably due to lack of catalase in the blood (acatalasaemia): report of nine cases. *Lancet* 1952;2:1101–4.
- [208] Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, Whittaker P, Ranganath V, Kumanduri V, McLaren W, Holm L, Lindh J, Rane A, Wadelius M, Deloukas P. Genome-wide association study confirms *VKORC1*, *CYP2C9*, and *CYP4F2* as principal genetic determinants of warfarin dose. *PLoS Genet* 2009;5:e1000433.
- [209] Tamm R, Magi R, Tremmel R, Winter S, Mihailov E, Smid A, Moricke A, Klein K, Schrappe M, Stanulla M, Houlston R, Weinshilboum R, Mlinaric Rascan I, Metspalu A, Milani L, Schwab M, Schaeffeler E. Polymorphic variation in *TPMT* is the principal determinant of TPMT phenotype: a meta-analysis of three genome-wide association studies. *Clin Pharmacol Ther* 2017;101:684–95.
- [210] Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S, Ito Y, Mita E, Tanaka E, Mo-chida S, Murawaki Y, Honda M, Sakai A, Hiasa Y, Nishiguchi S, Koike A, Sakaida I, Imamura M, Ito K, Yano K, Masaki N, Sugauchi F, Izumi N, Tokunaga K, Mizokami M. Genome-wide association of *IL28B* with response to pegylated interferon- α and ribavirin therapy for chronic hepatitis-C. *Nat Genet* 2009;41:1105–9.
- [211] Taylor DR, Ingvarsson PK. Common features of segregation distortion in plants and animals. *Genetica* 2003;117:27–35.
- [212] Teichert M, Eijgelsheim M, Rivadeneira F, Uitterlinden AG, van Schaik RH, Hofman A, De Smet PA, van GT, Visser LE, Stricker BH. A genome-wide association study of acenocoumarol maintenance dosage. *Hum Mol Genet* 2009;18:3758–68.

- [213] Teng YS. Human liver aldehyde dehydrogenase in Chinese and Asiatic Indians: gene deletion and its possible implications in alcohol metabolism. *Biochem Genet* 1981;19:107–14.
- [214] Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet* 2010;11:259–72.
- [215] Thomas DL, Thio CL, Martin MP, Qi Y, Ge D, O’Huigin C, Kidd J, Kidd K, Khakoo SI, Alexander G, Goedert JJ, Kirk GD, Donfield SM, Rosen HR, Tobler LH, Busch MP, McHutchison JG, Goldstein DB, Carrington M. Genetic variation in *IL28B* and spontaneous clearance of hepatitis C virus. *Nature* 2009;461:798–801.
- [216] Thompson AJ, McHutchison JG. Will *IL28B* polymorphism remain relevant in the era of direct-acting antiviral agents for hepatitis C virus? *Hepatology* 2012;56:373–81.
- [217] Townsend DM, Tew KD. The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* 2003;22:7369–75.
- [218] Traver RD, Horikoshi T, Danenberg KD, Stadlbauer TH, Danenberg PV, Ross D, Gibson NW. NAD(P)H:quinone oxidoreductase gene expression in human colon carcinoma cells: characterization of a mutation which modulates DT-diaphorase activity and mitomycin sensitivity. *Canc Res* 1992;52:797–802.
- [219] Trevino LR, Shimasaki N, Yang W, Panetta JC, Cheng C, Pei D, Chan D, Sparreboom A, Giacomini KM, Pui CH, Evans WE, Relling MV. Germline genetic variation in an organic anion transporter polypeptide associated with methotrexate pharmacokinetics and clinical effects. *J Clin Oncol* 2009;27:5972–8.
- [220] Turner ST, Bailey KR, Fridley BL, Chapman AB, Schwartz GL, Chai HS, Sicotte H, Kocher JP, Rodin AS, Boerwinkle E. Genomic association analysis suggests chromosome 12 locus influencing anti-hypertensive response to thiazide diuretic. *Hypertension* 2008;52:359–65.
- [221] Turner ST, Bailey KR, Schwartz GL, Chapman AB, Chai HS, Boerwinkle E. Genomic association analysis identifies multiple loci influencing anti-hypertensive response to an angiotensin II receptor blocker. *Hypertension* 2012;59:1204–11.
- [222] Turner ST, Boerwinkle E, O’Connell JR, Bailey KR, Gong Y, Chapman AB, McDonough CW, Beitelshes AL, Schwartz GL, Gums JG, Padmanabhan S, Hiltunen TP, Citterio L, Donner KM, Hedner T, Lanzani C, Melander O, Saarela J, Ripatti S, Wahlstrand B, Manunta P, Kontula K, Dominiczak AF, Cooper-DeHoff RM, Johnson JA. Genomic association analysis of common variants influencing anti-hypertensive response to hydrochlorothiazide. *Hypertension* 2013;62:391–7.
- [223] Uher R, Perroud N, Ng MY, Hauser J, Henigsberg N, Maier W, Mors O, Placentino A, Rietschel M, Souery D, Zagar T, Czerski PM, Jerman B, Larsen ER, Schulze TG, Zobel A, Cohen-Woods S, Pirlo K, Butler AW, Muglia P, Barnes MR, Lathrop M, Farmer A, Breen G, Aitchison KJ, Craig I, Lewis CM, McGuffin P. Genome-wide pharmacogenetics of anti-depressant response in the GENDEP project. *Am J Psychiatr* 2010;167:555–64.
- [224] van Leeuwen N, Nijpels G, Becker ML, Deshmukh H, Zhou K, Stricker BH, Uitterlinden AG, Hofman A, van ’t RE, Palmer CN, Guigas B, Slagboom PE, Durrington P, Calle RA, Neil A, Hitman G, Livingstone SJ, Colhoun H, Holman RR, McCarthy MI, Dekker JM, ’t Hart LM, Pearson ER. A gene variant near *ATM* is significantly associated with metformin treatment response in type-2 diabetes: replication and meta-analysis of five cohorts. *Diabetologia* 2012;55:1971–7.
- [225] Vasilevsky NA, Foster ED, Engelstad ME, Carmody L, Might M, Chambers C, Dawkins HJS, Lewis J, Della Rocca MG, Snyder M, Boerkoel CF, Rath A, Terry SF, Kent A, Searle B, Baynam G, Jones E, Gavin P, Bamshad M, Chong J, Groza T, Adams D, Resnick AC, Heath AP, Mungall C, Holm IA, Rageth K, Brownstein CA, Shefchek K, McMurphy JA, Robinson PN, Kohler S, Haendel MA. Plain-language medical vocabulary for precision diagnosis. *Nat Genet* 2018;50:474–6.
- [226] Vesell ES. Factors altering the responsiveness of mice to hexobarbital. *Pharmacology* 1968;1:81–97.
- [227] Vesell ES. Pharmacogenetics. *N Engl J Med* 1972;287:904–9.
- [228] Vesell ES, Page JG. Genetic control of drug levels in man: phenylbutazone. *Science* 1968;159:1479–80.
- [229] Vogel F. Moderne probleme der Humangenetik. In: Heilmeyer L, Schoen R, de Rudder B, editors. *Ergebnisse der Inneren Medizin und Kinderheilkunde*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1959. p. 52–125.
- [230] Voora D, Shah SH, Spasojevic I, Ali S, Reed CR, Salisbury BA, Ginsburg GS. The *SLCO1B1**5 genetic variant is associated with statin-induced side effects. *J Am Coll Cardiol* 2009;54:1609–16.
- [231] Wang K, Zhou B, Kuo YM, Zemansky J, Gitschier J. A novel member of a zinc transporter family is defective in acrodermatitis enteropathica. *Am J Hum Genet* 2002;71:66–73.
- [232] Wei J, Wang DW, Alings M, Fish F, Wathen M, Roden DM, George Jr AL. Congenital long-QT syndrome caused by a novel mutation in a conserved acidic domain of the cardiac Na^+ channel. *Circulation* 1999;99:3165–71.
- [233] Weiner M, Shapiro S, Axelrod J, Cooper JR, Brodie BB. The physiological disposition of dicumarol in man. *J Pharmacol Exp Therapeut* 1950;99:409–20.

- [234] Weinshilboum R. Phenol sulfotransferase inheritance. *Cell Mol Neurobiol* 1988;8:27–34.
- [235] Weinshilboum RM, Sladek SL. Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thiopurine methyltransferase activity. *Am J Hum Genet* 1980;32:651–62.
- [236] Wiencke JK, Pemble S, Ketterer B, Kelsey KT. Gene deletion of glutathione S-transferase theta: correlation with induced genetic damage and potential role in endogenous mutagenesis. *Cancer Epidemiol Biomark Prev* 1995;4:253–9.
- [237] Wirgin I, Roy NK, Loftus M, Chambers RC, Franks DG, Hahn ME. Mechanistic basis of resistance to PCBs in Atlantic tomcod from the Hudson River. *Science* 2011;331:1322–5.
- [238] Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. *Nature* 2007;447:279–83.
- [239] Woo SL, Lidsky AS, Guttler F, Chandra T, Robson KJ. Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. *Nature* 1983;306:151–5.
- [240] Woolhouse NM, Andoh B, Mahgoub A, Sloan TP, Idle JR, Smith RL. Debrisoquine hydroxylation polymorphism among Ghanaians and Caucasians. *Clin Pharmacol Ther* 1979;26:584–91.
- [241] Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TMF, Bacanu SA, Baekvad-Hansen M, Beekman AFT, Bigdeli TB, Binder EB, Blackwood DRH, Bryois J, Buttenschon HN, Bybjerg-Grauholm J, Cai N, Castelo E, Christensen JH, Clarke TK, Coleman JIR, Colodro-Conde L, Couvy-Duchesne B, Craddock N, Crawford GE, Crowley CA, Dashti HS, Davies G, Deary IJ, Degenhardt F, Derks EM, Direk N, Dolan CV, Dunn EC, Eley TC, Eriksson N, Escott-Price V, Kiadeh FHF, Finucane HK, Forstner AJ, Frank J, Gaspar HA, Gill M, Giusti-Rodriguez P, Goes FS, Gordon SD, Grove J, Hall LS, Hannon E, Hansen CS, Hansen TF, Herms S, Hickie IB, Hoffmann P, Homuth G, Horn C, Hottenga JJ, Hougaard DM, Hu M, Hyde CL, Ising M, Jansen R, Jin F, Jorgenson E, Knowles JA, Kohane IS, Kraft J, Kretschmar WW, Krogh J, Kutalik Z, Lane JM, Li Y, Li Y, Lind PA, Liu X, Lu L, MacIntyre DJ, MacKinnon DF, Maier RM, Maier W, Marchini J, Mbarek H, McGrath P, McGuffin P, Medland SE, Mehta D, Middeldorp CM, Mihailov E, Milanesechi Y, Milani L, Mill J, Mondimore FM, Montgomery GW, Mostafavi S, Mullins N, Nauck M, Ng B, Nivard MG, Nyholt DR, O'Reilly PF, Oskarsson H, Owen MJ, Painter JN, Pedersen CB, Pedersen MG, Peterson RE, Pettersson E, Peyrot WJ, Pistis G, Posthuma D, Purcell SM, Quiroz JA, Qvist P, Rice JP, Riley BP, Rivera M, Saeed Mirza S, Saxena R, Schoevers R, Schulte EC, Shen L, Shi J, Shyn SI, Sigurdsson E, Sinnamoni GBC, Smit JH, Smith DJ, Stefansson H, Steinberg S, Stockmeier CA, Streit F, Strohmaier J, Tansey KE, Teismann H, Teumer A, Thompson W, Thomson PA, Thorgeirsson TE, Tian C, Traylor M, Treutlein J, Trubetskoy V, Uitterlinden AG, Umbrecht D, Van der Auwera S, van Hemert AM, Viktorin A, Visscher PM, Wang Y, Webb BT, Weinsheimer SM, Wellmann J, Willemsen G, Witt SH, Wu Y, Xi HS, Yang J, Zhang F, eQTLgen, and Me, Arolt V, Baune BT, Berger K, Boomsma DI, Cichon S, Dannlowski U, de Geus ECJ, DePaulo JR, Domenici E, Domschke K, Esko T, Grabe HJ, Hamilton SP, Hayward C, Heath AC, Hinds DA, Kendler KS, Kloiber S, Lewis G, Li QS, Lucae S, Madden PFA, Magnusson PK, Martin NG, McIntosh AM, Metspalu A, Mors O, Mortensen PB, Muller-Myhsok B, Nordentoft M, Nothen MM, O'Donovan MC, Paciga SA, Pedersen NL, Penninx B, Perlis RH, Porteous DJ, Potash JB, Preisig M, Rietschel M, Schaefer C, Schulze TG, Smoller JW, Stefansson K, Tiemeier H, Uher R, Volzke H, Weissman MM, Werge T, Winslow AR, Lewis CM, Levinson DF, Breen G, Borglum AD, Sullivan PF, Major Depressive Disorder Working Group of the Psychiatric Genomics C. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 2018;50:668–81.
- [242] Wright E, Schofield PT, Molokhia M. Bisphosphonates and evidence for association with esophageal and gastric cancer: a systematic review and meta-analysis. *BMJ Open* 2015;5:e007133.
- [242a] Xu C, Goodz S, Sellers EM, Tyndale RF. *CYP2A6* genetic variation and potential consequences. *Advanc Drug Deliv Rev* 2002;54:1245–1256.
- [243] Xu H, Robinson GW, Huang J, Lim JY, Zhang H, Bass JK, Broniscer A, Chintagumpala M, Bartels U, Gururangan S, Hassall T, Fisher M, Cohn R, Yamashita T, Teitz T, Zuo J, Onar-Thomas A, Gajjar A, Stewart CF, Yang JJ. Common variants in *ACYP2* influence susceptibility to cisplatin-induced hearing loss. *Nat Genet* 2015;47:263–6.
- [244] Xu MQ, Sun WS, Liu BX, Feng GY, Yu L, Yang L, He G, Sham P, Susser E, St Clair D, He L. Prenatal malnutrition and adult schizophrenia: further evidence from the 1959-1961 Chinese famine. *Schizophr Bull* 2009;35:568–76.
- [245] Yamano S, Nhamburo PT, Aoyama T, Meyer UA, Inaba T, Kalow W, Gelboin HV, McBride OW, Gonzalez FJ. cDNA cloning and sequence and cDNA-directed expression of human P450 IIB1: identification of a normal and two variant cDNAs derived from the *CYP2B* locus on chromosome 19 and differential expression of the IIB mRNAs in human liver. *Biochemistry* 1989;28:7340–8.

- [246] Ylitalo P. Effect of exercise on pharmacokinetics. *Ann Med* 1991;23:289–94.
- [247] Yue Q, Jen JC, Thwe MM, Nelson SF, Baloh RW. De novo mutation in *CACNA1A* causes acetazolamide-responsive episodic ataxia. *Am J Med Genet* 1998;77:298–301.
- [248] Zhang G, Nebert DW. Personalized medicine: genetic risk prediction of drug response. *Pharmacol Ther* 2017;175:75–90.
- [249] Zhang G, Nebert DW, Chakraborty R, Jin L. Statistical power of association using the extreme discordant phenotype design. *Pharmacogenet Genom* 2006;16:401–13.
- [250] Zhao R, Min SH, Qiu A, Sakaris A, Goldberg GL, Sandoval C, Malatack JJ, Rosenblatt DS, Goldman ID. The spectrum of mutations in the *PCFT* gene, coding for an intestinal folate transporter, the basis for hereditary folate malabsorption. *Blood* 2007;110:1147–52.
- [251] Zhivotovsky LA, Rosenberg NA, Feldman MW. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* 2003;72:1171–86.
- [252] Zhou K, Yee SW, Seiser EL, van LN, Tavendale R, Bennett AJ, Groves CJ, Coleman RL, van der Heijden AA, Beulens JW, de Keyser CE, Zaharenko L, Rotroff DM, Out M, Jablonski KA, Chen L, Javorsky M, Zidzik J, Levin AM, Williams LK, Dujic T, Semiz S, Kubo M, Chien HC, Maeda S, Witte JS, Wu L, Tkac I, Kooy A, van Schaik RH, Stehouwer CD, Logie L, Sutherland C, Klovins J, Pirags V, Hofman A, Stricker BH, Motsinger-Reif AA, Wagner MJ, Innocenti F, Hart LM, Holman RR, McCarthy MI, Hedderson MM, Palmer CN, Florez JC, Giacomini KM, Pearson ER. Variation in the glucose transporter gene *SLC2A2* is associated with glycemic response to metformin. *Nat Genet* 2016;48:1055–9.
- [253] Zhou LX, Pihlstrom B, Hardwick JP, Park SS, Wrighton SA, Holtzman JL. Metabolism of phenytoin by the gingiva of normal humans: possible role of reactive metabolites of phenytoin in the initiation of gingival hyperplasia. *Clin Pharmacol Ther* 1996;60:191–8.

FURTHER READING

- Ahern TP. Pharmacoepidemiology in pharmacogenetics. *Adv Pharmacol* 2018;83:109–30.
- Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 2017;169:1177–86.
- Daly AK. Using genome-wide association studies to identify genes important in serious adverse drug reactions. *Annu Rev Pharmacol Toxicol* 2012;52:21–35.
- Doestzada M, Vila AV, Zhernakova A, Koonen DPY, Weersma RK, Touw DJ, Kuipers F, Wijmenga C, Fu J. Pharmacomicrobiomics: a novel route towards personalized medicine? *Protein Cell*. 2018;9:432–45.
- Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet* 2011;13:135–45.
- Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009;360:1696–8.
- Tornio A, Backman JT. Cytochrome P450 in pharmacogenetics: an update. *Adv Pharmacol* 2018;83:3–32.
- Zhang G, Nebert DW. Personalized medicine: genetic risk prediction of drug response. *Pharmacol Ther* 2017;175:75–90.

Note: Page numbers followed by “f” indicate figures, “t” indicate tables.

- A**
- AAV. *See* Adeno-associated virus (AAV)
- Abacavir-induced hypersensitivity, 460
- Aberrant miRNA in cancer, 100
- Abnormal phenotypes scope, 376–377
- homeostasis, 376–377
- Abnormal proteins due to different genes fusion, 165
- Absorption, distribution, metabolism, and excretion (ADME), 447
- Acetyl-CoA, 268–269
- N-Acetylation polymorphism, 454–456
- ACMG of Medical Genetics (ACMG). *See* American College
- Active principle, 446
- Acute lymphocytic leukemia (ALL), 458
- AD. *See* Alzheimer disease (AD)
- ADA gene. *See* Adenosine deaminase gene (ADA gene)
- Adenine (A), 55
- Adenine nucleotide (ADP/ATP) translocator (ANT1), 285
- Adenine nucleotide translocators (ANTs), 270
- Adeno-associated virus (AAV), 307
- gene therapy, 29
- Adenosine deaminase gene (ADA gene), 155
- ADME. *See* Absorption, distribution, metabolism, and excretion (ADME)
- ADPKD. *See* Autosomal dominant polycystic kidney disease (ADPKD)
- ADRs. *See* Adverse drug reactions (ADRs)
- Adulthood, genetic disease and, 50
- Advanced glycation end products, 419
- Advanced glycosylation end products (AGE products), 424–425
- Adverse drug reactions (ADRs), 446
- 446t
- genome-wide association studies, 464–467
- DILI and hypersensitivity, 467
- drug-induced QT-interval prolongation, 467
- osteonecrosis of jaw, 467
- statin-induced myopathy, 464–467
- indistinguishable from complex diseases, 463–464
- AFP genes. *See* α -fetoprotein genes (AFP genes)
- AGE products. *See* Advanced glycosylation end products (AGE products)
- Aging, 272, 415–429
- alterations
- in DNA, 425–428
- in lipids, 428–429
- in proteins, 424–425
- pro-longevity loci and “antigeroid” syndromes, 434–435
- progeroid syndromes of humans, 429–433
- AgNOR. *See* Silver NOR (AgNOR)
- AGPAT2, 433
- AGT gene. *See* Angiotensin gene (AGT gene)
- AHO. *See* Albright hereditary osteodystrophy (AHO)
- AI. *See* Artificial intelligence (AI)
- Albright hereditary osteodystrophy (AHO), 220–221
- Aldehyde dehydrogenase-2 (ALDH2), 461
- ALDH2. *See* Aldehyde dehydrogenase-2 (ALDH2)
- Alectinib, 25
- ALL. *See* Acute lymphocytic leukemia (ALL)
- Allele frequencies, 334, 360
- factors affecting, 366–370
- migration, 370
- mutation, 367–368
- random genetic drift, 366–367
- selection, 368–370
- Allele-specific DNA methylation (ASM), 104
- Allele-specific RNA expression (ASE), 104
- Allele-specific transcription factor binding (ASTF binding), 104
- Allelic association, 329–330
- Allelic heterogeneity, 168–169, 213, 452
- Alpha-1-antitrypsin. *See* Serine proteinase inhibitor clade A member 1 (SERPINA1)
- α -fetoprotein genes (AFP genes), 163–164
- α -Thalassemia/mental retardation X-linked syndrome (ATR-X syndrome), 96–97
- ALS. *See* Amyotrophic lateral sclerosis (ALS)
- Alu repeat, 138
- Alzheimer disease (AD), 303, 378, 417
- American College of Medical Genetics (ACMG), 126
- Amino acid sequence, variants affecting, 151
- Aminoacyl tRNA synthetases, 71
- Amyotrophic lateral sclerosis (ALS), 304
- Analysis of variance (ANOVA), 335
- Anaphase, 239–241
- Anatomic structure development, 377–379
- elaborateness of repair, 378
- life history, 378–379
- Androgenetic tumors, 88
- “Aneuploidy”, 260
- Angelman syndromes (AS), 90, 220
- Angiopoietin-like protein 3 (ANGPTL3), 434–435
- Angiotensin gene (AGT gene), 152
- Angular homeostasis, 377–379
- ANNOVAR, 338

- ANOVA. *See* Analysis of variance (ANOVA)
- ANT1. *See* Adenine nucleotide (ADP/ATP) translocator (ANT1)
- Antagonistic pleiotropy, 420–421
- Anti-Müllerian hormone, 257
- Antiapoptotic mechanisms, targeting, 306
- Anticipation, 207, 208f
- Anticoagulants, efficacy of, 467–468
- Antisense RNA, position effect by, 165
- ANTs. *See* Adenine nucleotide translocators (ANTs)
- APOE2 alleles, 417–418, 434
- APOE4 alleles, 418, 434
- Archetypal ADRs, 463, 465t
- Array hybridization, 340
- ARSA. *See* Arylsulfatase A gene (ARSA)
- Arsenite methyltransferase gene (AS3MT gene), 471–472
- ART. *See* Assisted reproductive technologies (ART)
- Artificial intelligence (AI), 30
- Arylsulfatase A gene (ARSA), 162
- AS. *See* Angelman syndromes (AS)
- AS3MT gene. *See* Arsenite methyltransferase gene (AS3MT gene)
- ASD. *See* Autism spectrum disorder (ASD)
- ASE. *See* Allele-specific RNA expression (ASE)
- ASM. *See* Allele-specific DNA methylation (ASM)
- Assisted reproductive technologies (ART), 102–103, 390
on epigenetic programming, 102–103
- Association methods/statistical analysis
discovery phase of genome-wide association study, 335
postanalysis quality control, 335–336
power and sample size calculations, 334–335
validation and replication phase, 336–337
- Assortative mating, 365–366
- ASTF binding. *See* Allele-specific transcription factor binding (ASTF binding)
- AT. *See* Ataxia-telangiectasia (AT)
- AT-rich DNA sequences, 249
- Ataxia-telangiectasia (AT), 432
- Ataxia-telangiectasia-mutated gene (ATM gene), 468
- Atelosteogenesis, 369
- Atherosclerosis, 420–421, 434–435
- ATM gene. *See* Ataxia-telangiectasia-mutated gene (ATM gene)
- ATP, 270
- ATP6 protein, 270
- ATR-X syndrome. *See* α -Thalassemia/mental retardation X-linked syndrome (ATR-X syndrome)
- ATRX gene, 96–97
- Atypical twinning, 389–391
chimeric twins, 390
CHM with coexistent twin, 391
fetus-in-fetu, 391
mirror-image twins, 390
polar body twins, 390
superfecundation, 391
superfetation, 391
VTS, 390
- AUG codon, 72
- Autism spectrum disorder (ASD), 101–102, 304–305
- Autosomal dominant inheritance, 206–210, 206f, 206t
anticipation, 207
expressivity, 207
gonadal or somatic mosaicism, 209–210
mechanisms of reduced penetrance and variable expressivity, 208–209
environmental factors, 208
genetic background, 208–209
somatic variants, 208
unstable DNA triplet-repeat sequences, 208
new dominant variants, 209
penetrance, 206–207
pleiotropy, 207
recurrence risks, 206
sex influence, 207
sex limitation, 207
- Autosomal dominant polycystic kidney disease (ADPKD), 208–209
- Autosomal dominant traits, 452
- Autosomal locus, 360–361
- Autosomal recessive inheritance, 210–213, 210t
consanguinity, 211
genetic heterogeneity, 212–213
new variants, 213
recurrence risks, 211
UPD, 213
- Autosomal recessive trait, 450–451
- Azathioprine, 458
- ## B
- B recognition elements sequence (BRE sequence), 67–68
- Balanced polymorphism, 368
- Bayesian framework, 338
- Bayesian hierarchical model, 339–341
- BCKDH. *See* Branched chain-keto acid dehydrogenase (BCKDH)
- BDNF gene, 95
- Beadle–Tatum one gene–one enzyme principle, 9–10
- Beckwith–Wiedemann syndrome (BWS), 89
- Beckwith–Wiedemann syndrome, 397
- Berardinelli–Seip congenital lipodystrophy 2 (BSC12), 433
- β -endorphin, 150–151
- β -globin gene cluster, 162
- β -melanocyte–stimulating hormone (β -MSH), 150–151
- β -satellite repeats, 141
- Betabinomial distributions, 340–341
- Biallelic markers, 229
- Bimodal distribution, 454, 457
- Binding protein (BiP), 74–75
- Biochemical genetics, 8
- Bioenergetic disease, 284
- Bioenergetics, 267
- Biological basis of linkage analysis, 228
- Biometrics inheritance school, 452
- BiP. *See* Binding protein (BiP)
- Bisulfiteconversion–based sequencing (BS-Seq), 340–341
- BMP2 gene. *See* Bone morphogenetic protein 2 gene (BMP2 gene)
- BMP15 genes. *See* Bone morphogenetic protein 15 genes (BMP15 genes)
- BMPR1B. *See* Bone morphogenetic protein receptor 1B (BMPR1B)
- Bone morphogenetic protein 15 genes (BMP15 genes), 396
- Bone morphogenetic protein receptor 1B (BMPR1B), 396

- Bone morphogenic protein 2 gene (*BMP2* gene), 143–144
- Bonferroni correction, 342–343
- BRACAnalysis CDx, 25–29
- Brain, 378
- Brain–gut–microbiome, 448
- Branched chain-keto acid dehydrogenase (BCKDH), 268–269
- Branching pathways, 380
- BrdU. *See* Bromodeoxyuridine (BrdU)
- BRE sequence. *See* B recognition elements sequence (BRE sequence)
- Breakpoint junction cluster, 146–147
- British DDD study, 167
- Bromodeoxyuridine (BrdU), 247
- BS-Seq. *See* Bisulfiteconversion–based sequencing (BS-Seq)
- BSCL2*. *See* Berardinelli–Seip congenital lipodystrophy 2 (*BSCL2*)
- BWS. *See* Beckwith–Wiedemann syndrome (BWS)
- C**
- C-band-positive regions, 249–250
- Ca²⁺-activated cyclophilin D (cypD), 270
- Calcium flux modulation, 306
- Calorically restricted rodents, 425
- Cancer epigenetics, 98–101
- cancer
- aberrant miRNA and lncRNA expression, 100
 - abnormalities of histones and histone modifications, 100
 - DNA hypermethylation, 99
 - DNA hypomethylation, 99
 - therapies targeting epigenetic modifications, 100–101
- Cap site variants, 158–159
- Capture-based sequencing, 340–341
- CAR-T therapy. *See* Chimeric antigen receptors T-cell therapy (CAR-T therapy)
- Carbohydrate metabolism disorders in segmental progeroid phenotypes, 433
- Carcinogens, 49–50
- Cardiovascular diseases, 24
- GWAS for treatment, 469
- Cascade screening, 23
- Causal variant identification for common diseases, 371–372
- CD/CV hypothesis. *See* Common disease–common variant hypothesis (CD/CV hypothesis)
- CD/RV hypothesis. *See* Common disease–rare variant hypothesis (CD/RV hypothesis)
- CDKN2A* gene, 99
- Cell division, chromosomes in, 239–244
- meiosis, 241–243
 - mitosis, 239–241
- Cellular antioxidant pathway stimulation, 306
- Cellular consequences of trinucleotide repeat expansions, 163
- CENP-A, 255
- Central metabolic pathways, 268–269
- Centre d'Etude du PolymorphismeHumain (CEPH), 330
- Centromere (C), 61, 244–245, 254–255
- Centromere-banded karyotype of male cell, 248f
- Centromeric probes (CEPs), 250–251
- CEPH. *See* Centre d'Etude du PolymorphismeHumain (CEPH)
- CEPs. *See* Centromeric probes (CEPs)
- Cerebroretinal microangiopathy with calcifications and cysts-1 (CRMCC1). *See* COATS plus disease
- Ceritinib, 25
- CFH* gene. *See* Complement factor H gene (*CFH* gene)
- “CG to TG or CA rule”, 131
- CGH. *See* Comparative genomic hybridization (CGH)
- Chaperone, 74–75
- Characterized genomic variants, 125–126
- Charcot–Marie–Tooth disease type 2 (CMT2), 431
- CHARGE syndrome, 96
- CHD7* gene, 96
- Chemical individuality, 7
- Chiasmata, 243
- Childhood, genetic disease and, 50
- Chimera/chimeric/chimerism, 390
- twins, 390
- Chimeric antigen receptors T-cell therapy (CAR-T therapy), 25
- ChIP. *See* Chromatin immunoprecipitation (ChIP)
- CHM. *See* Complete hydatidiform mole (CHM)
- Chorionicity, 393–394
- Chromatin, 80–83, 98, 254
- chromatin-modifying enzymes, 100
 - human disorders due to mutations in chromatin remodelers, 96–97
 - marks, 84–85
- Chromatin immunoprecipitation (ChIP), 94, 98
- Chromosomal microarray analysis (CMA), 22, 252–254
- Chromosomal/chromosomes, 237
- abnormalities, 260–262
 - numerical chromosome abnormalities, 260–261
 - structural chromosome abnormalities, 261–262
- banding, 244–249, 250t
- reveals genome sequence organization, 249–250
- basis of inheritance
- chromosome abnormalities, 260–262
 - chromosome structure, 237–239
 - functional organization of chromosomes, 254–256
 - methods for studying human chromosomes, 244–254
 - sex chromosomes and sex determination, 256–259
 - uniparental disomy and imprinting, 259–260
- chromosome-conformation capture method, 471
- disorders, 47–48, 222–223, 223f
- DNA, 57
- functional organization, 254–256
- organization, 237–239
- polymorphisms, 248
- regions, 227–228
- segment, 227
- structure, 237–239
- chromosomes in cell division, 239–244
 - levels of DNA packaging in cell, 238f
 - spermatogenesis and oogenesis, 243–244

- Chromothripsis, 146–147
 Chronic disease, 400
 Chronic progressive external
 ophthalmoplegia (CPEO), 268
 ChunkChromosome tool, 333–334
 Cigarette smoking, 471
Cis-acting regulatory elements, 69
 Citrate, 268–269
 Classic twin design, 400–401
 Classical WS, 430
 “Classical” evolutionary biological
 theory of aging, 416
 CLF. *See* Coexisting live fetus (CLF)
 Clinical Pharmacogenomics
 Implementation Consortium
 (CPIC), 464–467
 Clinical pharmacology, fundamental
 aspects of, 446–450
 extrahepatic PGx differences and
 endogenous functions, 448–449
 genetics of drug response, 449–450
 PK and PD, 447–448
 plasma clearance of drug, 448
 therapeutic index or window, 449
 Clinical Sequencing Evidence-
 Generating Research (CSER), 30
 Clinical utility, 467, 473
 ClinSeq project, 24
 Closely spaced multiple mutations
 (CSMMs), 146
 Clusters, 65, 336
 CMA. *See* Chromosomal microarray
 analysis (CMA)
 CMT2. *See* Charcot–Marie–Tooth
 disease type 2 (CMT2)
 CNCs. *See* Conserved noncoding
 elements (CNCs)
 CNP. *See* Copy number polymorphism
 (CNP)
 CNVs. *See* Copy number variants/
 variation (CNVs)
 COATS plus disease, 433
 Cochran–Armitage test, 335
 Cockayne syndrome (CS), 432
 Codeine (parent drug), 446–447, 458
 Coding and noncoding regions
 functional scores, 338
 Coefficient of inbreeding, 364–365
 Coenzyme Q10 (CoQ), 269
 analogue EPI-743, 306
 Coexistent twin, CHM with, 391
 Coexisting live fetus (CLF), 391
 Coherence, source of, 11
 Cohort allelic sums test method,
 337–338
 COL2A1 gene, 168–169
 Collagens, 419
 Common disease–common variant
 hypothesis (CD/CV hypothesis),
 168, 330
 Common disease–rare variant
 hypothesis (CD/RV hypothesis),
 168
 Comparative genomic hybridization
 (CGH), 252
 Compensatory variants, 149–150
 Complement factor H gene (*CFH*
 gene), 453
 Complete hydratidiform mole (CHM),
 88, 391
 with coexistent twin, 391
 Complex diseases, mitochondrial
 etiology of, 285–305
 Complex PGx traits, 450
 GWAS of, 468–469
 cardiovascular disease treatment,
 469
 type II diabetes treatment,
 468–469
 Complex rearrangement, 145–146
 Concordance rate, 326
 Consanguinity, 211, 364–365
 Conserved noncoding elements
 (CNCs), 164–165
 Constitutive heterochromatin, 60
 “Contact–first” hypothesis, 136
 Conventional nosology, 11
 Copy number polymorphism (CNP),
 128–129
 Copy number variants/variation
 (CNVs), 63–64, 128–130, 136,
 252–254, 304–305, 371
 in association with disease,
 143–144
 Copy-and-paste mechanism, 61–62
 CoQ. *See* Coenzyme Q10 (CoQ)
 Corus CAD, 24
 Coupling AB/ab phase, 229
 COX. *See* Cytochrome c oxidase
 (COX)
 CPEO. *See* Chronic progressive
 external ophthalmoplegia
 (CPEO)
 CpG. *See* Cytosine–guanine (CpG)
 CPIC. *See* Clinical Pharmacogenomics
 Implementation Consortium
 (CPIC)
 Craniosynostosis, 217
 Crick strand, 55
 Cristae, 268–269
 Crizotinib, 25
 CRYGEP1 pseudogene-reactivating
 variant, 157
 CS. *See* Cockayne syndrome (CS)
 CSER. *See* Clinical Sequencing
 Evidence-Generating Research
 (CSER)
 CSMMs. *See* Closely spaced multiple
 mutations (CSMMs)
 cSNP, 371
 CYP21. *See* Steroid 21-hydroxylase
 (CYP21)
 CYP2C8 gene, 467
 CYP2C9 gene, 460–461
 CYP2C19 gene, 459
 CYP2D6
 1 allele, 458
 CYP2D6 gene, 456–458
 CYP3A4 gene, 448
 CYP4F2 gene, 460–461
 cypD. *See* Ca²⁺-activated cyclophilin D
 (cypD)
 Cystic fibrosis, 211, 323
 Cytochrome c oxidase (COX),
 269
 Cytogenetic abnormalities, 216
 Cytokinesis, 241
 Cytosine (C), 55, 131, 131f
 methylation, 131
 Cytosine–guanine (CpG), 399
 dinucleotides, 258
 hypermethylation in gene
 promoters, 99
 islands, 81
 Cytosolic acetyl-CoA, 268–269

D
 d3. *See* Deletion of exon 3 (d3)
 DAAs. *See* Direct antiviral agents
 (DAAs)
 DAB. *See* Diaminobenzidine (DAB)
 Damage-associated molecular patterns
 (DAMPs), 305
 DAT. *See* Dementias of the Alzheimer
 Type (DAT)
 Data Sciences, 30

- DCM1A. *See* Dilated cardiomyopathy type 1A (DCM1A)
- De novo mutation, 47
- De novo-database, 468
- Debrisoquine/sparteine oxidation polymorphism, 456–458
- Deletion of exon 3 (d3), 434
- Deletion polymorphisms, 129
- Dementias of the Alzheimer Type (DAT), 434
- DESeq2*, 339–340
- “Developmental origins of health and disease” hypothesis (DOHAD hypothesis), 101
- DFNA50. *See* Dominant progressive hearing loss (DFNA50)
- Diabetics, 419
- Diaminobenzidine (DAB), 58–60
- Diathesis, 4–5
- Dichorionic twin pairs, 394
- Dictyotene, 243
- Differentially methylated regions (DMRs), 87–88, 340
- Digenic inheritance, 169–170, 221
- Digital health tools, 31
- Dihydropyrimidine dehydrogenase polymorphism (*DPYD* gene), 460
- Dilated cardiomyopathy type 1A (DCM1A), 431
- DILI. *See* Drug-induced liver injury (DILI)
- Direct antiviral agents (DAAs), 468
- Disease
- definition, 4–5
 - principles, 2–4
 - variants, 148
- Disease-causing variants
- CNV in association with disease, 143–144
 - complex rearrangement, 145–146
 - duplications, 143
 - expansion/CNV of trinucleotide, 134–135
 - frequency of disease-producing variants, 148
 - functional characteristics of human disease genes, 148–149
 - gene conversion, 144–145
 - germline epimutations, 147
 - gross deletions, 137–140
 - Indels, 145
 - inversions, 142–143, 142f
 - large insertion of repetitive and other elements, 141–142
 - large retrotranspositional insertions, 140–141
 - location of repeat expansion, 135f
 - mechanisms of gross genomic rearrangement, 136–137
 - microdeletions and microinsertions, 132–134
 - molecular misreading, 147
 - multiple simultaneous mutations, 146–147
 - mutation/variant nomenclature, 149
 - mutations in gene evolution, 149
 - nature of genomic variants, 130
 - nucleotide substitutions, 130–132
 - size distribution, 133f
 - slipped mispairing model, 132f
 - synonymous nucleotide substitutions, 132
- Disease-producing variants frequency, 148
- Dizygotic twins (DZ twins), 326, 387
- genetic causes, 395–396
 - incidence, 392
- DM1. *See* Type 1 myotonic dystrophy (DM1)
- DMD. *See* Duchenne muscular dystrophy (DMD)
- DMEs. *See* Drug-metabolizing enzymes (DMEs)
- DMRs. *See* Differentially methylated regions (DMRs)
- DNA, 5–6, 237–239. *See also* RNA.
- alterations in, 425–428
 - germline mutations, 428
 - mtDNA, 427–428
 - nuclear DNA, 425–427
 - telomeric DNA, 427
- DNA–protein interactions, 249
- duplex, 237–239
- hypermethylation in cancer, 99
 - hypomethylation in cancer, 99
- looping, 64–65
- marker alleles, 370–371
- methylation, 30, 80–83, 102, 219–220, 259, 340
- mapping, 97
- polymorphisms, 126–129
- replication, 55–56, 55f
- sequence, 399
- synthesis, 57
 - transposons, 61–62
- DNA methyltransferase 1 (DNMT1), 80–81
- DNase I-resistant, 162
- DNMT1. *See* DNA methyltransferase 1 (DNMT1)
- DOHAD hypothesis. *See* “Developmental origins of health and disease” hypothesis (DOHAD hypothesis)
- Dominance, 203–204
- codominance, 204
 - incomplete dominance, 204
 - mechanisms, 204–206
 - gain-of-function variants, 204–205
 - loss-of-function variants, 204
 - in relation to underlying variants, 167–168
- Dominant progressive hearing loss (DFNA50), 157
- Dominant variants, 209
- “Dominant-versus-recessive” classical genetics, 450
- Dominant–negative variants, 205
- Dose-independent ADRs, 446
- Double helix structure, 55, 55f
- Double heterozygosity, 212–213
- Double homeobox 4 gene (*DUX4*), 163–164
- Double-strand breaks (DSBs), 63–64, 136, 146–147
- Double-stranded DNA, 55f
- Downstream promoter element (DPE), 67–68
- Downstream sequence (DSS), 145
- DPE. *See* Downstream promoter element (DPE)
- DPYD* gene. *See* Dihydropyrimidine dehydrogenase polymorphism (*DPYD* gene)
- Drosophila*, 6–8
- Drug efficacy, GWAS of, 467–468
- anticoagulants, 467–468
 - hepatitis C virus infection, 468
 - mutational landscape of GPCR drug targets, 468
- Drug metabolism, ethnic differences in, 461–463
- Drug responses, 24, 445–446
- genetics, 449–450
- Drug-induced liver injury (DILI), 467

- Drug-induced QT-interval
prolongation, 467
- Drug-metabolizing enzymes (DMEs),
448
- DSBs. *See* Double-strand breaks
(DSBs)
- DSS. *See* Downstream sequence (DSS)
- Duchenne muscular dystrophy
(DMD), 137, 216, 323
- Duplication-inverted triplication-
duplication (DUP-TRP/INV-
DUP), 146
- Duplicational polymorphisms,
128–129
- Duplications, 143. *See also* Segmental
duplications (SDs).
- Duplicons, 138, 139f
- DUX4*. *See* Double homeobox 4 gene
(*DUX4*)
- Dyshomeostasis, 14
- DZ twins. *See* Dizygotic twins (DZ
twins)
- E**
- E3 ubiquitin ligase, 428
- Early PGx examples, 454–463
- abacavir-induced hypersensitivity,
 460
- debrisoquine/sparteine oxidation
 polymorphism, 456–458
- DPYD* gene, 460
- ethnic differences in drug
 metabolism, 461–463
- GST* genes, 459
- N*-acetylation polymorphism,
 454–456
- S*-mephenytoin polymorphism, 459
- TPMT* gene, 458–459
- UGT1A1* gene, 460
- warfarin polymorphisms, 460–461
- Ectopic or temporally altered
messenger RNA expression, 205
- edger* (Bioconductor R package),
339–340
- EDP. *See* Extreme discordant
phenotype (EDP)
- EF1. *See* Elongation factor 1 (EF1)
- Effectors of gene intention, 6
- EFMR. *See* Epilepsy in females with
mental retardation (EFMR)
- EGFR. *See* Epidermal growth factor
receptor (EGFR)
- Electron transport chain (ETC), 269,
306
- Electronic medical records (EMRs),
21
- Electronic Medical Records and
Genomics (eMERGE), 30
- Elongation factor 1 (EF1), 72–73
- EM. *See* Extensive-metabolizer (EM)
- Embryonic development, 214–215
- Embryonic ovary, 244
- Embryonic stem cells (ES cells), 83
- eMERGE. *See* Electronic Medical
Records and Genomics
(eMERGE)
- EMRs. *See* Electronic medical records
(EMRs)
- Encrustation theory, 376
- Endogenous
 functions, 448–449
- gestational environment impact,
 102–103
- influences, 445, 471
- Endoplasmic reticulum membrane
(ER membrane), 73–74
- Enhancer of Zeste, Drosophila, homolog
2 (EZH2)*, 94–95
- Enhancers, 69
- Environmental carcinogens, genetic
resistance to, 435
- Environmental factors, 208, 445,
471–472
- Enzymopathy, 383
- Epidermal growth factor receptor
(EGFR), 25
- Epigenes
 genetic disorders caused by
 mutations in, 93–97, 93t–94t
- human disorders
 due to abnormal readers of
 epigenetic marks, 95–96
- due to mutations in chromatin
 remodelers, 96–97
- due to mutations in erasers of
 epigenetic marks, 95
- due to mutations in writers of
 epigenetic marks, 94–95
- Epigenetic(s), 79, 258, 449, 470–471
- cancer epigenetics, 98–101
- chromatin organization, 81f
- differences within MZ twin pairs,
 399
- effects, 445
- environmental influences on
 epigenome, 101–104
- events, 425–426
- genomic imprinting, 86–93
- interactions between genome and
 epigenome, 104
- marks, 87–88
- mechanisms, 80–83, 219–221
- methods for studying epigenetic
 marks, 97–98
- regulation of X inactivation,
 84–86
- reprogramming, 83–84
- signatures, 104
- Epigenome, 30, 340
- environmental influences on,
 101–104
- endogenous gestational
 environment impact and
 ART, 102–103
- exogenous exposures, 102
- social environmental exposures
 and impact, 103–104
- interactions between genome and,
 104
- projects, 79
- Epigenome-wide association studies
(EWAS), 97–98
- Epigenomics, 105
- Epilepsy in females with mental
retardation (EFMR), 86
- Epileptic disorder Dravet syndrome,
398–399
- Epimutations, 147
- EPM1. *See* Myoclonus epilepsy type 1
(EPM1)
- eQTL analysis. *See* Expression
quantitative trait analysis (eQTL
analysis)
- ER membrane. *See* Endoplasmic
reticulum membrane (ER
membrane)
- “ER stress”, 379
- ERR receptors. *See* Estrogen receptor-
related receptors (ERR receptors)
- Erythroid precursor cells, 205
- ER β . *See* Estrogen receptor β (ER β)
- ES cells. *See* Embryonic stem cells (ES
cells)
- ESS. *See* Exon splicing silencer (ESS)
- Estrogen receptor gene, 425–426
- Estrogen receptor β (ER β), 306

- Estrogen receptor–related receptors (ERR receptors), 307
- ETC. *See* Electron transport chain (ETC)
- Ethnic differences/diversity
in drug metabolism, 461–463
in drug response, 471
of rare disease alleles, 370
- Etiology of twinning, 395–398
- Euchromatin, 60–61
- Evolutionary biology, 415
- Evolutionary patterns, 370–371
- EWAS. *See* Epigenome-wide association studies (EWAS)
- Exogenous exposures, 102
- Exon splicing silencer (ESS), 155
- Exons, 65–71
- Expansion/CNV of Trinucleotide, 134–135
- Expression quantitative trait analysis (eQTL analysis), 343
- Expressivity, 207
- Extensive-metabolizer (EM), 456–457
- Extracellular aging, 419
- Extrahepatic PGx differences, 448–449
- Extreme discordant phenotype (EDP), 470
- EZH2*. *See* Enhancer of Zeste, Drosophila, homolog 2 (*EZH2*)
- F**
- F508del mutation, 367
- F8* gene. *See* Factor VIII gene (*F8* gene)
- Facioscapulohumeral muscular dystrophy (FSHD), 163–164
- FSHD2, 221
- Factor VIII gene (*F8* gene), 161
- Facultative heterochromatin, 60, 254
- False discovery rate (FDR), 342–343
- Familial aggregation, 326–327
- Familial hypercholesterolemia, 370, 383
- “Familial MZ twinning”, 395
- Family health history (FHH), 23, 31
- Farnesyltransferase inhibitors, 431
- FASTQ files, 65, 339–340
- FASTQC tool, 339–340
- Fatal familial insomnia (FFI), 169
- FBN1*. *See* Fibrillin gene (*FBN1*)
- FDA. *See* US Food and Drug Administration (FDA)
- FDR. *See* False discovery rate (FDR)
- Female carriers of X-linked recessive disorders, 215–216
- mechanisms of nonrandom X inactivation, 215–216
- Fetoplacental unit, 400
- Fetus-in-fetu, 391
- FFI. *See* Fatal familial insomnia (FFI)
- FGFR2* gene. *See* Fibroblast growth factor receptor 2 gene (*FGFR2* gene)
- FHH. *See* Family health history (FHH)
- Fibrillin gene (*FBN1*), 161
- Fibroblast growth factor receptor 2 gene (*FGFR2* gene), 148, 168–169
- Fibroblast growth factor receptor 3 gene (*FGFR3* gene), 209
- First-pass elimination, 448
- FISH. *See* Fluorescence in situ hybridization (FISH)
- Fisher’s exact test, 335, 340–341
- Fitness of individual, 368
- Fluorescence in situ hybridization (FISH), 250–251, 251f
- FMRP. *See* Fragile X mental retardation protein (FMRP)
- Follicle-stimulating hormone (FSH), 380–381
- Follicle-stimulating hormone receptor (FSHR), 396
- Fork stalling and template switching (FoSTes), 63–64, 137
- Forkhead box class O (FOXO), 435
- FoSTes. *See* Fork stalling and template switching (FoSTes)
- Founder effect, 366–367
- FOXO. *See* Forkhead box class O (FOXO)
- FPKM. *See* Fragments per kilobase of exon per million fragments mapped (FPKM)
- Fragile sites, 250
- Fragile X (FRAXA), 396
- Fragile X mental retardation protein (FMRP), 163
- Fragile X syndrome, 48
- Fragments per kilobase of exon per million fragments mapped (FPKM), 339–340
- Frameshift variants, 160–161
- FRAXA. *See* Fragile X (FRAXA)
- Frequency of genetic disease, 47–50
- Friedreich ataxia (FXN), 134
- FSH. *See* Follicle-stimulating hormone (FSH)
- FSHD. *See* Facioscapulohumeral muscular dystrophy (FSHD)
- FSHR. *See* Follicle-stimulating hormone receptor (FSHR)
- “Full mutation”, 134–135
- Functional organization of chromosomes, 254–256
- centromere, 254–255
- telomere, 255–256
- “Functional SNPs”, 127
- “Functionome”, 158
- FXN. *See* Friedreich ataxia (FXN)
- G**
- G + D + E. *See* Genistein + daidzien + equol (G + D + E)
- G × E interactions. *See* Gene–environment interactions (G × E interactions)
- G × G interactions. *See* Gene × gene interactions (G ×; G interactions)
- G-protein-coupled receptor genes (*GPCR* genes), 468
- mutational landscape, 468
- Gain-of-function variants
- dominant–negative variants, 205
- ectopic or temporally altered messenger RNA expression, 205
- increased gene dosage, 204–205
- increased protein activity, 205
- new protein functions, 205
- toxic protein alterations, 205
- Galton–Fisher theory, 376–377
- Gas chromatography (GC), 342–343
- Gastrointestinal polyps, 433
- Gastrointestinal tract (GI tract), 448
- GBA*. *See* Glucocerebrosidase (*GBA*)
- gBGC. *See* GC-biased gene conversion (gBGC)
- GC. *See* Gas chromatography (GC)
- GC-biased gene conversion (gBGC), 145
- GDF9*. *See* Growth differentiation factor 9 (*GDF9*)
- GEEs. *See* Generalized estimating equations (GEEs)
- Gene × gene interactions (G × G interactions), 453
- “Gene for” phenylalanine hydroxylase, 9

- Gene Ontology (GO), 342
- Gene(s), 47, 138, 139f
- clusters, 53–54
 - conversion, 144–145
 - dosage, 204–205
 - expression, 67, 339–340
 - families, 62–63
 - in families
 - autosomal dominant inheritance, 206–210
 - autosomal recessive inheritance, 210–213
 - chromosomal disorders, 222–223
 - dominance and recessiveness, 203–206
 - isolated cases, 224
 - nontraditional inheritance, 219–222
 - partial sex linkage, 218–219
 - pedigree construction, 201–202
 - polygenic and multifactorial inheritance, 223–224
 - sex-linked inheritance, 213–218
 - unifactorial inheritance/single gene disorders, 202–203
 - flow, 370
 - mutations in mismatch repair, 165–166
 - structure, 65–71
 - enhancers and *cis*-acting regulatory elements, 69
 - introns and splice junctions, 69–71
 - transcription, 67–68
 - 3'-untranslated sequences and transcriptional termination, 71
 - 5'-untranslated sequences, 69
 - therapy treatment of somatic tissues, 307
 - variants affecting gene expression, 151
- Gene-environment interactions ($G \times E$ interactions), 453
- Generalized estimating equations (GEEs), 401
- Generalized linear mixed models (GLMMs), 401
- Genetic disease
 - caused by mutations in epigenes, 93–97, 93t–94t
 - frequency, 47–50
 - chromosomal disorders, 47–48
 - mitochondrial disorders, 49
 - multifactorial disorders, 49
 - single-gene disorders, 48–49
 - somatic cell genetic disorders, 49–50
 - morbidity and mortality, 50
- Genetic Power Calculator, 334–335
- Genetic(s)
 - anticipation, 134
 - architecture, 449
 - of complex diseases, 168
 - assessment, 201
 - causes
 - of DZ twinning, 395–396
 - of MZ twinning, 395
 - code, 71–72
 - component of trait determination, 326–330
 - familial aggregation, 326–327
 - linkage analysis, 328–329
 - segregation analysis, 327–328
 - transmission disequilibrium test and association analysis, 329–330
 - counseling process, 201
 - determinism, hedge against, 9–10
 - differences within pairs of MZ twins, 398–399
 - of drug response, 449–450
 - etiology, 375–376
 - heterogeneity, 212–213, 233
 - linkage analysis
 - extending parametric linkage analysis, 231–233
 - linkage analysis, 227–231
 - linkage analysis for complex and quantitative traits, 233–235
 - of mtDNA genes, 270–272
 - human mtDNA map, 271f
 - of nDNA mitochondrial genes, 272–273
 - pathogenesis, 376
 - recombination, 228
 - resistance to environmental carcinogens, 435
 - screening, 22–23
 - targeted approaches, 21
 - therapies of mitochondrial diseases, 307–308
- Genistein + daidzein + equol ($G + D + E$), 306
- Genome
 - analysis, 65
 - DNA and RNA synthesis, 57
 - DNA looping and TADs, 64–65
 - DNA replication, 55–56
 - double helix structure, 55
 - interactions with epigenome, 104
 - interindividual variations in human genome, 63–64
 - meiotic recombination, 56–57
 - nuclear human genome, 53–54
 - organization of genomic DNA, 57–65
 - RNA translation into protein, 71–76
 - sequencing studies, 47
 - transcription, 56
 - variation, 371
- Genome-wide association studies (GWAS), 49, 98, 235, 330–334, 453
 - of ADRs, 464–467
 - batch effects, 332
 - discovery phase, 335
 - Marker Allele frequency and HWE filter, 333
 - marker and sample genotyping efficiency or call rate, 332
 - population stratification, 332
 - QC, 331
 - relatedness and Mendelian errors, 332
 - sex inconsistency, 332
 - statistical fine mapping of GWAS data sets, 338
 - study designs, 331
- Genome-wide markers, 227
- Genome-wide PGx studies, 450
- Genomic(s), 22
 - control method, 332
 - disorders, 261
 - DNA organization, 57–65
 - centromeres and telomeres, 61
 - euchromatin and heterochromatin, 60–61
 - nucleosomes and higher order chromatin structure, 58–60
 - repeat content of human genome, 61–62
 - era, 452
 - genomic instability, gene mutations in mismatch repair with, 165–166

- imprinting, 83, 86–93, 104, 219–221
 androgenetic and gynogenetic tumors, 88
 diagnostic testing and recurrence risk, 92–93
 and human developmental disorders, 88–92
 medicine, 125–126
- Genotype, 445
 penetrance values, 231–232
- Genotype frequencies, 360
 factors affecting, 364–366
 assortative mating, 365–366
 consanguinity, 364–365
 inbreeding, 364–365
 stratification, 366
- Genotype–phenotype correlations, 168–170, 213
 digenic inheritance, 169–170
 one disorder caused by variants in more than one gene, 169
 polypheny, 169
 variants in same gene responsible for more than one disorder, 168–169
 variants in same gene rise to distinct dominant and recessive forms, 170
- Germline
 comparison of somatic mutational spectra and, 166
 DMR, 87–88
 epimutations, 147
 mosaicism, 166
 mutations, 428
- GH1* gene. *See* Growth hormone gene (*GH1* gene)
- GI tract. *See* Gastrointestinal tract (GI tract)
- Giemsa (G)
 banding, 244–245
 Giemsa-banded karyotype of male cell, 245f
 staining, 57, 244–245
- GLMMs. *See* Generalized linear mixed models (GLMMs)
- Global collaboration in twin research, 403
- Globin clusters, 53–54
- Glucocerebrosidase (*GBA*), 144
- Glucocorticoid receptor gene (*GR* gene), 103
- Glucocorticoid-suppressible hyperaldosteronism (*GSH*), 165
- Glutathione S-transferase polymorphisms (*GST* genes), 459
- Glycation, 424–425
- GNAS complex locus (*GNAS*), 157
- GO. *See* Gene Ontology (GO)
- Gompertz–Makeham equation, 416
- Gonadal mosaicism, 209–210, 216
- Gonadotropin-releasing hormone, 380–381
- Gonosomal mosaics, 166
- GPCR* genes. *See* G-protein-coupled receptor genes (*GPCR* genes)
- GR gene. *See* Glucocorticoid receptor gene (*GR* gene)
- GRaBD. *See* Gross Rearrangement Breakpoint Database (GRaBD)
- Gross deletions, 137–140
- Gross genomic rearrangement mechanisms, 136–137
- Gross Rearrangement Breakpoint Database (GRaBD), 138–139
- Growth differentiation factor 9 (*GDF9*), 396
- Growth hormone gene (*GH1* gene), 155, 170
- GSH. *See* Glucocorticoid-suppressible hyperaldosteronism (*GSH*)
- GST* genes. *See* Glutathione S-transferase polymorphisms (*GST* genes)
- Guanine (G), 55
- GWAS. *See* Genome-wide association studies (GWAS)
- Gynogenetic tumors, 88
- Gyrate atrophy, 370
- ## H
- H3K9me. *See* Histone 3-lysine 9 monomethylation (H3K9me)
- H3–mK9. *See* Histone H3 at lysine 9 (H3–mK9)
- Hamilton's theory, 419
- HapMap project. *See* Human Haplotype Map project (HapMap project)
- Hardy–Weinberg equilibrium filter (HWE filter), 333, 451
- Hardy–Weinberg law, 359–362
 autosomal locus, 360–361
 factors affecting, 364–370
 allele frequencies, 366–370
 genotype frequencies, 364–366
 two loci, 362–364
 X-linked locus, 362
- Haseman–Elston method, 329
- HATs. *See* Histone acetyl transferases (HATs)
- HBB*. *See* Human β -globin gene (*HBB*)
- HCV infection. *See* Hepatitis C virus infection (HCV infection)
- HDAC inhibitors (HDACi), 100–101
- HDACs. *See* Histone deacetylases (HDACs)
- Helix-loop-helix motif, 68
- Helix-turn-helix motif, 68
- Hemizygous, 203
- Hemoglobinopathies, 367
- Hepatic first-pass kinetics, 448
- Hepatitis C virus infection (HCV infection), 468
 efficacy, 468
- Heptanucleotide CCCCTG, 133–134
- Hereditary neuropathy with liability to pressure palsies (HNPP), 138
- Hereditary nonpolyposis colon cancer (HNPCC), 165–166
- Hereditary persistence of fetal hemoglobin (HPFH), 151–152, 163–164
- Hereditary persistence of α -fetoprotein (HPAFP), 163–164
- Heritability, 326, 400–401
 index, 448
- Heterochromatin, 60–61
- Heterogeneity, 213
- Heterogeneous nuclear RNA (hnRNA), 69–71
- Heteropaternal superfecundation, 391
- Heteroplasmy, 49, 222
- Heterozygosity, 334
- Heterozygote, 206
- Heterozygous state, 204, 210
- HGMD. *See* Human Gene Mutation Database (HGMD)
- HGP. *See* Human Genome Project (HGP)
- HGPS. *See* Hutchinson–Gilford progeria syndrome (HGPS)
- HGVS. *See* Human Genome Variation Society (HGVS)
- HiC assay, 64–65

- Histone
 abnormalities and modifications in
 cancer, 100
 acetylation, 219–220
 code hypothesis, 80
 methylation, 219–220
 modifications, 80–83, 82f
 mapping, 98
 mRNAs, 71
- Histone 3-lysine 9 monomethylation
 (H3K9me), 60
- Histone acetyl transferases (HATs),
 307
- Histone deacetylases (HDACs), 80,
 307
- Histone H3 at lysine 9 (H3–mK9), 258
- HIV. *See* Human immunodeficiency
 virus (HIV)
- HLA. *See* Human leukocyte antigen
 (HLA)
- HLA-B* 57:01 allele, 467
- 5hmC. *See* 5-Hydroxymethylcytosine
 (5hmC)
- HMP. *See* Human Microbiome Project
 (HMP)
- HNPCC. *See* Hereditary nonpolyposis
 colon cancer (HNPCC)
- HNPP. *See* Hereditary neuropathy
 with liability to pressure palsies
 (HNPP)
- hnRNA. *See* Heterogeneous nuclear
 RNA (hnRNA)
- Hole migration, 132
- Homeorhesis, 377
- Homeostasis, 5–6, 376–377
- Homoplasmy, 49
- Homozygous state, 210
- Hormonal differences in PK and PD
 traits, 471
- Hormones, 208
- “Housekeeping genes”, 67
 expression, 75–76
- “How” questions, 5–11
 qualities of unit step of homeostasis,
 6–11
- HOX* clusters, 53–54
- HPA axis. *See* Hypothalamic–
 pituitary–adrenal axis (HPA axis)
- HPAFP. *See* Hereditary persistence of
 α -fetoprotein (HPAFP)
- HPFH. *See* Hereditary persistence of
 fetal hemoglobin (HPFH)
- Human allelic variants homologous
 to pro-longevity genes in model
 organisms, 435
- Human chromosomes, 53–54, 54t, 66t,
 244–254
 chromosome banding, 244–249
 reveals genome sequence
 organization, 249–250
 identification, 244
 molecular cytogenetics, 250–254
- Human *COMT* gene, 132
- Human developmental disorders,
 88–92
- Human disease genes, functional
 characteristics of, 148–149
- Human disorders
 due to abnormal readers of
 epigenetic marks, 95–96
 due to mutations
 in chromatin remodelers, 96–97
 in erasers of epigenetic marks, 95
 in writers of epigenetic marks,
 94–95
 position effect in, 164–165
- Human DNA polymorphisms, 129
- Human Gene Mutation Database
 (HGMD), 126–128, 130, 130f,
 151–152
- Human gene variants, 126
- Human genetic diseases, X inactivation
 relevant to, 85–86
- Human genome, 60, 125, 237
 repeat content of, 61–62
- Human Genome Project (HGP), 3,
 21–22, 53, 227, 233–234
 goal, 10–11
- Human Genome Variation Society
 (HGVS), 149
- Human genomic variants, 125–126
 consequences of mutations, 149–168
 disease-causing variants, 130–149
 general principles of genotype–
 phenotype correlations,
 168–170
 molecular mechanisms of variants
 causing human inherited
 disease
 “neutral variation”/DNA
 polymorphisms, 126–129
 nonsense SNPs, 129–130
 nucleotide diversity, 128f
 mutation study, 170–171
- Human Haplotype Map project
 (HapMap project), 63, 127, 168
- Human immunodeficiency virus
 (HIV), 460
- Human leukocyte antigen (HLA), 390,
 393
 Loci, 460
- Human Microbiome Project (HMP),
 30
- Human mutation rates, 166–167
- Human origins, mtDNA and, 273–274
- Human β -globin gene (*HBB*), 161
- Humans, typical twinning in, 388–389
- Hunter syndrome, 75
- Huntington disease, 62, 323, 420
- Hurler syndrome, 75
- Hutchinson–Gilford progeria
 syndrome (HGPS), 429–431, 431f
- HWE filter. *See* Hardy–Weinberg
 equilibrium filter (HWE filter)
- Hydatidiform moles, 88
- Hydrocodone, 458
- Hydroxymethylation, 83–84
- 5-Hydroxymethylcytosine (5hmC), 83
- Hypersensitivity, 467
- Hypertension treatment, GWAS
 dilemma of, 469
- Hypothalamic–pituitary–adrenal axis
 (HPA axis), 103–104
- I**
- IBD. *See* Identity-by-descent (IBD)
- IC. *See* Imprinting center (IC)
- ICM. *See* Inner cell mass (ICM)
- ICOMBO. *See* International Council
 for Multiple Birth Organisations
 (ICOMBO)
- Identity-by-descent (IBD), 327, 329
- IDH. *See* Isocitrate dehydrogenase
 (IDH)
- “Idiogram”, 248
- Idiosyncratic dose-independent drug
 reactions, 446
- IFN. *See* Interferon (IFN)
- IGF1. *See* Insulin-like growth factor
 (IGF1)
- Igf2r* gene, 86–87
- IGLL1*. *See* Immunoglobulin λ -like
 polypeptide 1 (*IGLL1*)
- IGNITE. *See* Implementing GeNomics
 In PracTice (IGNITE)
- IHH*. *See* Indian hedgehog (*IHH*)

- IL. *See* Interleukin (IL)
- IL28B*. *See* Interleukin-28B gene (*IL28B*)
- Illumina Infinium methylation arrays, 340
- Illumina system, 65
- Immunoglobulin λ -like polypeptide 1 (*IGLL1*), 144
- Immunoglobulins, 62–63
- Implementing GeNomics In PracTice (IGNITE), 30
- Imprinted differentially methylated regions. *See* Epigenetic marks
- Imprinting center (IC), 87–88
- Imprinting disorders, diagnostic testing and recurrence risk in, 92–93
- Imprinting dysregulation, 88
- Imputation process, 333–334
- Inappropriate gene expression, variants to, 163–164
- Inborn errors, 15–18
- Inborn-errors-of-metabolism, 450–451
- Inbred mouse strains, 228–229
- Inbreeding, 364–365
- Incidentalome, 453
- Incomplete dominance, 204
- Incomplete penetrance, 467
- “Indel hotspot” GTAAAGT, 133–134
- Indels. *See* Insertion–Deletions (Indels)
- Indian hedgehog* (*IHH*), 64–65
- Individual drug response, 445
- Individualized drug therapy, 445
- Infinitesimal model, 454
- Inflation factor, 335–336
- Inner cell mass (ICM), 84
- Inosine triphosphatase (*ITPA*), 459
- Insertion–Deletions (Indels), 145
- Insudation theory, 376
- Insulators, 80
- Insulin-like growth factor (IGF1), 422–423
- Insulin-like growth factor 2 (*IGF2*), 89–90
- Interferon (IFN), 150
- Interindividual variations/variability in drug response, 445–446
- Interindividual variations in human genome, 63–64
- Interleukin (IL), 305
- Interleukin-28B gene (*IL28B*), 468
- Internal ribosome entry sites (IRES), 72
- International Council for Multiple Birth Organisations (ICOMBO), 395
- International HapMap project, 330–331
- International Society of Twin Studies (ISTS), 395
- Interspersed repeats. *See* Transposon-derived repeats
- Intramolecular disulfide bond formation, 75
- Intraspecific variations, 423–424
- Intron-splicing processing element (ISPE), 155
- Introns, 65–66
- Inversions, 142–143, 142f
- IRES. *See* Internal ribosome entry sites (IRES)
- Iron-sulfur (FeS), 269
- Isochores, 60
- Isochromosome, 262
- Isocitrate dehydrogenase (IDH), 268–269
- Isoniazid acetylation polymorphism, 454
- Isoniazid *N*-acetyltransferase variability, 454
- ISPE. *See* Intron-splicing processing element (ISPE)
- ISTS. *See* International Society of Twin Studies (ISTS)
- ITPA*. *See* Inosine triphosphatase (*ITPA*)
- Ivacaftor, 25
- K**
- Kabuki syndrome, 95
- Kearns-Sayre syndrome (KSS), 283–284, 428
- Keytruda, 25
- Kinetochore, 254–255
- Kozak consensus sequence GCCA/GCCCAUGG, 69, 159
- L**
- L1-retrotransposition, 141
- L1CAM gene, 168–169
- Lactase gene (*LCT*), 162–163
- LADs. *See* Lamina-associated domains (LADs)
- LAMA3* gene, 161–162
- Lamina-associated domains (LADs), 99
- Large insertion of repetitive, 141–142
- Large organized chromatin K9 modifications (LOCKS), 99
- Large retrotranspositional insertions, 140–141
- Late-life disorders, 421
- LC. *See* Liquid chromatography (LC)
- LCR. *See* Locus control region (LCR)
- LCT*. *See* Lactase gene (*LCT*)
- LD. *See* Linkage disequilibrium (LD)
- LDL. *See* Low-density lipoprotein (LDL)
- LDL-C. *See* Low-density-lipoprotein-cholesterol (LDL-C)
- LDLR*. *See* Low-density lipoprotein receptor (*LDLR*)
- Leber hereditary optic neuropathy (LHON), 221–222, 274, 428 mutations, 274
- Leigh syndrome, 284
- Lens crystallines, 419
- Leri–Weill dyschondrosteosis, 218–219
- Leucine zipper motif, 68
- LGMD1B. *See* Limb-girdle muscular dystrophy type 1B (LGMD1B)
- LHON. *See* Leber hereditary optic neuropathy (LHON)
- Liability distribution, 324–325, 325f
- Likelihoods, 229–230
- Limb-girdle muscular dystrophy type 1B (LGMD1B), 431
- LINE. *See* Long interspersed nuclear element (LINE)
- LINE retrotransposition, 140–141, 140f
- Linkage analysis, 227–231, 232t, 328–329
- biological basis, 228
- for complex and quantitative traits, 233–235
- linkage analysis of quantitative traits, 234–235
- model-free linkage analysis, 234
- future directions, 235–236
- model-based analysis, 328
- model-free linkage analysis, 329
- parametric linkage analysis, 229–231
- simplified, 228–229
- Linkage analysis of quantitative traits, 234–235

- Linkage disequilibrium (LD), 228, 324–325, 362–363, 467
- LINKAGE software, 233
- Linkage studies, 453
- “Linker” DNA, 237–239
- lin Var*, 126
- Lipid
- alterations in, 428–429
 - metabolism disorders in segmental progeroid phenotypes, 433
- Lipid mediators (LMs), 448, 472
- Lipofuscins, 419
- Liquid chromatography (LC), 341–342
- LMNA* mutations, 431
- LMs. *See* Lipid mediators (LMs)
- lncRNAs. *See* Long noncoding RNAs (lncRNAs)
- Location
- of repeat expansion, 135f
 - scores, 233
- LOCKS. *See* Large organized chromatin K9 modifications (LOCKS)
- Locus control region (LCR), 64–65, 162
- Locus heterogeneity, 169, 212–213
- LocusExplorer, 336
- LocusZoom, 336, 336f
- LOD score, 229–230
- LOF. *See* Loss-of-function (LOF)
- Long interspersed nuclear element (LINE), 60
- Long noncoding RNAs (lncRNAs), 56, 80–83
- expression in cancer, 100
- Long terminal repeat (LTR), 61–62
- Loss-of-function (LOF), 435
- mutations, 304–305
 - phenotypes, 383
 - variants, 204
- Low copy repeats (LCRs). *See* Segmental duplications (SDs)
- Low-copy repeats, 261
- Low-density lipoprotein (LDL), 25, 434
- Low-density lipoprotein receptor (*LDLR*), 138
- Low-density-lipoprotein-cholesterol (LDL-C), 469
- LTA* gene. *See* Lymphotoxin- α gene (*LTA* gene)
- LTR. *See* Long terminal repeat (LTR)
- LUC7L* gene, 165
- Lymphotoxin- α gene (*LTA* gene), 453
- Lyonization, 214–215
- M**
- Macrophage receptors, 424–425
- MAF. *See* Minor allele frequency (MAF)
- Major depressive disorder (MDD), 470
- Major histocompatibility complex (MHC), 460
- Mandibular hypoplasia, deafness, progeroid features, lipodystrophy syndrome (MDPL syndrome), 431–432
- Manhattan plots, 335–336
- MANTRA method, 338–339
- MAPT* gene. *See* Microtubule-associated protein tau gene (*MAPT* gene)
- Marfan syndrome, 161, 323
- Marker
- allele frequency, 333
 - chromosome, 262
 - genotypes, 228
 - and sample genotyping efficiency, 332
- Markov chain Monte Carlo methods (MCMC methods), 401
- MARs. *See* Matrix attachment regions (MARs)
- Mass spectrometry (MS), 341–342
- Massively parallel sequencing, 98
- Matrix attachment regions (MARs), 237–239
- MatrixEQTL, 343–344
- Maximum likelihood estimation, 229–230
- 5mC. *See* 5-Methylcytosine (5mC)
- MC-Seq. *See* Methyl-capture sequencing (MC-Seq)
- MCMC methods. *See* Markov chain Monte Carlo methods (MCMC methods)
- MCU. *See* Mitochondria Ca^{2+} uniporter (MCU)
- MDD. *See* Major depressive disorder (MDD)
- MDPL syndrome. *See* Mandibular hypoplasia, deafness, progeroid features, lipodystrophy syndrome (MDPL syndrome)
- MDS. *See* Myelodysplastic syndrome (MDS)
- Medicine, 1
- disease definition, 4–5
 - “how” questions, 5–11
 - prevention and treatment, 15–18
 - principles of disease, 2–4
 - “why” questions, 12–15
- MedSeq project, 24
- Meiotic/meiosis, 239, 241–243, 242f
- metaphase, 243
 - recombination, 56–57
- MELAS. *See* Mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS)
- Mendelian diseases, 334
- Mendelian errors, relatedness and, 332
- Mendelian genetics, 241
- Mendelian inheritance, 201
- resolution of multifactorial traits with, 452
 - school, 452
- Mendelian Inheritance in Man*, 126
- Mendelian phenotypes, 48
- Mendelian traits, 126–127, 228, 231
- S-Mephenytoin polymorphism, 459
- 6-Mercaptopurine (6MP), 458
- MERRF. *See* Myoclonic epilepsy and ragged red fiber (MERRF)
- Messenger RNA (mRNA), 56, 65–71
- splicing mutants, 152–156
 - splicing of, 70f
 - translation into protein, 73f
- Met358Arg, 150
- Metabolic and Molecular Basis of Inherited Disease (MMBID), 1–2
- Metabolic process, 381
- Metabolic therapies of mitochondrial diseases, 306–307
- Metabolite, 446–447
- data, 342–343
- Metabolome, 342–343
- Metabolomics, 342–343
- Metal-catalyzed oxidation systems, 424
- Metaphase, 239
- chromosome scaffold, 237–239
 - fluorescence in situ hybridization analysis, 252f
- Metformin Genetics Consortium (MetGen Consortium), 469
- Methyaminomethyl-2-thiouridylate-methyltransferase (TRMU), 284–285

- Methyl-capture sequencing (MC-Seq), 97
- Methylation data, 340
- from arrays, 340–341
- Methylation quantitative trait loci (mQTL), 104
- Methylation-sensitive multiplex ligation-dependent probe amplification (MS-MLPA), 92
- 5-Methylcytosine (5mC), 80–81, 131, 131f
- Methylguanosine (mG), 71
- Methyltetrahydrofolate reductase (MTHFR), 102
- MeTree, 31
- mG. *See* Methylguanosine (mG)
- MHC. *See* Major histocompatibility complex (MHC)
- Microarray-based assays/methods, 98, 340
- Microbiome, 30
- differences, 445, 472–473
- Microcephalic osteodysplastic primordial dwarfism type I (MOPD I), 157
- Microdeletions, 132–134
- syndromes, 251–252
- “Microhomology-dependent BIR” model, 137
- Microhomology-mediated break-induced replication (MMBIR), 137
- Microhomology-mediated end joining (MMEJ), 137
- Microinsertions, 132–134
- microRNAs (miRNAs), 56, 82–83, 156–157
- microRNA-binding sites, variants in, 156–157
- Microsatellites, 62, 63f, 137
- Microtubule-associated protein tau gene (*MAPT* gene), 155
- Microtubule-binding repeats, 155
- Migration, 370
- Minor allele frequency (MAF), 333, 451
- miRNAs. *See* microRNAs (miRNAs)
- Mirror-image twins, 390
- Missense variant, 150–151, 162
- Missing heritability, 337–338, 454
- Missing lesions, 158
- Mitochondria Ca^{2+} uniporter (MCU), 270
- Mitochondria(l), 270
- aldehyde dehydrogenase-2 deficiency, 461
- carrier family. *See* Solute carrier family 25 (SLC25)
- chromosome, 49
- disorders, 49, 268
- dysfunction, 268, 285–303, 305
- inheritance, 221–222
- inner membrane, 268–269
- ribosomes, 267
- Mitochondrial biology-mitochondrial medicine
- genetic therapies of mitochondrial diseases, 307–308
- metabolic therapies of mitochondrial diseases, 306–307
- mitochondrial biochemistry, 268–270
- mitochondrial diseases diagnosis, 305–306
- mitochondrial etiology of complex diseases, 285–305
- mitochondrial genetics, 270–273
- genetics of mtDNA genes, 270–272
- genetics of nDNA mitochondrial genes, 272–273
- nDNA coded mitochondrial diseases, 284–285
- Mitochondrial DNA (mtDNA), 49, 221, 267, 370–371, 427–428
- coded mitochondrial diseases, 274–284
- ancient adaptive mtDNA variants, 283
- developmental and somatic mtDNA mutations, 283–284
- maternally inherited mtDNA diseases, 274–283
- disease phenotypes range in MITOMAP and MITOMASTER, 284
- genetics, 270–272
- and human origins, 273–274
- mutations, 275t–276t
- polypeptide genes, 267
- Mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS), 268
- Mitochondrial permeability transition pore (mtPTP), 270
- Mitochondrion, 268–269
- MITOMAP, mtDNA disease phenotype range in, 284, 285f
- MITOMASTER, mtDNA disease phenotype range in, 284
- Mitosis, 239–241, 240f
- Mixed effects models. *See* Generalized linear mixed models (GLMMs)
- MLID. *See* Multilocus imprinting defect (MLID)
- MMBID. *See* Metabolic and Molecular Basis of Inherited Disease (MMBID)
- MMBIR. *See* Microhomology-mediated break-induced replication (MMBIR)
- MMEJ. *See* Microhomology-mediated end joining (MMEJ)
- Mn superoxide dismutase (MnSOD), 306
- Model-based analysis, 328
- Model-free linkage analysis, 234, 329
- Molecular cytogenetics, 250–254
- Molecular misreading, 147, 427
- Molecular pathogenetics, 381–384
- Monochorionic diamniotic twins, 392
- Monochorionic monoamniotic twins, 392
- Monogenic disease, 323
- Monogenic traits, 450–452
- Monopaternal superfecundation, 391
- Monozygotic twinning (MZ twinning), 215–216, 326, 387, 395, 398–400
- epigenetic differences within, 399
- genetic causes, 395
- genetic differences within, 398–399
- incidence, 392
- nonshared environment and chronic disease, 400
- sex ratio in, 392
- MOPD I. *See* Microcephalic osteodysplastic primordial dwarfism type I (MOPD I)
- Mosaicism, 166, 260
- gonadal, 209–210
- somatic, 209–210
- mQTL. *See* Methylation quantitative trait loci (mQTL)
- 6MP. *See* 6-Mercaptopurine (6MP)
- mRNA. *See* Messenger RNA (mRNA)

- MS. *See* Mass spectrometry (MS)
- MS-MLPA. *See* Methylation-sensitive multiplex ligation-dependent probe amplification (MS-MLPA)
- β -MSH. *See* β -melanocyte-stimulating hormone (β -MSH)
- mtDNA. *See* Mitochondrial DNA (mtDNA)
- MTHFR. *See* Methyltetrahydrofolate reductase (MTHFR)
- mtPTP. *See* Mitochondrial permeability transition pore (mtPTP)
- Mucopolysaccharidoses, 75
- Multi-tilt scanning electron microscopy tomography, 58–60
- Multiallelic single tandem repeats, 227
- Multifactorial disorders, 49
- Multifactorial inheritance, 223–224
analysis of rare variants using new technologies, 337–338
association methods/statistical analysis, 334–337
data types and analysis methods
epigenome and methylation data, 340
gene expression and RNA-Seq data, 339–340
metabolome and metabolite data, 342–343
methylation data from arrays, 340–341
proteome and protein data, 341–342
determining genetic component of trait, 326–330
expected phenotype distribution for trait, 325f
frequency distribution of genotypic values, 324t
future directions/integration, 343–344
genotypic values of two-locus genotypes, 324t
GWAS, 331–333
imputation, 333–334
international HapMap project, 330–331
phenotype distribution for trait with single causal locus, 324f
statistical fine mapping of GWAS data sets, 338
transethnic meta-analysis, 338–339
- Multifactorial traits, 452
- Multilocus imprinting defect (MLID), 90
- Multiple genetic disorders in family, 222
- Multiple mechanisms, 422
- Multiple simultaneous mutations, 146–147
- Multiple-compartment models, 381
- Multiple-hit processes, 381
- Multipoint parametric linkage analysis, 233
- Multivariate methods, 342–343
- Mus musculus*, 416
- Mutalyzer, 149
- Mutant strain, 384
- Mutation, 283, 367–368
accumulation, 420
events, 426–427
frequency
within genes, 148
within human populations, 148
in gene evolution, 149
nomenclature, 149
study, 170–171
- Mutation consequences
abnormal proteins due to different genes fusion, 165
cap site variants, 158–159
cellular consequences of
trinucleotide repeat expansions, 163
comparison of germline and somatic mutational spectra, 166
dominance and recessiveness in relation to underlying variants, 167–168
exon skipping due to nonsense, missense, and silent mutations, 156f
frameshift variants, 160–161
gene mutations in mismatch repair with genomic instability, 165–166
genetic architecture of complex diseases, 168
human mutation rates, 166–167
mosaicism, 166
mRNA splicing mutants, 152–156
nonsense variants, 161–162
position effect
by antisense RNA, 165
in human disorders, 164–165
promoter (transcription regulatory) variants, 151–152
RNA cleavage-polyadenylation mutants, 156
of splice junctions recorded in HGMD, 154f
splicing abnormalities in introns of human genes, 153f
termination codon (“nonstop”) variants, 160
translational initiation codon variants, 159–160
unstable protein mutants, 162
variants
affecting amino acid sequence of predicted protein, 149–151
affecting gene expression, 151
to inappropriate gene expression, 163–164
in microRNA-Binding Sites, 156–157
in noncoding regions of functional significance, 157–158
in non-protein-coding genes, 157
in 3' regulatory regions, 159
in remote gene regulatory elements, 162–163
in 5' UTRs, 159
- Mutational heterogeneity, 213
- Myelodysplastic syndrome (MDS), 83
- Myocardial infarction, 234–235, 453
- Myoclonic epilepsy and ragged red fiber (MERRF), 268
- Myoclonus epilepsy type 1 (EPM1), 135
- Myotonic dystrophy type 2 (*ZNF9*), 134
- MZ twinning. *See* Monozygotic twinning (MZ twinning)
- N**
- N-linked glycosylation, 75
- NAD-dependent histone deacetylation, 426
- NADH. *See* Nicotinamide dinucleotide (NADH)
- NAHR. *See* Nonallelic homologous recombination (NAHR)
- NAT2 gene, 454–456
- Nature of genomic variants, 130

- NCF1*. See Neutrophil cytosolic factor p47-*phox* (*NCF1*)
- NDD. See Neurodevelopmental disorder (NDD)
- nDNA. See Nuclear DNA (nDNA)
- Negative binomial model, 339–340
- neo-TADs. See New TADs (neo-TADs)
- Neocentromeres, 255
- Neodarwinism, 9
- NER. See Nucleotide excision repair (NER)
- Neurodevelopmental disorder (NDD), 86
- Neurofibromatosis type 2, 210
- Neuropharmacogenomics, 472
- Neuropsychiatric disease, 305
- Neurospora*, 8
- Neutral variation, 126–129
- Neutrophil cytosolic factor p47-*phox* (*NCF1*), 144
- New TADs (neo-TADs), 158
- Newborn Sequencing in Genomic Medicine and Public Health (NSIGHT), 30
- Next-generation sequencing (NGS), 65, 97, 213
- NextGen. See Next-generation sequencing (NGS)
- NGS. See Next-generation sequencing (NGS)
- NHEJ. See Nonhomologous end joining (NHEJ)
- Nicotinamide dinucleotide (NADH), 268–269
- NIPT. See Noninvasive prenatal testing (NIPT)
- Nitric oxide synthase (NO synthase), 268–269
- Nitroglycerin, 461–462
- NMD. See Nonsense-mediated mRNA decay (NMD)
- NMR. See Nuclear magnetic resonance (NMR)
- NO synthase. See Nitric oxide synthase (NO synthase)
- NOISeg*, 339–340
- Nonallelic heterogeneity, 169
- Nonallelic homologous recombination (NAHR), 63–64, 136
- Noncoding DNA, 371
- mRNA sequences, 65–66
- regions of functional significance, variants in, 157–158
- Nonhomologous (illegitimate) recombination, 138–139
- Nonhomologous end joining (NHEJ), 63–64, 136, 146–147
- Noninvasive prenatal testing (NIPT), 22
- Nonparametric linkage analysis, 328
- Non-protein-coding genes, variants in, 157
- Nonrandom X inactivation mechanisms, 215–216
- Nonsense mutations, 72
- SNPs, 129–130
- variants, 161–162
- Nonsense-mediated decay, 72
- Nonsense-mediated mRNA decay (NMD), 129, 161
- Nonshared environment and chronic disease, 400
- Non-small cell lung cancer (NSCLC), 25, 100
- Nontraditional inheritance, 219–222
- digenic inheritance, 221
- genomic imprinting and epigenetic mechanisms, 219–221
- mitochondrial inheritance, 221–222
- multiple genetic disorders in family, 222
- NORs. See Nucleolar organizing regions (NORs)
- NSCLC. See Non-small cell lung cancer (NSCLC)
- NSD1* gene, 94
- NSIGHT. See Newborn Sequencing in Genomic Medicine and Public Health (NSIGHT)
- Nuclear DNA (nDNA), 267
- coded mitochondrial diseases, 284–285
- nonstructural nuclear genes, 289t–296t
- structural nuclear genes, 286t–287t
- epigenetic events, 425–426
- gene mutations, 268
- mitochondrial genes genetics, 272–273
- molecular misreading, 427
- mutational events, 426–427
- Nuclear human genome, 53–54
- human chromosomes, 54t
- Nuclear magnetic resonance (NMR), 342–343
- Nucleolar organizing regions (NORs), 244
- Nucleosome, 80, 237–239
- eukaryotic gene structure and the pathway of gene expression, 59f
- and higher order chromatin structure, 58–60
- Nucleotide excision repair (NER), 432
- Nucleotide substitutions, 130–132
- Nudix hydrolase-15 (*NUDT15*), 459
- Null alleles, 459
- Numerical chromosome abnormalities, 260–261
- O**
- O-linked glycosylation, 75
- OCM cycle. See One-carbon metabolism cycle (OCM cycle)
- Odds ratios (ORs), 305, 460
- Olaparib, 25
- Oligoarray comparative genomic hybridization analysis, 253f
- Oligogenic disorders, 323
- omics technologies, 30
- OMIM. See Online Mendelian Inheritance in Man (OMIM)
- Oncology, 24
- One-carbon metabolism cycle (OCM cycle), 102
- Online Mendelian Inheritance in Man (OMIM), 168
- Oogenesis, 243–244
- Open reading frames (ORFs), 140–141
- Ordered subset analysis approach, 235
- ORFs. See Open reading frames (ORFs)
- ORs. See Odds ratios (ORs)
- Orthogenetics, 376
- Osteogenesis imperfecta, 323
- Osteonecrosis of jaw, 467
- Outer mitochondrial membrane, 268
- Ovarian teratomas, 88, 259
- Overdominance, 368
- Ovulation, 396
- Ovum, 237
- Oxidative phosphorylation (OXPHOS), 267, 269f

- OXPHOS. *See* Oxidative phosphorylation (OXPHOS)
- Oxycodone, 458
- Oxygen toxicity, 427
- P**
- p.R153P variant, 432–433
- p53 protein, 384
- PAH*. *See* Phenylalanine hydroxylase (*PAH*)
- PAINTOR, 338–339
- Panomics, 30
- Paracentric inversion, 262
- Paragangliomas, 220, 220f
- Parametric linkage analysis, 229–233
- designing and conducting, 231
- genetic heterogeneity, 233
- incomplete penetrance and phenocopies, 232
- likelihoods, maximum likelihood estimation, and statistical significance, 229–230
- LOD scores, 230
- modeling traits with penetrance functions, 230–231
- multipoint parametric linkage analysis, 233
- Parent drug, 446–447
- Parental imprinting. *See* Genomic(s)—imprinting
- Parkinson disease (PD), 303, 428
- PARP inhibitor. *See* Poly(ADP)ribose polymerase inhibitor (PARP inhibitor)
- Partial karyotype of acrocentric chromosomes, 246f
- Partial least squares (PLS), 342–343
- Partial sex linkage, 218–219
- Paternal lineage, 202
- Paternal uniparental disomy (UPD), 88
- Pathogenesis, 376
- Pathogenetics, 376
- development of anatomic structures, 377–379
- molecular pathogenetics, 381–384
- pathogenetic tree for Mendelian condition, 382f
- pathogenetics of refined traits, 379
- pathways and multiple-stage processes, 379–381
- branching pathways, 380
- multiple-compartment models, 381
- multiple-hit processes, 381
- pathways with feedback, 380–381
- simple pathways, 379–380
- scope of abnormal phenotypes, 376–377
- Pathogenic variants, 125, 165
- PAX3, 204
- PCA. *See* Principal components analysis (PCA)
- PCBs. *See* Polychlorinated biphenyls (PCBs)
- PCMs. *See* Potentially compensated mutations (PCMs)
- PCR. *See* Polymerase chain reaction (PCR)
- PCs. *See* Principal components (PCs)
- PCSK9. *See* Proprotein convertase subtilisin/kexin type 9 (PCSK9)
- PD. *See* Parkinson disease (PD); Pharmacodynamics (PD)
- PDH. *See* Pyruvate dehydrogenase (PDH)
- PDI. *See* Protein disulfide isomerase (PDI)
- Pearson's χ^2 tests, 335
- Pediatrics, 22–23
- Pedigrees, 227
- construction, 201–202, 202f–203f
- Penetrance, 206–207
- functions, 230–231
- incomplete, 232
- Pentanucleotide repeat (ATTCT)_n, 135
- Pentosidine, 424–425
- PEO. *See* Progressive external ophthalmoplegia (PEO)
- Pep3D* tool, 341–342
- PeptideProphet* tool, 341–342
- Percutaneous transvenous endomyocardial biopsy, 24
- Pericentric inversion, 262
- Peripheral myelin protein 22 gene (*PMP22* gene), 143, 204–205
- Peromyscus leucopus*, 416
- Peroxisome proliferator-activated receptor gamma (PPARG), 396
- Peroxisome proliferator-activated receptors (PPARs), 307
- Personalized medicine, 445
- PGA, 334–335
- PGC-1 α . *See* PPAR γ -coactivator-1 (PGC-1 α)
- PGD. *See* Preimplantation genetic diagnosis (PGD)
- PGx. *See* Pharmacogenetics (PGx)
- Ph chromosome. *See* Philadelphia chromosome (Ph chromosome)
- Pharmacodynamics (PD), 447–448
- Pharmacogenetics (PGx), 445
- differences, 447
- FDA recommendations for PGx genotyping, 473
- GWAS of unsuccessful PGx examples, 469–470
- dilemma of hypertension treatment, 469
- dilemma of psychotropic drugs, 469–470
- history of genetics relevant to, 450–454
- beginning of genomics era, 452
- GWAS, 453
- monogenic traits, 450–452
- resolution of multifactorial traits with Mendelian inheritance, 452
- SNPs and SNVs, 453
- variance explained vs. “missing heritability”, 454
- multifactorial traits, 450
- profile, 445
- testing, 24
- traits, 449
- Pharmacogenomics, 445, 463–470
- ADRs, 446, 446f
- indistinguishable from complex diseases, 463–464
- early PGx examples, 454–463
- FDA recommendations for PGx genotyping, 473
- fundamental aspects of clinical pharmacology, 446–450
- genome-wide association studies of ADRs, 464–467
- complex PGx traits, 468–469
- drug efficacy, 467–468
- unsuccessful PGx examples, 469–470
- history of genetics relevant to PGx, 450–454
- response to drugs other than genotype of patient, 470–473

- endogenous influences, 471
- environmental factors, 471–472
- epigenetics, 470–471
- microbiome differences, 472–473
- types of drug responses, 445–446
- Pharmacokinetics (PK), 447–448
- Phenocopies, 232
- Phenotype, 204, 222
- Phenotypic effects, 248
- Phenylalanine hydroxylase (*PAH*), 162, 452
- Phenylketonuria (PKU), 359, 380, 452
- Philadelphia chromosome (Ph chromosome), 165
- Phosphomannomutase (*PMM2*), 144
- Phytohemagglutinin, 244
- PINK1* (serine/threonine kinase), 428
- PISRT1* gene, 162
- PK. *See* Pharmacokinetics (PK)
- PKD1*. *See* Polycystic kidney disease (*PKD1*)
- PKU. *See* Phenylketonuria (PKU)
- Placentation, 391–392
- Plasma clearance of drug, 448
- Plasticity, 424
- Pleiotropy, 207
- PLS. *See* Partial least squares (PLS)
- PM. *See* Poor-metabolizer (PM)
- PMM2*. *See* Phosphomannomutase (*PMM2*)
- PMP22* gene. *See* Peripheral myelin protein 22 gene (*PMP22* gene)
- PNT. *See* Pronuclear transfer (PNT)
- Poisson distribution, 339–340
- Polar body twins, 390
- POLD1* gene, 432
- POLG*. *See* Polymerase γ (*POLG*)
- Poly(ADP)ribose polymerase inhibitor (PARP inhibitor), 25
- Polychlorinated biphenyls (PCBs), 462
- Polycomb repressive complex-2 (PRC2), 82–83
- Polycystic kidney disease (*PKD1*), 144
- Polygenic basis, 422
- Polygenic inheritance, 223–224, 323
- Polygenic model, 323–324
- Polymerase chain reaction (PCR), 24, 128–129
- PCR-based method, 427
- Polymerase γ (*POLG*), 285
- Polymorphisms, 152
- “Polypheny”, 169
- “Polyploidy”, 260–261
- POMC* gene. *See* Pro-opiomelanocortin gene (*POMC* gene)
- Poor-metabolizer (PM), 456–457
- Population genetics, 359. *See also* Epigenetics.
- applications in, 370–372
- causal variant identification for common diseases, 371–372
- ethnic diversity of rare disease alleles, 370
- evolutionary patterns, 370–371
- genome variation, 371
- Hardy–Weinberg law, 359–362
- Population stratification, 332, 470–471
- Postanalysis quality control, 335–336
- Posttranslational modification, 75
- Potentially compensated mutations (PCMs), 149–150
- Power and sample size calculations, 334–335
- PPAR γ -coactivator-1 (PGC-1 α), 307
- PPARG. *See* Peroxisome proliferator-activated receptor gamma (PPARG)
- PPARs. *See* Peroxisome proliferator-activated receptors (PPARs)
- PPMGG. *See* Principles and Practice of Medical Genetics and Genomics (PPMGG)
- Prader–Willi syndrome (PWS), 90, 220, 259–260
- PRC2. *See* Polycomb repressive complex-2 (PRC2)
- Precision drug development, 25–29
- approved oncology drugs, 27t–29t
- genes, evidence level, and type of information, 26t
- Precision medicine, 21–22
- applications across lifespan and clinical specialties, 22–25
- practice, 30–32
- precision drug development, 25–29
- research, 29–30
- Preconceptual and Prenatal Screening, 22
- Predictive value, 454, 469
- Predominantly oligogenic traits, 450
- Pregnancy, genetic disease and, 50
- Preimplantation genetic diagnosis (PGD), 22
- Premeiotic mutations, 167
- Premutation, 134–135
- Prenatal diagnosis, 22
- Principal components (PCs), 332
- Principal components analysis (PCA), 332
- Principles and Practice of Medical Genetics and Genomics (PPMGG), 1
- Prion protein polymorphism (*PRNP* polymorphism), 148
- PRNP* polymorphism. *See* Prion protein polymorphism (*PRNP* polymorphism)
- Pro-longevity loci and “antigeroid” syndromes, 434–435
- atherosclerosis, 434–435
- DAT, 434
- genetic resistance to environmental carcinogens, 435
- human allelic variants homologous to pro-longevity genes, 435
- Pro-opiomelanocortin gene (*POMC* gene), 150–151
- PROC*. *See* Protein C gene (*PROC*)
- Progeroid syndromes of humans, 429–433
- disorders of lipid and carbohydrate metabolism in segmental progeroid phenotypes, 433
- HGPS, 430–431, 431f
- MDPL syndrome, 431–432
- miscellaneous disorders in segmental progeroid phenotypes, 433
- rare genomic instability disorders in segmental progeroid phenotypes, 432–433
- WS, 429–430
- Progressive external ophthalmoplegia (PEO), 428
- Proinsulin, 75
- Prometaphase, 239
- Promoter (transcription regulatory) variants, 151–152
- Pronuclear transfer (PNT), 307–308
- Prophase, 239
- Proprotein convertase subtilisin/kexin type 9 (PCSK9), 25, 434
- Protein C gene (*PROC*), 162
- Protein disulfide isomerase (PDI), 74–75

- Proteins, 205, 383
 activity, 205
 alterations in, 424–425
 data, 341–342
 gene product, 8–9
 localization, 73–75
 new protein functions, 205
 product as unit of selection, 9
 protein-coding genes, 56
 protein–DNA complex, 254–255
 RNA translation into, 71–76
 synthesis, 72–73
 error catastrophe theory of aging, 424
 mRNA translation into protein, 73f
- Proteoglycans, 75
- Proteome, 341–342
 Exchange project, 342
- Proteomics, 341–342
- Proton gradient, 270
- Pseudoautosomal regions, 218–219
- Pseudodiploids, 255
- Pseudodominance, 211, 212f
- Pseudoexons, 155–156
- Pseudogenes, 61–62
- Psychouridine, 71
- Psychotropic drugs, GWAS dilemma of, 469–470
- Putative pathological variant, 157
- PWS. *See* Prader–Willi syndrome (PWS)
- Pyloric stenosis, 223
- Pyruvate dehydrogenase (PDH), 268–269
- Q**
- QC. *See* Quality control (QC)
- QQ plots, 335–336
- QTL. *See* Quantitative trait loci (QTL)
- Quality control (QC), 331
- Quantitative genetic theory, 387
- Quantitative trait loci (QTL), 235, 329
- Quantitative traits, 323–324
 linkage analysis, 235
- Quinacrine (Q), 244–245
 quinacrine-banded karyotype of male cell, 246f
- R**
- Random genetic drift, 366–367
- Random mating, 360
- Rapid acetylators, 454
- Rare genomic instability disorders, 432–433
- Rare variant model, 454
- RD. *See* Restrictive dermatopathy (RD)
- Reactive oxygen species (ROS), 427
 scavenging, 306
- Reads per kilobase per million mapped reads (RPKM), 339–340
- Recessiveness, 203–206
 recessive variants with dominant effects, 205–206
 in relation to underlying variants, 167–168
- Recurrence risks, 206, 207f, 211, 211f–212f, 214, 215f, 216–217, 218f
- “Red-cell TPMT activity”, 458
- Reduced representation bisulfite sequencing (RRBS), 97, 340–341
- Reference genome, 143–144
- Regulatory single nucleotide polymorphism (rSNP), 165
- Remote gene regulatory elements, variants in, 162–163
- Replication
 phase, 336–337
 replication-based mechanisms, 137
 slippage, 137
- Repulsion Ab/aB phase, 229
- Restrictive dermatopathy (RD), 431
- Retinitis pigmentosa, 221
- Rett syndrome, 86, 95–96, 218, 384
- Reverse genetics, 227
- Reverse-banded karyotype of female cell, 247f
- “Reverse” antagonistic pleiotropy, 420–421
- Ribosomal RNAs (rRNAs), 56, 96–97, 244
- Ribosome, 72–73
- Risk assessment, 23
- RNA. *See also* DNA.
 cleavage-polyadenylation mutants, 156
 molecules, 67–68
 polymerase II, 82–83
 polymerases, 56
 RNA-Seq data, 339–340
 synthesis, 57
 translation into protein
- expression of housekeeping and tissue-specific genes, 75–76
 genetic code, 71–72
 posttranslational modification, 75
 protein localization, 73–75
 protein synthesis, 72–73
- RNA-Seq by expectation maximization (sRSEM), 339–340
- RNASeqPower (Bioconductor package), 339–340
- RNAseqPS tool, 339–340
- Robertsonian translocation, 261
- ROS. *See* Reactive oxygen species (ROS)
- RPKM. *See* Reads per kilobase per million mapped reads (RPKM)
- RRBS. *See* Reduced representation bisulfite sequencing (RRBS)
- rRNAs. *See* Ribosomal RNAs (rRNAs)
- rSNP. *See* Regulatory single nucleotide polymorphism (rSNP)
- Russell–Silver syndrome (RSS), 90
- S**
- Saccharomyces cerevisiae*, 254–255
- Satellite DNA, 62
- SC. *See* Synaptonemal complex (SC)
- SCE. *See* Sister chromatid exchange (SCE)
- Schiff base, 424–425
- Schizophrenia, 101–102
- Sclerostin gene (*SOST*), 164–165
- SCN2A* gene, 95
- Segmental duplications (SDs), 61, 63–64, 136, 249–250, 261
- Segmental progeroid
 large, 62
 mutations, 429
 phenotypes
 disorders of lipid and carbohydrate metabolism, 433
 miscellaneous disorders, 433
 rare genomic instability disorders, 432–433
- Segregation, 203
- Segregation analysis, 327–328
- Seipin. *See* Berardinelli–Seip congenital lipodystrophy 2 (*BSCL2*)
- Selection, 368–370
- Senile dementia, 378
- Sense strand, 56

- Sequence kernel association test, 337–338
- Serial replication slippage (SRS), 137
- Serine proteinase inhibitor clade A member 1 (*SERPINA1*), 396
- Sertoli cells, 257
- Sex
chromosomes, 53–54, 256–259
determination, 256–259
X chromosome, 257–259
Y chromosome, 257
inconsistency, 332
influence, 207
limitation, 207
ratio in MZ twinning, 392
- Sex-determining region Y gene (*SRY* gene), 84
- Sex-linked inheritance, 213–218
X-linked dominant inheritance, 216–218
X-linked recessive inheritance, 213–216
Y-linked (holandric) inheritance, 218
- SHH*. See Sonic hedgehog gene (*SHH*)
- Short interspersed nuclear element (SINE), 61
- Short sequence repeats (SSRs), 128–129
- Sib-pair allele sharing linkage methods, 234
- Sibling relative risk, 327
- Signal recognition particle (SRP), 74
- Silver NOR (AgNOR), 248
- SINE. See Short interspersed nuclear element (SINE)
- Single base-pair substitutions within “splicing enhancer” sequences, 155
- Single nucleotide polymorphisms (SNPs), 7, 63, 126–127, 227, 252–254, 327, 333, 371, 398–399, 445, 453
arrays, 104
CpG dinucleotides, 131
genotype arrays, 231
markers, 233
nonsense, 129–130
- Single-gene disorders, 48–49, 222, 323
- Single-nucleotide variants (SNVs), 445, 453
- siRNAs. See Small interfering RNAs (siRNAs)
- Sister chromatid exchange (SCE), 248–249
protocol, 248–249
- Sister chromatid exchanges, 249f
- Skewed X-chromosomal inactivation, 397
- SLC25. See Solute carrier family 25 (SLC25)
- SLC2A2* transporter gene, 469
- SLCO1B1* gene, 464
- SLCOB1* gene, 464–467
- Slow acetylators, 454
- Small interfering RNAs (siRNAs), 82–83
- Small nuclear RNAs (snRNAs), 157
- Small-effect DNA variants, 454
- Smallest region of overlap (SRO), 90–92
- SMC. See Structural maintenance of chromosomes (SMC)
- SNPs. See Single nucleotide polymorphisms (SNPs)
- snRNAs. See Small nuclear RNAs (snRNAs)
- snRNPs, 69–71
- SNVs. See Single-nucleotide variants (SNVs)
- Social environmental exposures, 103–104
- SOD1* mutations. See Superoxide dismutase mutations (*SOD1* mutations)
- Solenoid model, 237–239
- Solute carrier family 25 (SLC25), 268–269
- Somatic abnormalities, 85–86
- Somatic cell(s), 270
genetic disorders, 49–50
mutation, 49–50
telomeres, 427
- Somatic mosaicism, 166, 209–210
- Somatic mtDNA mutations, developmental and, 283–284
- Somatic mutational spectra
comparison of germline and, 166
- Somatic variants, 208
- Sonic hedgehog gene (*SHH*), 162–163
- SOST*. See Sclerostin gene (*SOST*)
- Sotos syndrome, 94
- Species specificity, 422–423
- Sperm, 237
- Spermatogenesis, 243–244
- Spindle transfer (ST), 307–308
- Spinocerebellar ataxia, 62
- Splice-mediated insertional inactivation, 155–156
- Spliceosomes, 69–71
- SRO. See Smallest region of overlap (SRO)
- SRP. See Signal recognition particle (SRP)
- SRS. See Serial replication slippage (SRS)
- sRSEM. See RNA-Seq by expectation maximization (sRSEM)
- SRY* gene. See Sex-determining region Y gene (*SRY* gene)
- SSRs. See Short sequence repeats (SSRs)
- ST. See Spindle transfer (ST)
- Statin-induced myopathy, 464–467
- Statistical fine mapping of GWAS data sets, 338
- Statistical power, 334
- Steroid 21-hydroxylase (*CYP21*), 144
- Stochastic processes, 421–422
- Stratification, 366
- Structural chromosome abnormalities, 261–262
- Structural maintenance of chromosomes (SMC), 237–239
- Structure association testing, 332
- “Super-hotspot” motifs, 133–134
- Superfecundation, 391
- Superfetation, 391
- Superhelix, 237–239
- Supernumerary marker chromosomes, 262
- Superoxide dismutase mutations (*SOD1* mutations), 304
- Suppressor alleles, 420
- Surrogate variant, 338
- Sustained virological response (SVR), 468
- SVR. See Sustained virological response (SVR)
- Synaptonemal complex (SC), 241–243
- Synonymous nucleotide substitutions, 132

- T**
- T2D. *See* Type 2 diabetes (T2D)
- TADs. *See* Topologically associated domains (TADs)
- TARs. *See* Telomere-associated repeats (TARs)
- TATA box. *See* 5'-TATA-3' sequence
- 5'-TATA-3' sequence, 67–68
- Tay–Sachs disease, 367, 370
- TBFS. *See* Transcription factor binding sites (TBFS)
- TDF. *See* Testis-determining factor (TDF)
- TDT. *See* Transmission disequilibrium test (TDT)
- Technical validation, 336–337
- Telomerase, 256, 427
- Telomere-associated repeats (TARs), 256
- Telomeres, 61, 255–256, 427
- Telomeric DNA sequences, 62, 427
- Telophase, 241
- Template switching, 137
- Ten–Eleven–Translocation (TET), 83
- proteins, 83–84
- Termination codon (“nonstop”) variants, 160
- Testis, 244
- Testis-determining factor (TDF), 257
- TET. *See* Ten–Eleven–Translocation (TET)
- Tetranucleotide repeat expansion (CCTG), 135
- TFs. *See* Transcription factors (TFs)
- TG(A/G)(A/G)(G/T)(A/C) sequence, 132–133
- TGF- β . *See* Transforming growth factor- β (TGF- β)
- TGFB2R. *See* Type II TGF- β receptor (TGFB2R)
- Thalassemias, 370
- Thanatophoric dysplasia, 369
- Therapeutic index or window, 449
- 6-Thioguanine, 458
- Thiopurine methyltransferase polymorphism (TPMT gene), 458–459
- Threshold, 325
- Thrombopoietin (TPO), 159–160
- Thymidine (T), 248–249
- Thymine (T), 55, 131, 131f
- TIM complex. *See* Transport through inner mitochondrial membrane complex (TIM complex)
- Tissue-specific genes expression, 75–76
- TLS. *See* Translesion synthesis (TLS)
- TOM complex. *See* Transport through outer mitochondrial membrane complex (TOM complex)
- Topoisomerase II, 237–239
- Topologically associated domains (TADs), 64–65, 158
- Total genome, 445
- Toxic protein alterations, 205
- TPM. *See* Transcripts per kilobase million (TPM)
- TPMT gene. *See* Thiopurine methyltransferase polymorphism (TPMT gene)
- TPO. *See* Thrombopoietin (TPO)
- TPP⁺. *See* Triphenylphosphonium moiety (TPP⁺)
- Transcription factor binding sites (TBFS), 340–341
- Transcription factors (TFs), 56
- recognition element, 67–68
- 3D structure of DNA helix bound to, 68f
- Transcription(al), 55f, 56
- termination, 71
- units, 65–71
- Transcripts per kilobase million (TPM), 339–340
- Transethnic meta-analysis, 338–339
- Transfer RNA (tRNA), 56
- Transforming growth factor- β (TGF- β), 165–166
- TGF- β 1, 384
- Transgenerational epigenetic effect, 449–450
- Transgenerational mechanism, 470
- Transient hypermutability, 146
- Translational initiation codon variants, 159–160
- Translesion synthesis (TLS), 432
- Transmission disequilibrium test (TDT), 329–330
- Transport through inner mitochondrial membrane complex (TIM complex), 272
- Transport through outer mitochondrial membrane complex (TOM complex), 272
- Transposable elements, 61–62
- Transposon-derived repeats, 61
- Transposon-rich regions, 137
- Transthyretin (TTR), 162
- Triallelic inheritance, 169–170
- Trinucleotide expansion/CNV of, 134–135
- repeats, 134–135
- cellular consequences of trinucleotide repeat expansions, 163
- Triphenylphosphonium moiety (TPP⁺), 306
- Triploidy, 47–48, 260–261
- Trisomy 16, 47–48
- TRMU. *See* Methyaminomethyl-2-thiouridylate-methyltransferase (TRMU)
- tRNA. *See* Transfer RNA (tRNA)
- Truncated LINE, 140–141
- TTR. *See* Transthyretin (TTR)
- Twins/twinning, 387
- atypical twinning, 389–391
- etiology
- causes, 396–398
- genetic causes of DZ twinning, 395–396
- genetic causes of MZ twinning, 395
- incidence
- incidence of DZ twinning, 392
- incidence of monozygotic twinning, 392
- placentation, 391–392
- registries, 387
- and international collaboration, 403
- research, 400–403, 402t
- classic twin design, 400–401
- general statistical issues, 403
- types of twin designs, 401
- sex ratio in MZ twinning, 392
- structural defects in, 392–393
- structural defects in monozygotic twins, 393t
- typical twinning in humans, 388–389
- zygosity determination, 393–395, 394f
- Type 1 myotonic dystrophy (DM1), 163
- Type 2 diabetes (T2D), 98, 303, 433
- GWAS of treatment, 468–469

- Type 2 diabetes mellitus (T2DM). *See* Type 2 diabetes (T2D)
- Type I errors, 453
- Type II errors, 453
- Type II TGF- β receptor (TGFB2R), 165–166
- Typical twinning in humans, 388–389
chorionicity and amnionicity by time, 388t
formation of types of twins, 389f
monozygotic twinning during postfertilization, 388f
- Typological thinking, 3
- U**
- UCEs. *See* Ultraconserved elements (UCEs)
- UCSC. *See* University of California at Santa Cruz (UCSC)
- UDP glucuronosyltransferase-1A1 polymorphism (*UGT1A1* gene), 460
- UGT2B17* gene, 144
- Ultraconserved elements (UCEs), 56, 164–165
- Ultrarapid metabolizer (UM), 457
- Underdominance, 368
- Unifactorial inheritance/singlegene disorders, 202–203
- Uniparental disomy (UPD), 213, 252–254 and imprinting, 259–260
- Unit steps, 6
as effectors of disease, 8–9
qualities of homeostasis, 6–11
effectors of gene intention, 6
goal of HGP, 10–11
hedge against genetic determinism, 9–10
protein product as unit of selection, 9
social impact, 11
source of coherence, 11
unit of development, 6–7
unit of history, 6
unit of individuality, 7
- University of California at Santa Cruz (UCSC), 79, 333
- Unstable DNA triplet-repeat sequences, 208
- Unstable protein mutants, 162
- 3'-Untranslated region (3' UTRs), 71
variants in, 159
- 5' Untranslated region (5' UTRs), 69
variants in, 159
- UPD. *See* Paternal uniparental disomy (UPD); Uniparental disomy (UPD)
- Upstream ORFs (uORFs), 151
- US Food and Drug Administration (FDA), 25, 100–101, 473
recommendations for PGx genotyping, 473
- USH1C* gene, 135
- V**
- Validation phase, 336–337
- Valproic acid (VPA), 101
- Vanishing twin syndrome (VTS), 390
- Variance, 454
component models, 400–401
- Variants, 213, 216
- VDAC. *See* Voltage-dependent anion channels (VDAC)
- VKORC1* gene, 460–461
- Voltage-dependent anion channels (VDAC), 270
- von Willebrand factor (VWF), 144
deficiency, 170
- Voriconazole, 459
- VPA. *See* Valproic acid (VPA)
- VTS. *See* Vanishing twin syndrome (VTS)
- VWF. *See* von Willebrand factor (VWF)
- W**
- Warfarin polymorphisms, 460–461
- Watson strand, 55
- Werner syndrome (WS), 94–95, 375–376, 429–430
- Western medicine, 267
- WGBS. *See* Whole-genome bisulfite sequencing (WGBS)
- WGS. *See* Whole-genome sequencing (WGS)
- Whole-chromosome paint probes, 250–251
- Whole-exome sequencing, 222
- Whole-genome bisulfite sequencing (WGBS), 97, 340–341
- Whole-genome sequencing (WGS), 201, 222, 254
- “Why” questions, 12–15
- WRN* protein, 430
- WS. *See* Werner syndrome (WS)
- X**
- X chromosome, 257–259
- X inactivation, 214–215
epigenetic regulation, 84–86
special aspects relevant to human genetic diseases, 85–86
- X-chromosome inactivation (XCI), 84
- X-inactivation center (XIC), 84, 258
- X-inactivation-specific transcript gene (*XIST* gene), 60, 84–85, 258
- X-linked diseases, 168
- X-linked dominant
inheritance, 216–218, 216t, 217f
recurrence risks, 216–217
lethal alleles, 217–218
traits, 452
- X-linked locus, 362
- X-linked recessive inheritance, 213–216, 214f, 214t
gonadal mosaicism, 216
manifesting female carriers of X-linked recessive disorders, 215–216
new variants, 216
recurrence risks, 214
X inactivation, 214–215
- X-linked recessive traits, 452
- XCI. *See* X-chromosome inactivation (XCI)
- XCMS, 342–343
- Xeroderma pigmentosum (XP), 432
- XIC. *See* X-inactivation center (XIC)
- XIST* gene. *See* X-inactivation-specific transcript gene (*XIST* gene)
- XP. *See* Xeroderma pigmentosum (XP)
- XPF/progeroid syndrome, 432–433
- Y**
- Y chromosome, 257
- Y-linked (holandric) inheritance, 218, 219f
- Z**
- Zero-order kinetics, 381
- ZNF9*. *See* Myotonic dystrophy type 2 (*ZNF9*)
- Zygosity, 393–395, 394f
importance of zygosity knowledge, 394–395
incorrect assumptions about, 394