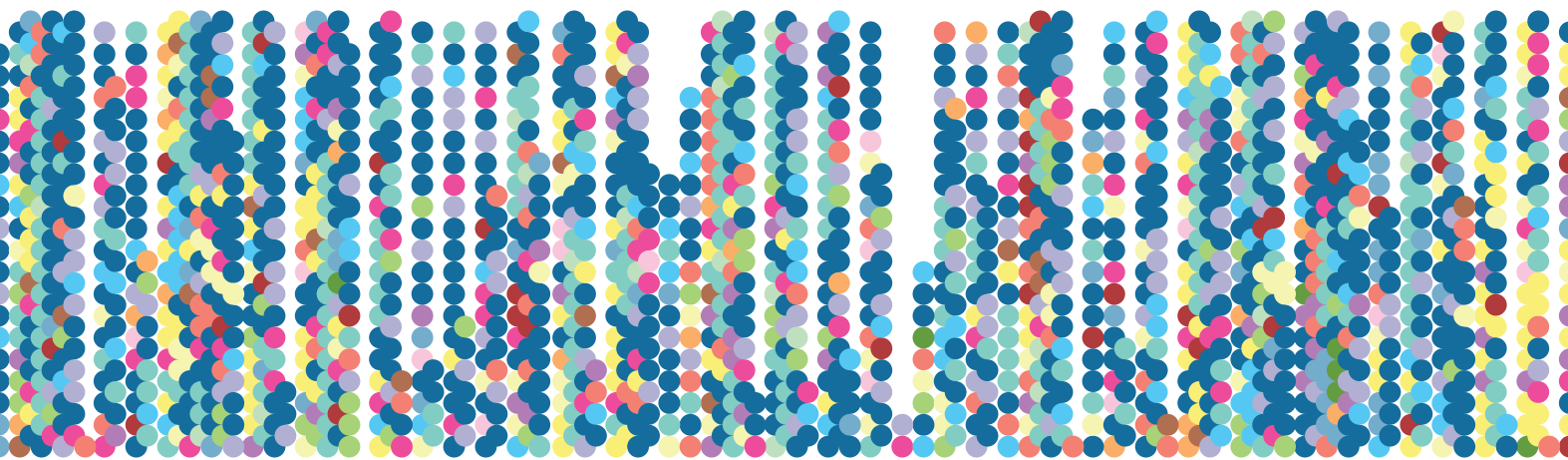


FOURTH EDITION

NEW CLINICAL GENETICS

A GUIDE TO GENOMIC MEDICINE



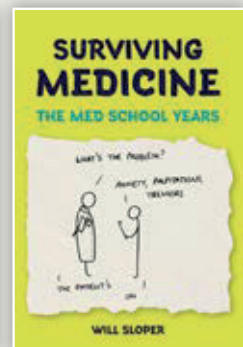
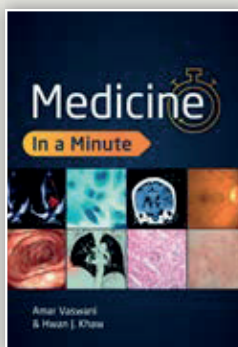
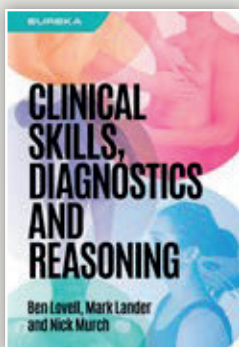
ANDREW READ AND DIAN DONNAI

NEW CLINICAL GENETICS

A GUIDE TO GENOMIC MEDICINE

FOURTH EDITION

Other titles from Scion



For more information on these and other titles, see www.scionpublishing.com



@ScionMedical



@scionpub



scionpublishing

FOURTH EDITION

NEW CLINICAL GENETICS

A GUIDE TO GENOMIC MEDICINE

Andrew Read and Dian Donnai

Manchester Centre for Genomic Medicine, Manchester Academic Health Science Centre,
University of Manchester, St Mary's Hospital, Manchester, UK



© **Scion Publishing Ltd, 2021**

First edition published 2007, reprinted 2009, 2010

Second edition published 2011, reprinted 2012, 2014

Third edition published 2015, reprinted 2016, 2017, 2020

Fourth edition published 2021

All rights reserved. No part of this book may be reproduced or transmitted, in any form or by any means, without permission.

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

ISBN 9781911510703

Scion Publishing Limited

The Old Hayloft, Vantage Business Park, Bloxham Road, Banbury OX16 9UX, UK

www.scionpublishing.com

Important Note from the Publisher

The information contained within this book was obtained by Scion Publishing Limited from sources believed by us to be reliable. However, while every effort has been made to ensure its accuracy, no responsibility for loss or injury whatsoever occasioned to any person acting or refraining from action as a result of information contained herein can be accepted by the authors or publishers.

Readers are reminded that medicine is a constantly evolving science and while the authors and publishers have ensured that all dosages, applications and practices are based on current indications, there may be specific practices which differ between communities. You should always follow the guidelines laid down by the manufacturers of specific products and the relevant authorities in the country in which you are practicing.

Although every effort has been made to ensure that all owners of copyright material have been acknowledged in this publication, we would be pleased to acknowledge in subsequent reprints or editions any omissions brought to our attention.

Illustrations by Matthew McClements at Blink Studio Ltd, www.blink.biz

Typeset by Evolution Design and Digital, Kent, UK

Printed in the UK

Last digit is the print number 10 9 8 7 6 5 4 3 2

Contents

Preface to the fourth edition	xxi
Dedication	xxii
Abbreviations	xxiii
How to use this book	xxv

Chapter 1 – What can we learn from a family history?

1.1 Case studies	1
Case 1 – Ashton family	1
Case 2 – Brown family	2
Box 1.1 – The pleiotropic effects of cystic fibrosis	3
Case 3 – Kowalski family	3
Case 4 – Davies family	4
Case 5 – Elliot family	4
Case 6 – Fletcher family	5
1.2 Science toolkit	6
Box 1.2 – How to take a family history and draw a pedigree	6
1.3 Investigations of patients	7
Box 1.3 – How John Ashton came to the genetic clinic and issues the geneticist considered	7
Case 1 – Ashton family	8
Case 2 – Brown family	10
Case 3 – Kowalski family	10
Case 4 – Davies family	11
Case 5 – Elliot family	12
Case 6 – Fletcher family	13
1.4 Going deeper ...	14
The art of pedigree interpretation	14
Box 1.4 – Summary of modes of inheritance of monogenic characters	16
Penetrance and expressivity – pitfalls in inheritance and counseling	17
Rarer modes of inheritance	18
Some further problems in pedigree interpretation	20
Mosaicism	20
Disease box 1 – Type 1 Neurofibromatosis	21
1.5 References	22
Useful websites	22
1.6 Self-assessment questions	23

Chapter 2 – How can a patient's chromosomes be studied?

2.1 Case studies	25
Case 7 – Green family	25
Case 8 – Howard family	26
Case 9 – Ingram family	26

2.2	Science toolkit	27
	Why clinicians need to know about chromosomes	27
	How are chromosomes studied?	27
	Box 2.1 – Material for chromosome analysis	28
	Box 2.2 – Chromosomes and their abnormalities: nomenclature and glossary	30
	Chromosome abnormalities	31
	Box 2.3 – Syndromes due to numerical chromosome abnormalities	32
	Box 2.4 – Recurrent microdeletion and microduplication syndromes	33
	Why do we have chromosomes?	34
	Centromeres and telomeres	34
	The behavior of chromosomes during cell division	34
2.3	Investigations of patients	38
	Case 7 – Green family	39
	Case 8 – Howard family	39
	Case 9 – Ingram family	42
	Case 5 – Elliot family	43
2.4	Going deeper ...	47
	What are chromosomes?	47
	Numerical and structural chromosome abnormalities	48
	Copy number variants	49
	Balanced and unbalanced abnormalities	50
	Constitutional and mosaic abnormalities	52
	Disease box 2 – A microdeletion syndrome: Williams–Beuren syndrome	53
2.5	References	54
	Useful websites	54
2.6	Self-assessment questions	55

Chapter 3 – How do genes work?

3.1	Case studies	57
	Case 10 – O'Reilly family	57
3.2	Science toolkit	58
	Structure of nucleic acids	59
	Box 3.1 – A note on units	60
	The structure of genes: exons and introns	60
	Box 3.2 – 5' and 3' ends	61
	Splicing of the primary transcript	62
	Translation and the genetic code	63
	Box 3.3 – The reading frame	64
	Translation is not the end of the story	64
	Box 3.4 – Biosynthesis of collagens	65
3.3	Investigations of patients	66
	Case 1 – Ashton family	67
	Case 2 – Brown family	67
	Case 3 – Kowalski family	67
	Case 4 – Davies family	68
	Case 5 – Elliot family	68
	Case 6 – Fletcher family	69
	Case 7 – Green family	70
	Case 8 – Howard family	70
	Case 9 – Ingram family	70
	Case 10 – O'Reilly family	70
3.4	Going deeper ...	71
	Some chemistry	71

One gene often encodes more than one protein	71
Box 3.5 – Chemical formulae of A, G, C, T and U	71
Box 3.6 – Structure of proteins	72
Switching genes on and off – transcription and its controls	72
From gene to genome	74
Box 3.7 – How to use the ENSEMBL genome browser	75
An overview of the human genome	76
Looking at our non-coding DNA	77
Disease box 3 – From genes to diseases: the RASopathies	78
3.5 References	80
Useful websites	80
3.6 Self-assessment questions	81

Chapter 4 – How can a patient's DNA be studied?

4.1 Case studies	83
Case 11 – Lipton family	83
Case 12 – Meinhardt family	84
4.2 Science toolkit	85
Nucleic acid hybridization	86
Using hybridization as the basis for DNA testing	86
Box 4.1 – Principle of Southern blotting	88
Box 4.2 – Restriction endonucleases	88
Box 4.3 – Gel electrophoresis	89
Amplifying a sequence of interest: the polymerase chain reaction	93
Box 4.4 – Understanding PCR	95
4.3 Investigations of patients	97
Cases studied using a hybridization procedure	97
Case 7 – Green family	97
Case 4 – Davies family	98
Case 5 – Elliot family	100
Case 12 – Meinhardt family	101
Case 3 – Kowalski family	102
Cases studied using PCR	103
Case 9 – Ingram family	103
Case 1 – Ashton family	103
Case 11 – Lipton family	105
4.4 Going deeper ...	108
Quantitative PCR	109
Chromosome painting	110
Testing RNA	111
Testing protein	111
Disease box 4 – Diseases caused by expanding nucleotide repeats	112
4.5 References	114
Useful websites	114
4.6 Self-assessment questions	114

Chapter 5 – How can we check a patient's DNA for gene mutations?

5.1 Case studies	117
Case 13 – Nicolaides family	117
5.2 Science toolkit	118
Methods for detecting specific sequence changes	118

	Box 5.1 – A brief guide to nomenclature of variants	120
	Methods for scanning a gene for any sequence change	120
	DNA sequencing – the ultimate test	121
	Sanger (dideoxy) sequencing	122
	Next generation sequencing (NGS)	125
5.3	Investigation of patients	128
	The stories so far ...	128
	Case 13 – Nicolaides family	129
	Case 6 – Fletcher family	130
	Case 2 – Brown family	132
	Case 10 – O'Reilly family	134
	Case 3 – Kowalski family	134
5.4	Going deeper ...	135
	The three questions	135
	Where's it all going?	137
	Disease box 5 – Sudden arrhythmic death syndrome	138
5.5	References	140
	Useful websites	140
5.6	Self-assessment questions	141

Chapter 6 – What do mutations do?

6.1	Case studies	143
	Case 14 – Jenkins family	143
6.2	Science toolkit	145
	Box 6.1 – Words to describe DNA sequence variants	145
	An overview of types of variants	146
	Deletion or duplication of a whole gene	146
	Disruption of a gene	147
	Variants that affect the transcription of an intact coding sequence	147
	Variants that affect splicing of the primary transcript	149
	Variants that cause errors in translation	150
6.3	Investigations of patients	153
	Case 1 – Ashton family	153
	Case 2 – Brown family	154
	Case 3 – Kowalski family	155
	Case 4 – Davies family	156
	Case 6 – Fletcher family	157
	Case 10 – O'Reilly family	158
	Case 13 – Nicolaides family	159
	Case 14 – Jenkins family	160
6.4	Going deeper ...	160
	Loss of function and gain of function changes	161
	Dominant or recessive?	162
	Understanding the phenotype	164
	Genotype–phenotype correlations	164
	Box 6.2 – Genotype–phenotype correlation in mutations of the <i>FGFR</i> genes	165
	How do mutations arise?	168
	Disease box 6 – Mosaicism in clinical genetics	169
6.5	References	171
	Useful websites	172
6.6	Self-assessment questions	172
	Box 6.3 – Partial sequence of <i>PAX3</i> gene	173

Chapter 7 – Is cancer genetic?

7.1	Case studies	175
	Case 15 – Tierney family	175
	Case 16 – Wilson family	176
	Case 17 – Xenakis family	176
7.2	Science toolkit	177
	Natural selection and the evolution of cancer	177
	Overcoming the defenses	178
	Box 7.1 – Genomic instability in cancer cells	179
	Box 7.2 – Living for ever: the importance of telomeres	180
	Oncogenes	181
	Tumor suppressor genes	184
	The normal functions of tumor suppressor genes	186
	Box 7.3 – The G1–S checkpoint	187
	Mismatch repair and microsatellite instability	188
	The multistage development of cancer	189
7.3	Investigations of patients	190
	Case 15 – Tierney family	190
	Case 16 – Wilson family	191
	Box 7.4 – A scoring system for assessing the likelihood of a <i>BRCA1/2</i> mutation	192
	Box 7.5 – Liquid biopsies	195
	Case 24 – Xenakis family	195
7.4	Going deeper ...	197
	Getting the complete picture: whole genome studies	197
	Box 7.6 – Multi-omics approaches to cancer	197
	Genomics-based classification of tumors	200
	Disease box 7 – Chronic myeloid leukemia	202
7.5	References	204
	Useful websites	204
7.6	Self-assessment questions	205

Chapter 8 – How do researchers identify genes for mendelian diseases?

8.1	Case studies	207
	Case 18 – Choudhary family	207
8.2	Science toolkit	208
	Associating a phenotype with a DNA sequence variant	209
	Identifying a gene through its product	209
	Identifying a gene through a chromosomal abnormality	210
	Identifying a gene through an animal model	210
	Identifying a gene by positional cloning	210
	Box 8.1 – Genetic markers	212
	Identifying a gene by autozygosity mapping	213
	Identifying a gene by exome or genome sequencing	214
	Demonstrating why a variant causes a phenotype	215
8.3	Investigations of patients	216
	Case 18 – Choudhary family	216
8.4	Going deeper ...	220
	What happens if exome sequencing does not identify a candidate gene?	223
	Disease box 8 – The Deciphering Developmental Disorders (DDD) Study	225

8.5	References	226
	General background	227
	Useful websites	227
8.6	Self-assessment questions	227

Chapter 9 – Why are some conditions common and others rare?

9.1	Case studies	231
	Case 19 – Ulmer family	231
9.2	Science toolkit	232
	Box 9.1 – The Hardy–Weinberg distribution	233
	Using Hardy–Weinberg to calculate carrier risks	233
	Changing allele frequencies	234
	Factors determining allele frequencies	235
	Heterozygote advantage	237
	Heterozygote advantage or founder effect?	238
9.3	Investigations of patients	239
	Case 19 – Ulmer family	239
	Box 9.2 – The risk a healthy sib is a carrier	239
	Case 18 – Choudhary family	240
	Box 9.3 – Calculating the effects of inbreeding	241
9.4	Going deeper ...	242
	What is the chance the offspring of a consanguineous marriage will have a recessive disease?	244
	Can we abolish genetic disease?	244
	Box 9.4 – Should treated people repay their debt to society by not having children?	245
	Identifying relatives – and criminals	245
	Disease box 9 – Jewish diseases and Finnish diseases	247
9.5	References	248
9.6	Self-assessment questions	249

Chapter 10 – How do our genes affect our metabolism, drug responses and immune system?

10.1	Case studies	251
	Case 20 – Vlasi family	251
	Case 21 – Portillo family	252
	Box 10.1 – Types and functions of lymphocytes	252
10.2	Science toolkit	253
	Inborn errors of metabolism	253
	Box 10.2 – Some history	255
	Pharmacogenetics	256
	Box 10.3 – Robert Smith's debrisoquine misadventure	257
	Immunogenetics	258
	Box 10.4 – Recombination and gene conversion	260
10.3	Investigations of patients	261
	Case 15 – Tierney family	261
	Case 20 – Vlasi family	262
	Case 21 – Portillo family	263
10.4	Going deeper ...	266
	Inborn errors of metabolism	266
	Box 10.5 – Inability to make vitamin C – a universal inborn error in humans	268

Box 10.6 – Lactose intolerance – a common metabolic polymorphism	268
Pharmacogenetics	269
Immunogenetics	271
Disease box 10 – Disorders of the spliceosome	273
10.5 References	275
Useful websites	275
10.6 Self-assessment questions	275

Chapter 11 – How are genes regulated?

11.1 Case studies	277
Case 22 – Qian family	277
Case 23 – Rogers family	278
11.2 Science toolkit	278
Two definitions of epigenetics	279
X-inactivation: a classic epigenetic process	280
Imprinting – why you need a mother and a father	283
11.3 Investigations of patients	285
Case 4 – Davies family	285
Case 9 – Ingram family	285
Case 21 – Portillo family	286
Cases 22 – Qian family and 23 – Rogers family	287
11.4 Going deeper ...	291
DNA methylation	292
Studying DNA methylation	294
Box 11.1 – DNA methylation and CpG islands	295
Histone modifications	295
Chromatin conformation	296
Box 11.2 – Higher-level chromatin organization	297
How far do epigenetic effects determine individual differences?	298
Box 11.3 – Can epigenetic effects operate across generations?	298
Disease box 11 – Chromatin diseases	299
11.5 References	302
Useful website	303
11.6 Self-assessment questions	303

Chapter 12 – When is screening useful?

12.1 Case studies	305
Case 24 – Smit family	305
12.2 Science toolkit	306
Screening versus diagnostic tests	306
Box 12.1 – Parameters of a screening test	308
When might screening be done?	308
Who should be screened?	310
How should screening be done?	310
Antenatal screening for Down syndrome and other trisomies	310
Box 12.2 – What is the best prenatal diagnostic test for Down syndrome?	312
12.3 Investigations of patients	313
Case 2 – Brown family	313
Case 4 – Davies family	315

Case 8 – Howard family	315
Case 13 – Nicolaides family	316
Case 19 – Ulmer family	316
Case 20 – Vlasi family	317
Case 24 – Smit family	318
12.4 Going deeper ...	319
What conditions should we screen for?	319
Box 12.3 – The Population Attributable Risk	321
Box 12.4 – Key criteria used by the UK National Screening Committee	323
Incidental findings – a form of opportunistic screening	325
'Lifestyle' genetic testing	326
Disease box 12 – Non-invasive prenatal testing	329
12.5 References	330
Useful websites	331
12.6 Self-assessment questions	332

Chapter 13 – Should we be testing for genetic susceptibility to common diseases?

13.1 Case studies	333
Case 25 – Yamomoto family	333
Case 26 – Zuabi family	334
13.2 Science toolkit	335
Estimating heritability	336
Identifying genetic susceptibility factors	337
Box 13.1 – Linkage versus association	338
Genome-wide association studies	338
Box 13.2 – Effect sizes, odds ratios and allele frequencies	343
13.3 Investigations of patients	344
Case 25 – Yamomoto family	344
Box 13.3 – Measuring the performance of a test using the ROC curve	348
Case 26 – Zuabi family	349
13.4 Going deeper ...	352
Why have GWAS told us so little that is clinically useful?	352
The 'missing heritability' problem	353
Polygenic risk scores	354
So should we be testing for susceptibility to common diseases?	355
What is we knew everything?	356
Disease box 13 – Autism spectrum disorders	358
13.5 References	360
Useful websites	362
13.6 Self-assessment questions	362

Chapter 14 – What clinical services are available for families with genetic disorders?

14.1 The work of the genetics service	365
Reasons for genetic referral	366
Box 14.1 – Reproductive genetic issues	367
Box 14.2 – Concerns about a pregnancy	369
14.2 Diagnosis and testing	371
The importance of a diagnosis	371
Dysmorphology	372
Box 14.3 – Terminology used in dysmorphology	374

	Investigations	374
	Box 14.4 – Predictive testing for Huntington disease	376
	Box 14.5 – Obtaining fetal material	379
	Box 14.6 – Pre-implantation genetic diagnosis	381
14.3	Counseling and risk estimation	382
	Risk assessment	383
	Box 14.7 – An introduction to Bayesian calculations in genetics	384
14.4	Management and treatment	385
	Box 14.8 – Methods for inserting an exogenous gene into a cell	389
	Box 14.9 – Treatment of spinal muscular atrophy	391
	Management and treatment for Cases 1–26	395
	Box 14.10 – Immunotherapy of cancer	399
14.5	The evolving role of the genetics service	401
	The rise of acute genetics	401
	Major areas of change in genetic services in the last 5 years	402
14.6	References	402
	Recommended textbooks	404
14.7	Self-assessment questions	404

Chapter 15 – How to use linkage to map a disease gene

Online chapter – see www.scionpublishing.com/NCG4

Guidance for self-assessment questions

Chapter 1	407
Chapter 2	407
Chapter 3	408
Chapter 4	408
Chapter 5	408
Chapter 6	408
Chapter 7	409
Chapter 8	409
Chapter 9	410
Chapter 10	410
Chapter 11	411
Chapter 12	411
Chapter 13	412
Chapter 14	412

Glossary	415
-----------------	------------

Index	429
--------------	------------

Disease index	439
----------------------	------------

Case notes – Summary of cases and their page references

CASE 1 ASHTON FAMILY

1

8

67

103

153

395

- John, healthy 28-year-old son of Alfred Ashton – *Chapter 1*
- Family history of ? Huntington disease
- Autosomal dominant inheritance
- Need for diagnostic PCR test – *Chapter 3*
- PCR test confirms diagnosis in John's father – *Chapter 4*
- Pros and cons of predictive test
- Molecular pathology – *Chapter 6*
- Possibilities for therapy – *Chapter 14*

CASE 2 BROWN FAMILY

2

10

67

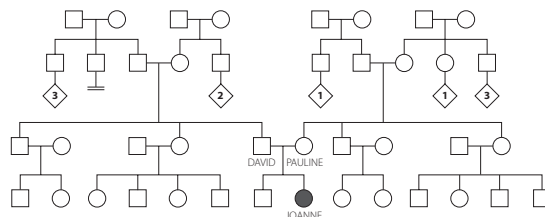
132

154

313

395

- Baby Joanne, recurrent infections, poor growth – *Chapter 1*
- Sweat test confirms she has cystic fibrosis
- Autosomal recessive inheritance
- Need for molecular test – *Chapter 3*
- *CFTR* variants identified – *Chapter 5*
- Molecular pathology – *Chapter 6*
- Approaches to screening – *Chapter 12*
- Possibilities for therapy – *Chapter 14*

**CASE 3 KOWALSKI FAMILY**

3

10

67

102

134

155

395

- Karol, first son of Kamil and Klaudia – *Chapter 1*
- Developmental delay, hypotonic, severe intellectual disability
- Difficulties of genetic testing in such cases
- Likely need for exome sequencing – *Chapter 3*
- Negative SNP chip test for microdeletions – *Chapter 4*
- Exome sequencing – *Chapter 5*
- *De novo ARID1B* variant identified – *Chapter 6*
- Possibilities for therapy – *Chapter 14*

CASE 4 DAVIES FAMILY

4

11

68

98

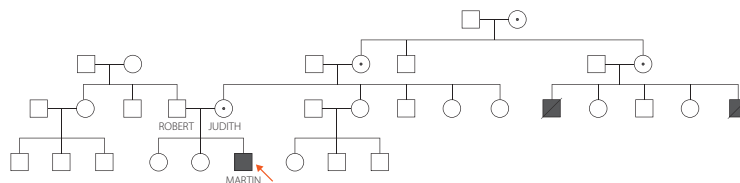
156

285

315

395

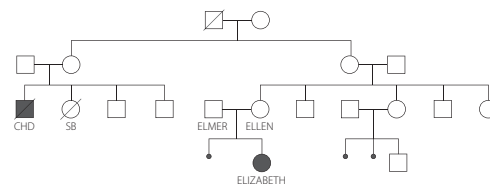
- Martin, aged 24 months, clumsy and slow to walk – *Chapter 1*
- Family history of muscular dystrophy
- X-linked recessive inheritance
- Problems of testing dystrophin gene – *Chapter 3*
- Exon 44–48 deletion identified by MLPA – *Chapter 4*
- Molecular pathology – *Chapter 6*
- Implications of X-inactivation – *Chapter 11*
- Screen all newborn boys? – *Chapter 12*
- Possibilities for therapy – *Chapter 14*



CASE 5 ELLIOT FAMILY

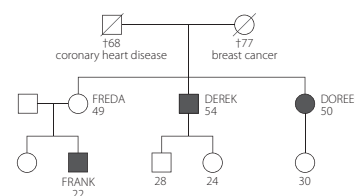
4 12 43 68 100 395

- Baby girl Elizabeth, parents Elmer and Ellen – *Chapter 1*
- Multiple congenital abnormalities
- Family history of reproductive problems
- ? Chromosome abnormality
- Ellen – balanced 1:22 translocation – *Chapter 2*
- Elizabeth – unbalanced segregation product
- Reciprocal translocation – *Chapter 3*
- Translocation identified by array-CGH – *Chapter 4*
- Possibilities for therapy – *Chapter 14*

**CASE 6 FLETCHER FAMILY**

5 13 69 130 157 395

- Frank, aged 22, with increasingly blurred vision – *Chapter 1*
- Family history of visual problems
- Possible mitochondrial inheritance
- ? Leber hereditary optic neuropathy
- Test mitochondrial genome – *Chapter 3*
- m.G3460A mutation identified – *Chapter 5*
- Molecular pathology – *Chapter 6*
- Possibilities for therapy – *Chapter 14*

**CASE 7 GREEN FAMILY**

25 39 70 97 395

- George, aged 3 years – *Chapter 2*
- Developmental delay, mildly dysmorphic
- Normal 46,XY karyotype but suspect microdeletion
- Test for microdeletions – *Chapter 3*
- 22q11 deletion identified by FISH – *Chapter 4*
- Possibilities for therapy – *Chapter 14*

CASE 8 HOWARD FAMILY

26 39 70 315 395

- Helen, newborn daughter of young parents – *Chapter 2*
- Down syndrome confirmed
- 47,XX,+21 karyotype
- Options for prenatal testing – *Chapter 3*
- Non-invasive prenatal test – *Chapter 12*
- Possibilities for therapy – *Chapter 14*

CASE 9 INGRAM FAMILY

26 42 70 103 285 395

- Isabel, 10 years old with small stature and possibly delayed puberty – *Chapter 2*
- ? Turner syndrome
- 45,X karyotype
- Risk of Y-chromosome DNA – *Chapter 3*
- PCR test for Y sequences negative – *Chapter 4*
- Questions around X-inactivation – *Chapter 11*
- Possibilities for therapy – *Chapter 14*

CASE 10 O'REILLY FAMILY

57

70

134

158

395

- Orla has severe myopia, short stature and hip problems – *Chapter 3*
- Family history of similar problems
- ? Stickler syndrome
- Test collagen II genes
- Sequencing identifies *COL2A1* variant – *Chapter 5*
- Molecular pathology – *Chapter 6*
- Possibilities for therapy – *Chapter 14*

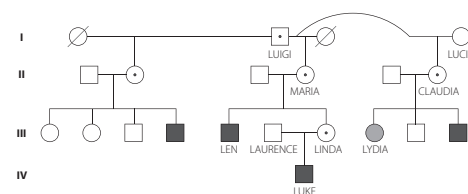
CASE 11 LIPTON FAMILY

83

105

395

- Baby boy, Luke, with developmental delay – *Chapter 4*
- Family history of learning difficulties
- Unusual features of Fragile-X pedigrees
- Caused by unstable repeat expansion
- Premutations, full mutations and normal transmitting males
- Measuring repeat expansions
- Possibilities for therapy – *Chapter 14*

**CASE 12 MEINHARDT FAMILY**

84

101

395

- Madelena, baby daughter of Margareta and Manfred – *Chapter 4*
- Multiple congenital abnormalities and developmental delay
- Normal 46,XX karyotype under the microscope
- 16p microdeletion identified by SNP chip
- Is this microdeletion pathogenic?
- ? Recurrence risk
- Possibilities for therapy – *Chapter 14*

CASE 13 NICOLAIDES FAMILY

117

129

159

316

395

- Spiros and Elena both carriers of β -thalassemia – *Chapter 5*
- Need to define mutations for prenatal diagnosis
- Allele-specific PCR shows Spiros carries the p.Gln39X variant
- Restriction digest shows Elena carries the c.316–106C>G variant
- Molecular pathology – *Chapter 6*
- Population screening for carriers – *Chapter 12*
- Possibilities for therapy – *Chapter 14*

CASE 14 JENKINS FAMILY

143

160

395

- James Jenkins, achondroplasia diagnosed in infancy – *Chapter 6*
- No family history
- Father was 58 years old when James conceived
- James's wife Joanne also has achondroplasia
- Obstetric problems and risks to children
- All cases have same *FGFR3* mutation
- Reasons for apparent high mutation rate
- Possibilities for therapy – *Chapter 14*

CASE 15 TIERNEY FAMILY

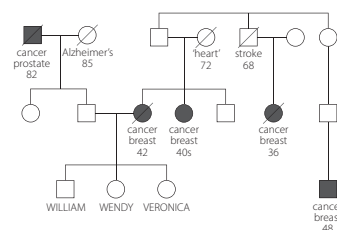
175 190 261 395

- 4-year-old boy, Jason – *Chapter 7*
- Pale with extensive bruising and tachycardia
- ? Acute lymphocytic leukemia
- Diagnosis of ALL confirmed with *TEL-AML1* fusion gene
- TPMT test prior to chemotherapy – *Chapter 10*
- Severe adverse reaction after false negative TPMT result – *Chapter 10*
- Possibilities for therapy – *Chapter 14*

CASE 16 WILSON FAMILY

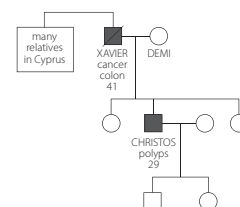
176 191 395

- Family history of breast cancer – *Chapter 7*
- Options for genetic testing
- Family *BRCA2* mutation identified
- Implications for relatives
- Possibilities for therapy – *Chapter 14*

**CASE 17 XENAKIS FAMILY**

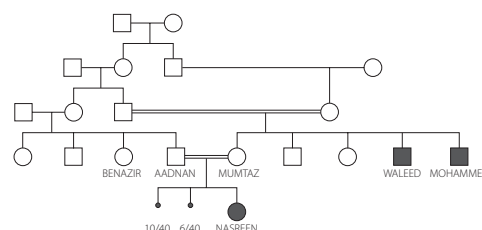
176 195 395

- Family history of bowel problems – *Chapter 7*
- ? Familial adenomatous polyposis
- *APC* mutation identified
- Risk to relatives
- How to manage his children?
- Possibilities for therapy – *Chapter 14*

**CASE 18 CHOUDHARY FAMILY**

207 216 240 395

- Baby girl Nasreen, healthy but deaf – *Chapter 8*
- Multiply consanguineous family
- Autozygosity mapping
- Exome sequencing
- A second recessive condition?
- Calculate coefficient of inbreeding – *Chapter 9*
- Possibilities for therapy – *Chapter 14*

**CASE 19 ULMER FAMILY**

231 239 316 395

- Hannah, 6-month-old baby girl, Ashkenazi Jewish background – *Chapter 9*
- Normal at birth but then increasing problems
- ? Tay-Sachs disease
- Enzyme test confirms diagnosis
- Test the sibs?
- Carrier screening – *Chapter 12*
- Possibilities for therapy – *Chapter 14*

CASE 20 VLASI FAMILY

251

262

317

395

396

- Valon, 6-year-old boy with serious learning problems – *Chapter 10*
- Small, microcephalic, blue eyes, fair skin and hair, eczema; hyperactive
- ? Phenylketonuria
- Testing for subsequent baby?
- Newborn screening – *Chapter 12*
- Possibilities for therapy – *Chapter 14*

CASE 21 PORTILLO FAMILY

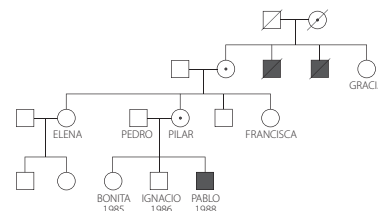
252

263

286

395

- Sickly boy, Pablo – *Chapter 10*
- Family history of similar problems – *Chapter 10*
- X-linked severe combined immunodeficiency
- Bone marrow transplantation
- Genetic cause defined
- Carrier tests for female relatives
- Implications of X-inactivation – *Chapter 11*
- Possibilities for therapy – *Chapter 14*

**CASE 22 QIAN FAMILY**

277

287

395

- Girl, Kai, aged 2 years – *Chapter 11*
- Developmental delay, seizures
- ? Angelman syndrome
- Causes and genetic tests
- Possibilities for therapy – *Chapter 14*

CASE 23 ROGERS FAMILY

278

287

395

- Baby boy, Robert, born to older parents – *Chapter 11*
- Normal 46,XY karyotype and pregnancy tests
- Severely hypotonic
- ? Prader–Willi syndrome
- Causes and genetic tests
- Possibilities for therapy – *Chapter 14*

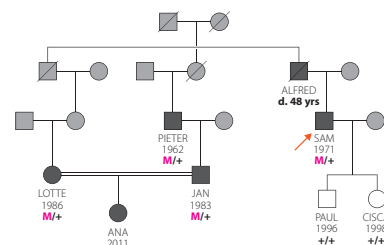
CASE 24 SMIT FAMILY

305

318

395

- Sam Smit, familial hypercholesterolemia – *Chapter 12*
- Identified through cascade screening
- LDLR mutation detected, treatment started
- Affected relatives, including a homozygote
- Conflict between privacy and cascade screening
- Possibilities for therapy – *Chapter 14*



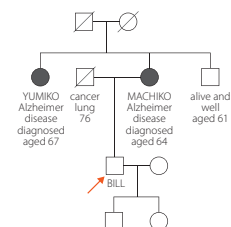
CASE 25 YAMOMOTO FAMILY

333

344

395

- Family history of dementia – *Chapter 13*
- Alzheimer disease
- Test for ApoE4?
- Genetic susceptibility to Alzheimer disease
- Possibilities for therapy – *Chapter 14*

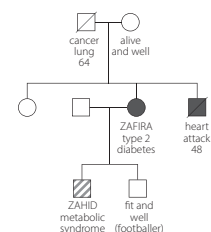
**CASE 26 ZUABI FAMILY**

334

349

395

- Zafra, woman aged 52 years – *Chapter 13*
- Overweight, sedentary lifestyle, insatiable thirst
- Type 2 diabetes
- Son's lifestyle and heredity put him at high risk
- Management of family
- Genetic susceptibility to Type 2 diabetes
- Possibilities for therapy – *Chapter 14*



Preface to the fourth edition

The science and technology, and their applications in genomic medicine, continue to evolve. For this edition we have therefore revisited every page of the previous edition, and amended or rewritten many to reflect new science, new techniques or new ways of thinking. We also rearranged the order of chapters, bringing the chapter on cancer forward so that it immediately follows *Chapter 6* where the idea of loss of function and gain of function changes is introduced. The final chapter, discussing services, has been considerably expanded, reflecting the increasing role of genomic medicine services in management of patients. Genomics is now relevant to the diagnosis and treatment of patients from many disciplines, and clinical geneticists and scientists play important roles in many multidisciplinary teams. The way these services have evolved is summed up in a series of bullet points at the very end of that chapter, making a fitting conclusion to the whole book. Genetics is now part of the curricula for professional training in medical specialties, nursing, midwifery and pharmacy and for many other professions allied to medicine. We hope that this edition will appeal to all who wish to learn the science and practice of modern genomics.

Big Data is ever more important in underpinning clinical applications. Over the past few years, thanks to initiatives such as the UK Biobank and the GnomAD database, we have vastly more information about the range of genetic variation found in normal healthy people, and this offers an essential background to assessing variants found in patients. This has allowed polygenic risk scores to start to become useful clinical service tools. There are still caveats about their general applicability, but it appears that at long last genetic analysis may have something useful to contribute to management of patients with common complex disorders, in addition to identifying monogenic subsets of such diseases. Other topics that have gained importance since our previous edition include non-invasive prenatal testing, companion diagnostics for prescribed drugs, liquid biopsies in cancer and preimplantation diagnosis.

To keep the book the same length despite these new matters, something had to give. We have reduced coverage of techniques that have been largely replaced by sequencing, in particular linkage analysis. However, we did not want to lose that topic altogether, so it is now in an online electronic supplement, www.scionpublishing.com/NCG4 “Resources”.

Some things have not changed. The book is still arranged around a set of clinical anecdotes, fictional but reflecting our long real-world experience in the clinic and the laboratory. Each chapter addresses a question that students may have about clinical genetics. As before, we give hints but not answers to the self-assessment questions – we want students to think, not just look up the answers. We hope you enjoy using the book, and come to share the fascination and enthusiasm that have motivated both of us for many years, and which continue to do so.

As always, we are grateful to the many colleagues who have helped with suggestions, data or pictures, and to Jonathan Ray of Scion Publishing for his tact and skill in turning our drafts into the finished article.

Andrew Read and Dian Donnai

Dedication

We dedicate this edition to our grandchildren in the confident expectation that, by the time they are grown up, much that we don't know now will be known then.

Ben	Lucille
Joe	Zac
Lois	Camille
James	
(APR)	(DD)

Abbreviations

ACTH	adrenocorticotrophic hormone	IRT	immunoreactive trypsin
ADR	adverse drug reaction	IVF	<i>in vitro</i> fertilization
AFP	alpha-fetoprotein	LDL	low density lipoprotein
AIS	androgen insensitivity syndrome	LHON	Leber hereditary optic neuropathy
ALL	acute lymphocytic leukemia	MAPK	mitogen-activated protein kinase
APP	amyloid precursor protein	MCAD	medium chain acyl-CoA dehydrogenase
ASO	allele-specific oligonucleotide	MDT	multidisciplinary team
ASP	affected sib pair	MH	malignant hyperthermia
CAH	congenital adrenal hyperplasia	MHC	major histocompatibility complex
CCC	chromosome conformation capture	ML	mucopolipidosis
CDK	cyclin-dependent kinase	MLPA	multiplex ligation-dependent probe amplification
CGH	comparative genomic hybridization	MMR	mis-match repair
cM	centiMorgan	MODY	maturity-onset diabetes in the young
CNVs	copy number variants	MSAFP	maternal serum alpha-fetoprotein
CVB	chorion villus biopsy	MZ	monozygotic
dHPLC	denaturing high performance liquid chromatography	NAHR	non-allelic homologous recombination
DMD	Duchenne muscular dystrophy	NGS	next generation sequencing
DMR	differentially methylated region	NIPT	non-invasive prenatal testing
DNA	deoxyribonucleic acid	NK	natural killer
DSH	dyschromatosis symmetrica hereditaria	NSC	National Screening Committee
DTC	direct-to-consumer	NTD	neural tube defect
DTDST	diastrophic dystrophy sulfate transporter	OLA	oligonucleotide ligation assay
DZ	dizygotic	OMIM	Online Mendelian Inheritance in Man
FAP	familial adenomatous polyposis	PCR	polymerase chain reaction
FDA	Food and Drug Administration	PKU	phenylketonuria
FH	familial hypercholesterolemia	PRS	polygenic risk scores
FISH	fluorescence <i>in situ</i> hybridization	PWS	Prader–Willi syndrome
GvH	graft versus host	QTL	quantitative trait locus
GWAS	genome-wide association studies	RB	retinoblastoma
HAT	histone acetyltransferase	RFLP	restriction fragment length polymorphism
HDAC	histone deacetylase	RNA	ribonucleic acid
HLA	human leukocyte antigen	RSTS	Rubinstein–Taybi syndrome
HNPCC	hereditary non-polyposis colon cancer	SNP	single nucleotide polymorphism
Ig	immunoglobulins	SSCP	single strand conformation polymorphism
iPSC	induced pluripotent stem cells	SUMF	sulfatase modifying factor

Abbreviations

T2D	type 2 diabetes	TPMT	thiopurine methyl transferase
TAD	topologically associating domain	TS	tumor suppressor
TCGA	the Cancer Genome Atlas	UPD	uniparental disomy
TDT	transmission disequilibrium test	VKOR	vitamin K epoxide reductase
TGF β	transforming growth factor β	X-SCID	X-linked severe combined immunodeficiency

How to use this book

Each chapter addresses a specific question that students may have about genetics, e.g. How can we study chromosomes? Is cancer genetic? etc. Chapters follow a common structure:

- *Learning points* – summarizing what the chapter should enable the student to achieve. These are chosen to cover the curriculum published by the American Society of Human Genetics and the list of competencies being developed by the UK NHS Genetics Education Centre.
- *Case studies* – all chapters except the last one have this section, which introduces a series of short clinically oriented descriptions of a family and the reasons why they sought genetic counseling or testing. The cases are all fictional and the photographs illustrate the condition rather than the patients described; but they are based on our long experience of dealing with real families.
- *Background* – an explanation of the methods and concepts that are necessary in the next section.
- *Investigations of patients* – using the case studies to illustrate the application of the methods and concepts in realistic scenarios. These form running stories that develop over several chapters.
- *Going deeper* – summarizes and extends the relevant science.

Students can choose different routes through the material:

CASE 1 ASHTON FAMILY

- John, healthy 28-year-old son of Alfred Ashton – *Chapter 1*
- Family history of ? Huntington disease
- Autosomal dominant inheritance

1 8 67 103 153 395

- Those who prefer a problem-based approach would move from the initial *Case studies* to the *Investigations of patients* sections. They should treat the other sections as reference material. The page design includes page references to enable students to follow cases that run through several chapters. For example, the Ashton family case is covered on pages 1, 8, 67, 103, 153 and 395. In addition, the bullets associated with each case serve as reminders as to what has been covered previously
- Students who prefer a more didactic approach can concentrate on the science in the *Background* and *Going deeper* sections. The other sections provide illustrations and examples.

Chapters finish with self-assessment questions so that students can check that they have mastered the material, regardless which route they chose to take through it.

The book is not intended as a diagnostic manual and so does not aim to describe all the major mendelian and chromosomal disorders. However, the cases and the *Disease boxes* help to show the range and variety of genetic conditions. For specific information you should refer to the reliable websites listed in the *References* section of relevant chapters. The first place to look should be the OMIM database: www.ncbi.nlm.nih.gov/omim – OMIM reference numbers are given throughout the book.

01

What can we learn from a family history?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Take a family history
- Draw a pedigree using the correct symbols
- Identify the most likely mode of inheritance, given a straightforward pedigree
- Describe how genes segregate in autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive, Y-linked and mitochondrial conditions
- Define penetrance and expressivity
- Show appreciation of the human and scientific issues raised by the conditions described

1.1. Case studies

CASE 1 ASHTON FAMILY

- John, healthy 28-year-old son of Alfred Ashton
- Family history of ?
Huntington disease

1

8

67

103

153

395

Alfred Ashton, aged 52, had been getting forgetful and was thought to be depressed after losing his job. He has been seeing a psychiatrist, who noted that Alfred was restless and had some choreiform movements (involuntary jerky movements of his fingers and shoulders and facial grimacing). Alfred had told the psychiatrist that he thought he was developing 'the family disease', though he was vague as to what this was. The psychiatrist suspected that Alfred had Huntington disease (OMIM 143100). His 28-year-old son John has been referred to the genetic clinic by his family doctor at the suggestion of the psychiatrist. John knows nothing about Huntington disease; he is preoccupied with other things, having recently married and bought a house.



Figure 1.1 – Huntington Disease.

(a) A patient in the advanced stages of the disease showing involuntary movements of the head and face. (b) Post mortem sections comparing normal brain (left) with brain from Huntington disease patient (right); note the loss of tissue in the Huntington disease brain. Photos courtesy of Dr David Craufurd, St Mary's Hospital, Manchester.

CASE 2 BROWN FAMILY

- Baby Joanne, recurrent infections, poor growth
- Sweat test confirms she has cystic fibrosis

2 10 67 132 154 313 395

Joanne is the second child born to David and Pauline Brown. Her older brother Jason is now 4 years old and very healthy, in fact his parents have to buy age 6 clothes for him. Joanne, however, is a different matter. She has worried her parents from the start. Although she took her bottle well she was very slow to put on weight and in her first few months seemed constantly to have a cold and a cough. At first Pauline and the doctor put this down to the fact that Jason had just started nursery and had a few colds himself. When she was 5 months old Joanne was really ill and was admitted to hospital with a chest infection. The nurses commented that her bowel motions were very bulky and offensive, and when her weight and height were plotted on the charts they were on lower centiles than they had been at birth and in her first month. The doctors suspected she might have cystic fibrosis (OMIM 219700) and therefore arranged for her to have a sweat test. This confirmed their suspicion (Cl^- level of 87 mmol/l, well above the normal upper limit of 40 mmol/l). The diagnosis came as a complete bombshell to Pauline and David. At their request, Joanne's pediatrician arranged for them to see a geneticist to talk things through.

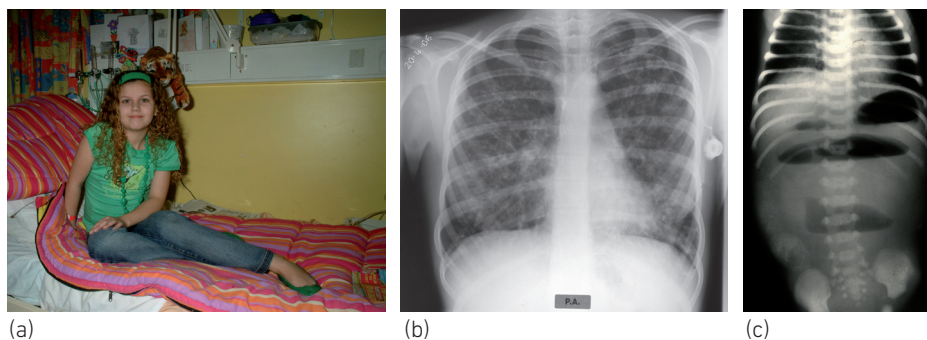


Figure 1.2 – Cystic fibrosis.

(a) The outlook for cystic fibrosis patients has improved over the years but they still need frequent hospital admissions, physiotherapy and constant medication. (b) Chest X-ray of lungs of cystic fibrosis patient. (c) Erect abdominal film of newborn with meconium ileus showing multiple fluid levels. Photos (a) and (b) courtesy of Dr Tim David, Royal Manchester Children's Hospital.

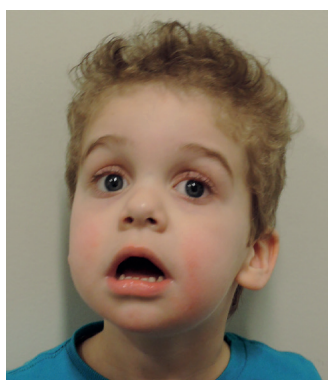
The pleiotropic* effects of cystic fibrosis

- **Lung** Abnormal mucus leading to infection and lung damage
- **Gastrointestinal tract** Meconium ileus
Distal intestinal obstruction
Rectal prolapse
- **Pancreas** Exocrine pancreas dysfunction
Malabsorption
- **Hepatobiliary tract** Biliary cirrhosis
Gallstones
- **Sweat glands** Elevated chloride and sodium in sweat
- **Reproductive tract** Congenital bilateral absence of vas deferens (CBAVD)
Thick cervical secretions

*Pleiotropic = having many effects.

CASE 3 KOWALSKI FAMILY

- Karol, first son of Kamil and Klaudia
- Developmental delay, hypotonic, severe intellectual disability



3 10 67 102 134 155 395

Karol was the first child born to his parents Kamil and Klaudia. His parents were worried about him from when he was only a few months old because all his milestones seemed slower than their friends' children. In particular, he seemed hypotonic (floppy) and was difficult to wean from milk to more solid food. However, there was nothing specific noted when he had a check-up at the clinic, although the doctor did comment that it was unusual that he had such long eyelashes and a rather hairy body whilst the hair on his head was very sparse. As time went on Kamil and Klaudia became even more concerned because Karol didn't walk until he was 30 months old and was extremely slow with his speech: he still had no clear words at over 3 years of age. He was referred to a pediatrician but, whilst waiting for an appointment, he had a seizure and was admitted to hospital. The pediatrician organized several tests and at first suspected he might have an enzyme deficiency because his face appeared rather coarse, he had a lot of body hair and his fingers were short with small nails. All the tests were normal, as was the microarray test undertaken to look for chromosomal abnormalities. An MRI (magnetic resonance imaging) scan of his brain showed that his corpus callosum was partially absent. Formal developmental assessment confirmed he had intellectual disability in the moderate to severe range.

Figure 1.3 – Severe intellectual disability case.

Despite his slightly unusual appearance, Karol was grossly normal and there was nothing in his phenotype that would suggest either an environmental cause or a specific genetic syndrome. Photo courtesy of Dr Gijs Santen, Leiden.

CASE 4 **DAVIES FAMILY**

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy

4

11

68

98

156

285

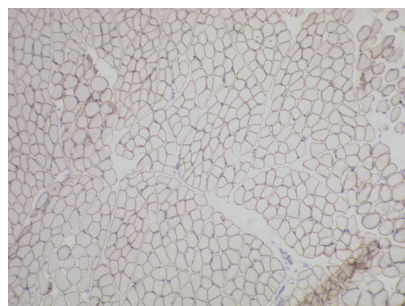
315

395

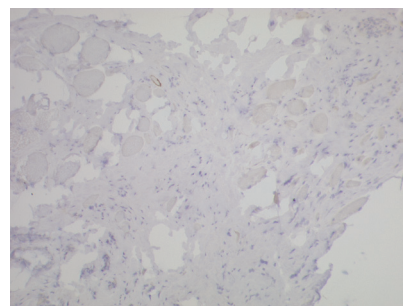
Martin is the first son born to Judith and Robert Davies; they already have two daughters Lisa and Jessica. The girls were very quick with their motor milestones and both were walking just before their first birthdays. Martin seemed slower in many ways and his mum thought it was just that he was a boy. However, when he still wasn't walking at 18 months she asked the doctor at the clinic for advice. The doctor didn't find anything when he examined Martin and arranged for another assessment in 6 months' time. Meanwhile Martin did take his first steps at 20 months but the doctor, at the next appointment, noticed Martin was very clumsy and that when he got up from the floor he had to hold on to a chair or push himself up by propping his hands on his legs. Both the doctor and Judith were very worried at this point because they knew that there was a family history of muscular dystrophy, and although there are many reasons why a little boy might be slow to walk and clumsy, these can also be early signs of that disease. They agreed that Martin needed to be referred to a neurologist, and Judith and Robert needed to see a geneticist.



(a)



(b)



(c)

Figure 1.4 – Duchenne muscular dystrophy.

(a) Affected boys stand up by bracing their arms against their legs (Gower's maneuver) because their proximal leg muscles are weak. (b) and (c) Muscle histology (Gomori trichrome stain). Normal muscle (b) shows a regular architecture of cells with dystrophin (brown stain) on all the outer membranes. (c) Shows muscle from a 10-year-old affected boy. Note the disorganization, invasion by fibrous tissue and complete absence of dystrophin. Histology photos courtesy of Dr Richard Charlton, Newcastle upon Tyne.

CASE 5 **ELLIOT FAMILY**

- Baby girl Elizabeth, parents Elmer and Ellen
- Multiple congenital abnormalities

4

12

43

68

100

395

Elmer and Ellen decided they wanted to start a family soon after they came back from their honeymoon in Jamaica where Elmer's parents were born. It wasn't long before Ellen became pregnant but 8 weeks after her last period she started bleeding and after that the pregnancy test was negative. The next pregnancy 5 months later seemed to go well at first but at 30 weeks' gestation a scan showed the baby to be small. This didn't worry the couple because Ellen was petite herself. However, when Elizabeth was born at 37 weeks' gestation it was clear that all was not well. Elizabeth was overall a small baby (her



Figure 1.5 – A baby with multiple congenital abnormalities and dysmorphic features.

length and weight were on the 3rd centiles) but her head circumference was significantly below the 3rd centile. She needed tube feeding and seemed to get breathless very easily. A heart murmur was detected and an echocardiogram showed she had a ventricular septal defect and a narrowing of her aortic valve. The pediatrician asked the geneticist to see Elizabeth and her parents and advise which investigations might be helpful. Ellen told the geneticist her sister had had two early miscarriages, and her aunt who lives in Trinidad had a baby who died with a congenital heart defect and a stillborn baby before she had two healthy children.

CASE 6 FLETCHER FAMILY

- Frank, aged 22, with increasingly blurred vision
- Family history of visual problems

5 13 69 130 157 395

Frank was an electrician who, at 22 years old, had just finished his college course. He enjoyed going out with his friends and he tended to drink quite a lot of alcohol at weekends. He had always had good eyesight but one week he noticed that his vision was blurred and the colors of the wires he was working on seemed paler than usual. When this didn't improve he went to the optician who noticed changes in Frank's retina and made an urgent referral to the Eye Hospital. There they found he had disk swelling (pseudoeedema of the nerve fiber layer), and increased tortuosity of the retinal vessels. Gradually over the next few months his central vision became much worse and he had to give up his job. His mother, Freda, was a healthy woman but her only brother had been registered blind since 28 years of age with what she remembered to be optic atrophy. Her sister Doreen also has serious visual difficulties which came on when she was around 45 years old. The ophthalmologist referred Frank to the genetic clinic because of this family history.

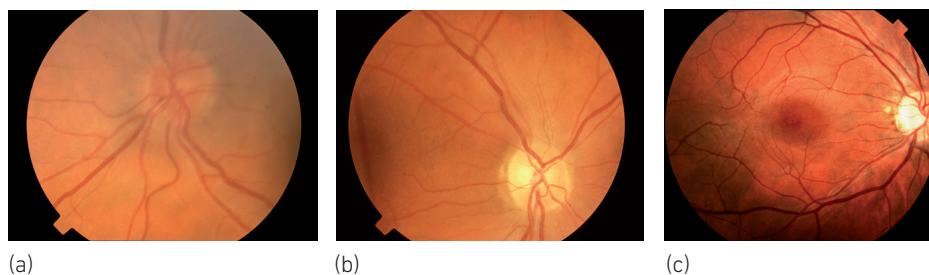


Figure 1.6 – Leber hereditary optic neuropathy.

(a) Optic disc 3 weeks after patient noted reduction in vision; note hyperemia of disc with blurred margins. (b) Retina of uncle; vision lost several years previously; note pallor of disc particularly temporally corresponding to optic atrophy. (c) Normal retina. Photos courtesy of Mr Graeme Black, St Mary's Hospital, Manchester.

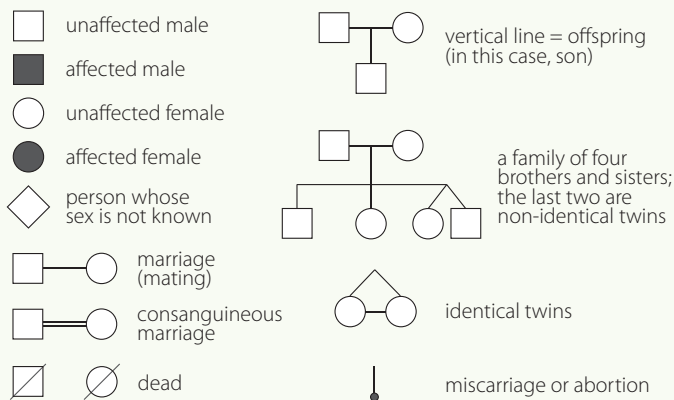
1.2. Science toolkit

The first thing to do with any of these patients is to take a family history. Even when the referral letter gives a family history, it is important for the geneticist to go carefully through, following the protocol in Box 1.2. The family history can give important clues about the genetic diagnosis; it also forms a necessary background for genetic diagnosis and counseling.

How to take a family history and draw a pedigree

Take a special pedigree proforma or a sheet of blank paper and rule four lines across the long dimension. Start with the *consultand* (the person who is referred to the clinic) in the middle of the next-to-bottom line (or the bottom line for a child). Draw in the appropriate symbol (see below) and note his or her name, date of birth and any relevant clinical features. Next record details of the spouse(s) or partners, if any, then proceed systematically through their children, parents, and brothers and sisters (sibs). For each sib record their partners and children. Next ask about the sibs of each parent and their partners and children. Finally document all four grandparents and their sibs. Unless there is some reason to do so, it should not be necessary to go beyond that. Be careful to ask systematically about each person, and for each person note their name, date of birth, if dead the date and cause of death, and any relevant clinical information. Ask about miscarriages or reproductive problems and, if appropriate, about the place of origin of each person. Even if the consultands are convinced that they know which side of the family is the source of the problem, it is wise to collect full details of both sides – family myths can be very misleading. Many institutions use computerized systems that take you through a series of questions similar to those above, and draw the pedigree for you.

Build up the pedigree as you go along, keeping each generation on one horizontal line. List sibs in the order of their birth if this can be done without too many lines crossing on the drawing. If the pedigree is at all complicated you will probably need to re-draw it neatly later. You can use the pedigrees in this chapter as models. The appropriate symbols to use are shown in Box figure 1.1.



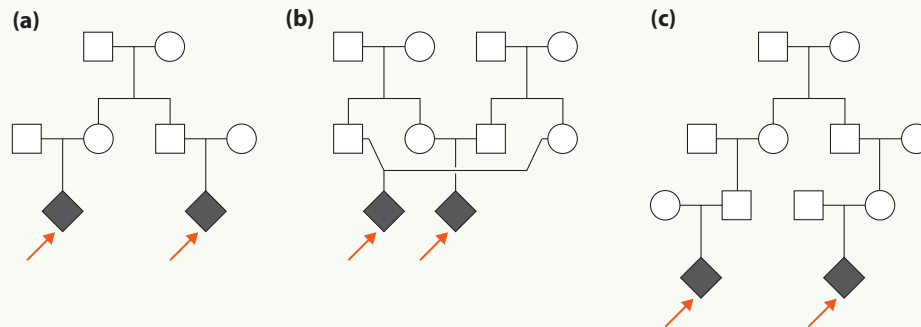
Box figure 1.1 – Pedigree symbols.

Describing relationships

Sibs (or **siblings**) means brothers or sisters, regardless of sex

Cousins: Jack and Jill are **first cousins** (Box figure 1.2) if one of Jack's parents is a sib of one of Jill's parents. If both of Jack's parents are sibs of both of Jill's parents, Jack and Jill are **double first cousins** (Box figure 1.2b). They are **second cousins** (Box figure 1.2c) if one of Jack's parents is first cousin of one of Jill's parents.

In some cultures the word 'cousin' is used much more loosely to mean 'kinsman'. For genetic purposes it should be used in the strict sense defined above. More complex relationships are best defined by drawing or describing the pedigree details. We will see in *Chapter 9* how to calculate exact degrees of relationship and inbreeding.



Box figure 1.2 – Relationships.

The arrowed individuals are (a) first cousins, (b) double first cousins, (c) second cousins.

BOX 1.2 – continued

1.3. Investigations of patients

In each of our cases the initial investigation was to construct a detailed pedigree. The pedigrees below illustrate some of the typical features of autosomal dominant (Huntington disease), autosomal recessive (cystic fibrosis) and X-linked recessive (Martin's muscular dystrophy) conditions. The other cases (Kowalski, Elliot and Fletcher families) raise issues of interpretation that will be considered towards the end of this chapter.

How John Ashton came to the genetic clinic and issues the geneticist considered

Letter from John's family doctor to the genetics service:

'Would you arrange an appointment for John Ashton aged 28 years who recently married? John appears healthy but he is extremely concerned because the psychiatrist investigating his father's depression suggested there may be an inherited cause. Alfred, John's father, apparently has some unusual movements and is very forgetful, and several family members in previous generations developed dementia in their 50s. John and his wife have recently taken out a mortgage and were about to start a family, but understandably want to clarify whether there are any implications for John or any children they may have'.

The genetic counselor's notes in preparation for seeing John:

Information needed at/before the clinic appointment

- Details of Alfred's clinical state, and results of investigations from the psychiatrist (need to ensure permission from Alfred)
- Family pedigree noting members with dementia and ages of onset

BOX 1.3

Consider differential diagnosis

- Huntington's disease (HD)
- Other inherited dementias
- Parkinson's disease in Alfred with no relevance to family history of dementia

Establishing a diagnosis in Alfred

- When a patient has clinical symptoms which suggest HD, a DNA **diagnostic test** can be done (see *Chapter 4* for a description of the test). This is specific for HD and if positive confirms the clinical diagnosis. If negative, then HD is excluded and other causes need to be considered.

Establishing risks for John

- If HD is confirmed in Alfred then John can be offered a **predictive test**. This is a much more ethically sensitive situation because John will have a 1 in 2 chance of carrying the gene mutation and, if he does, he will develop the disease. He would also have a 1 in 2 chance of passing the affected gene to a child. John and his wife will require sensitive genetic counseling and information about the implications of a positive test, and the options for tests in pregnancy.

Hopes for treatments

- John and his wife will also likely want to know what research is being done and whether there are any promising treatments in the pipeline. The Genetics Center will be able to keep them informed.

CASE 1 ASHTON FAMILY

- John, healthy 28-year-old son of Alfred Ashton
- Family history of
 - ? Huntington disease
- Autosomal dominant inheritance

1

8

67

103

153

395

John's mother comes to the clinic with John to help construct a family tree. As well as John, she has a daughter who has two young sons. Alfred Ashton's father Frederick (John's grandfather) was killed in an industrial accident aged 38 years but Alfred's mother is alive and well at 82 years. Alfred's paternal grandmother became demented in her 50s and needed institutional care, as did one of her two brothers and her only sister. Although the family have lost touch with the extended family, since they became aware of the diagnosis in John's father, they have heard that other people in the family have Huntington disease. They are very worried about Alfred's sister who lives in Australia and who has been demonstrating jerky movements like her brother.

This pedigree shows autosomal dominant inheritance. People with Huntington disease and other serious dominant conditions are normally heterozygous – that is, they have one copy of the abnormal gene, inherited from an affected parent, and one copy of the normal gene, inherited from the other parent. The child of somebody affected by Huntington disease has a 1 in 2 chance of having inherited the Huntington disease gene. If they do inherit the gene, they will inevitably develop Huntington disease unless they die of something else first, like Alfred's father did. The disease may develop at any age from childhood to 70+, but most usually when the person is in their 40s.

Severe late-onset genetic conditions like Huntington disease cause agonizing dilemmas for at-risk people. John has just got married and bought a house. His wife and he are thinking of starting a family. Although he shows no symptoms of the disease, if he does carry the Huntington disease gene he will inevitably develop this severe disease later in

his life, and any child he has would be at a 1 in 2 risk. A DNA test is available that could tell John definitively whether or not he carries the mutation. Deciding whether or not to take the test will be one of the biggest decisions of his life. In the UK about 80% of people in this situation opt, after full counseling, not to take the test. Genetic counselors must always respect a patient's right **not** to know, in this or any other situation, if that is their choice. Before embarking on this whole process, it is crucial to confirm that the family disease really is Huntington disease and not either an unrelated neurodegenerative condition or one of the rare autosomal dominant diseases that can resemble Huntington disease. The DNA test is specific for Huntington disease, and the result would be irrelevant if the disease in this family were in fact something else. In this case the psychiatrist is confident of the diagnosis for Alfred, but a test of Alfred's DNA (see *Chapter 4* for a description of the type of test) is used to make sure. Note that testing somebody who is already ill constitutes a **diagnostic test** which is not as ethically sensitive as a **predictive test** on a healthy person like John, although a positive result in Alfred would confirm the poor prognosis.

The counselor told John about a new drug, IONIS-HTT, that was showing promise in clinical trials (Fischbeck and Wexler, 2019). It lowered the level of the mutant HTT protein in cerebrospinal fluid; whether this translated into clinical improvement remained to be seen. John agreed that, if the drug fulfilled its clinical promise, then it would be in his interest to take the predictive test; however, at age 28 he felt he could delay that decision for several more years and wait to see the results of further clinical trials currently in progress.

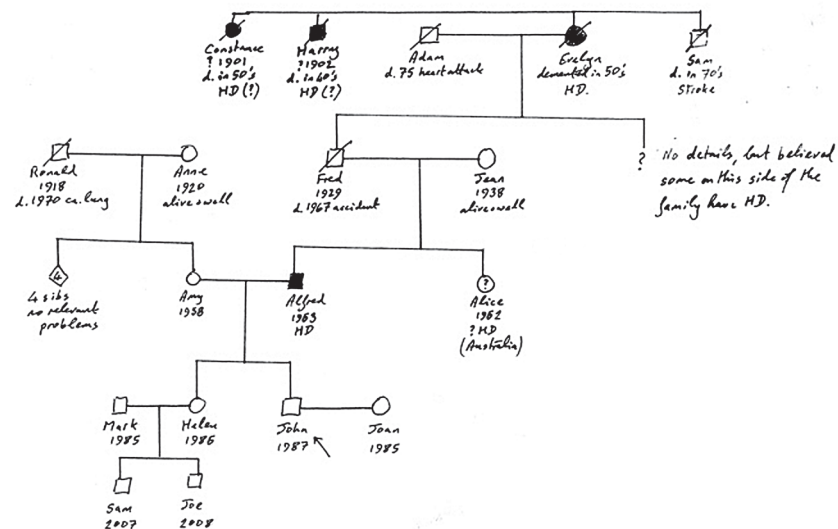


Figure 1.7 – Pedigree of John Ashton's family.

This is shown as it might be recorded in the clinic but because the cases and families in this book are fictional, all subsequent pedigrees show only information that is relevant to following the case and understanding the genetics.

CASE 2 BROWN FAMILY

- Baby Joanne, recurrent infections, poor growth
- Sweat test confirms she has cystic fibrosis
- Autosomal recessive inheritance

2

10

67

132

154

313

395

The pedigree shows that there is no family history of cystic fibrosis or other evident genetic problems. “How can it be genetic, when nobody in either family has ever had it?” Pauline asked. In fact this pedigree is typical of the way autosomal recessive diseases often present in societies where consanguinity is not common. The affected child is usually the only affected case born to a non-consanguineous couple with no relevant family history. Thus the pedigree gives no clue that the condition is genetic. Identifying the cause usually starts with making a clinical diagnosis. Sometimes the condition is so unmistakable, and the genetics so unambiguous, that a clinical diagnosis provides an adequate degree of certainty. More often the clinical diagnosis is really a hypothesis, more or less plausible, but requiring confirmation by a DNA or biochemical test where this is possible.

In most cases of cystic fibrosis the clinical history and a positive sweat test (showing characteristically raised chloride) make the diagnosis fairly secure. Genetically, cystic fibrosis is always autosomal recessive, and always caused by mutations in the *CFTR* gene on chromosome 7 (see Chapter 3). A molecular test to demonstrate the mutation is needed for advising relatives about their carrier status, for prenatal testing, and to confirm the diagnosis in atypical cases. Some newly developed drugs that can ameliorate the symptoms of the disease also require the precise mutations to be identified, because they target *CFTR* genes carrying certain specific mutations, and are effective against only those specific variants.

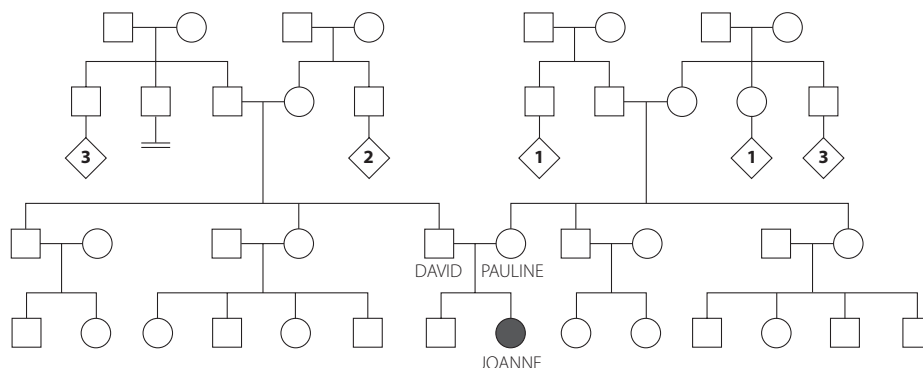


Figure 1.8 – Pedigree of Joanne Brown's family.

Note the complete absence of any family history of cystic fibrosis. Autosomal recessive conditions commonly present as a single isolated case. Numbers inside pedigree symbols specify the number of offspring, sex not shown.

CASE 3 KOWALSKI FAMILY

- Karol, first son of Kamil and Klaudia
- Developmental delay, hypotonic, severe intellectual disability
- Difficulties of genetic testing in such cases

3

10

67

102

134

155

395

As with the family of Joanne Brown, described above, the family history was completely negative. Unlike Joanne Brown, Karol's clinical examination gave no clue as to the possible cause of his severe problems. The pregnancy and birth had been entirely uncomplicated – there was no suggestion of infection in pregnancy, birth asphyxia or any other environmental explanation for Karol's problems. When he was admitted to hospital following his first seizure the pediatrician organized several tests, as described above. All those tests were negative. Formal developmental assessment confirmed he had intellectual disability in the moderate to severe range.

In the absence of alternative explanations, it seemed likely that his problems were genetic. His phenotype – hypotonia, developmental delay and intellectual disability – was severe but non-specific: there were no distinctive features that would suggest a specific genetic diagnosis. Our brain is the most complicated part of our body, and so there are innumerable ways in which it might go wrong and produce nonsyndromic intellectual disability. Perhaps there was a *de novo* chromosomal abnormality, although the lack of syndromic features made this less likely; alternatively, he might have an autosomal recessive condition, like Joanne Brown, or a *de novo* mutation producing an autosomal dominant or X-linked condition. Assuming no chromosomal cause could be found (see *Chapter 2*), Karol's case is typical of a situation where until recently no genetic test could be offered because there was no specific genetic hypothesis to test. Advances in DNA sequencing technology have finally made such cases tractable, as described in *Chapter 5*.

CASE 4 DAVIES FAMILY

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy
- X-linked recessive inheritance

4

11

68

98

156

285

315

395

The pedigree shows that Judith's aunt (her mother's sister) had had two boys who died in their teens having had a progressive muscle disease. There were no other affected relatives in the family. The pedigree is consistent with X-linked recessive inheritance. If Martin does indeed have the same condition as his two dead relatives, then X-linked inheritance is highly likely, but at this stage this is still not certain.

It was crucial to check the exact diagnosis of these two boys. There are many different degenerative muscle diseases. While these might have roughly similar implications for management of the patient, the implications for the wider family depend critically on a precise genetic diagnosis to establish the mode of inheritance. The geneticist will request the case notes, paying particular attention to the detailed course of the boys' decline and to any reports of muscle histology. Given the dates of their deaths it is unlikely that any DNA test results would be available. As described in *Chapter 4*, testing Martin's DNA will play a central part in the investigations.

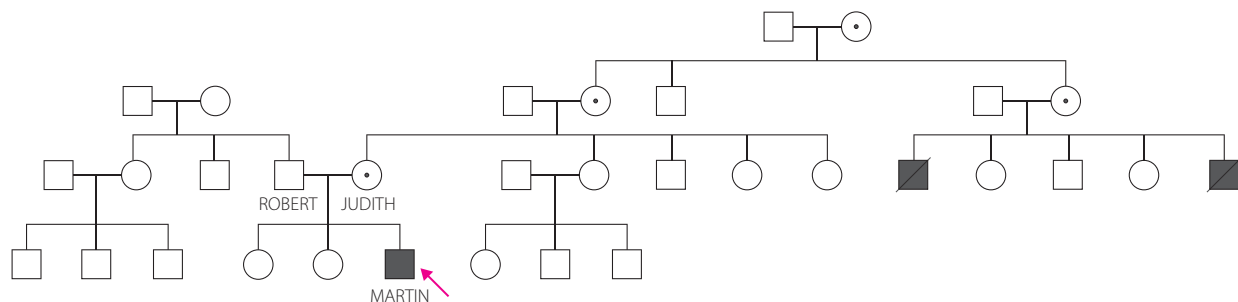


Figure 1.9 – Pedigree of Martin Davies's family.

Assuming this is X-linked muscular dystrophy, the women marked with dots are **obligate carriers** of the disease gene – that is, they must be carriers because they have both parents and offspring who are affected or carriers. Other females (e.g. the sisters of affected boys) may or may not be carriers.

CASE 5 ELLIOT FAMILY

- Baby girl Elizabeth, parents Elmer and Ellen
- Multiple congenital abnormalities
- Family history of reproductive problems
- ? Chromosome abnormality

4

12

43

68

100

395

The family history of miscarriages, stillbirth and liveborn infants with malformations in several generations might suggest an autosomal dominant condition with reduced penetrance (see *Section 1.4* for an explanation of penetrance). However, clinical experience suggests the most likely explanation is a balanced chromosome structural abnormality. The next step is to perform chromosome analysis on the parents and the surviving abnormal child (see *Chapter 2*).

After the birth of a child with malformations or other problems, parents experience a whole range of emotions including shock, anxiety, denial and confusion. They often also experience a sense of loss for the normal baby they hoped to have. Naturally other family members have their own responses to the difficult situation and tensions can be raised further, particularly if one 'side' blames the other. The role of the pediatric team caring for the baby is to keep the family fully informed about their findings in the baby and about tests and consultations that are being arranged. The roles of the clinical geneticist are to help establish a diagnosis as soon as possible and to convey complex results to the parents in a comprehensible way, and also to explain what they mean for the baby's future. Genetic counseling about recurrence risks and implications for the extended family is often arranged later after urgent clinical management issues have been dealt with.

Over several sessions the counselor built up the pedigree shown below. The poor reproductive outcomes (miscarriages or abnormal babies) in this family appear to show an autosomal dominant pattern. As explained below, this is not incompatible with the suspicion that the problems were caused by a chromosomal abnormality. Blood was therefore taken from the baby and both parents for chromosome analysis. The results and implications are discussed in *Chapter 2*.

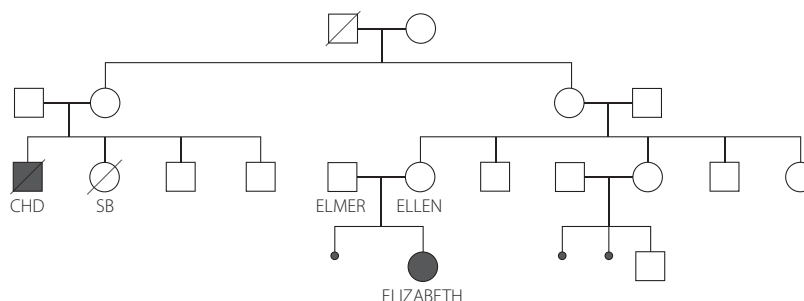


Figure 1.10 – Pedigree of the Elliot family.

This shows the family history of reproductive problems. SB denotes stillbirth and CHD congenital heart disease.

CASE 6 FLETCHER FAMILY

- Frank, aged 22, with increasingly blurred vision
- Family history of visual problems
- Possible mitochondrial inheritance
- ? Leber hereditary optic neuropathy

5

13

69

130

157

395

Frank attended the genetic clinic with his mother. He was still very shocked after having the diagnosis and poor prognosis explained to him by the ophthalmologist. He hadn't even started to come to terms with the major changes this would bring in his life, including losing his job and not being able to drive. It had, however, begun to occur to him that because of the family history his own children might be at risk. When he came to the clinic he very much wanted to discuss this aspect of things since his girlfriend and he were very worried.

The clinical geneticist first of all drew up a detailed family tree with the help of Frank's mother Freda. Her brother, Derek, was registered blind and had been diagnosed as having optic atrophy. Like Frank, he had first noticed problems in his 20s; encouragingly he had attended college after his diagnosis and now worked for a large company in their telephone sales department. He had also told Frank about the national organization for people with visual disabilities and about all the help and advice they could offer. Freda's sister Doreen didn't develop any problems until she was in her 40s and her visual deterioration was slower, although still affecting her central vision. At a recent check-up she had been found to have an unusual heart rhythm and had requested Freda to find out in the clinic if that was linked to the eye problems.

The following clues led the geneticist to a hypothesis about the condition in Frank's family:

- the nature of the eye problems
- the rapid progression of symptoms in the affected males
- the later onset and slightly milder disease in Doreen
- Doreen's heart rhythm problems

Based on these observations, the geneticist thought the condition could be Leber hereditary optic neuropathy (LHON). LHON (OMIM 535000) has a wide range of symptoms and ages of onset. It is normally caused by a mutation in mitochondrial DNA (mtDNA – see *Chapter 3*), although it has recently been shown that in rare cases the cause can be mutation of a gene on chromosome 7, giving an autosomal recessive inheritance pattern. The geneticist therefore arranged for blood samples to be taken from Frank and Freda. If this type of inheritance is confirmed then it is good news for any children Frank and his girlfriend may have, since a male does not pass on his mtDNA to his children.

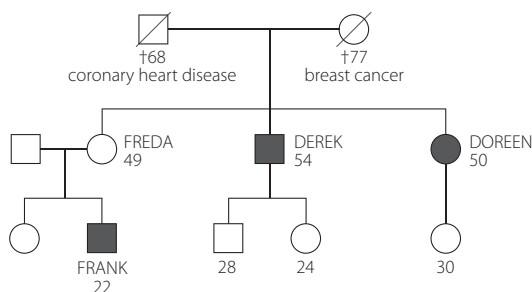


Figure 1.11 – Pedigree of Frank Fletcher.

This pedigree was difficult to interpret. At first sight the geneticist considered X-linked recessive inheritance because of two similarly affected males linked through a normal female. However, Doreen was also affected, which made X-linked inheritance less likely. This nevertheless remained a possibility because sometimes in such families there are 'manifesting female carriers' with milder problems than affected males (see *Chapter 11*). Other types of inheritance such as X-linked dominant and mitochondrial inheritance have to be considered when affected people are linked through the female line.

1.4. Going deeper...

The art of pedigree interpretation

In laboratory animals like fruit flies or mice, the mode of inheritance of a character can always be established beyond doubt by breeding experiments. With humans we have to take pedigrees as we find them, and they are rarely large enough to define the mode of inheritance unambiguously. For research purposes, collections of pedigrees can be analyzed statistically, using the tools of segregation analysis to work out the most likely mode of inheritance. In the clinic, pedigree interpretation is an art as much as a science. It is more a matter of forming hypotheses for subsequent investigation. The hypothesis might involve any of the following causes:

- a chromosomal abnormality
- an autosomal dominant condition, fully or partially penetrant
- an autosomal recessive condition
- an X-linked condition, dominant or recessive
- a condition caused by a defect in the mitochondrial DNA
- a multifactorial condition
- a non-genetic cause

Box 1.4 summarizes the features of the main mendelian pedigree patterns. To form a hypothesis, individual pedigrees are checked to see what mode of inheritance gives the best fit. The initial hypothesis is based on two questions:

- Does the pedigree show that affected people have at least one affected parent? If the answer is yes, the condition is most likely dominant; if no, it is most likely recessive. A dominant condition is manifest in anybody who carries the relevant gene. An affected person must have inherited the gene from one parent, who should therefore also be affected. Bear in mind, however, that many cases of severe dominant (or X-linked) conditions are new (*'de novo'*) mutations, with no previous family history.
- Are there any sex effects? For example, does the condition affect both sexes, and can it be passed on by a parent of either sex to a child of either sex? If there are no sex effects, the condition is most likely autosomal. Look especially for male-to-male transmission: this is a powerful pointer against X-linked inheritance because a father should never transmit his X chromosome to a son. If there appear to be sex effects, then it may be X-linked, although as explained below, there can be other reasons for a sex bias. Unless the pedigree is exceptionally large, apparent effects may be just random fluctuations.

Having arrived at an initial hypothesis, the next step is to test it by writing in genotypes, as in *Figure 1.12*. Given sufficient coincidences (new mutations, carriers of the disease happening to marry into the family, etc.) almost any pedigree can be made to fit almost any mode of inheritance. The most likely mode of inheritance is the one that requires the least number of such coincidences – preferably none. If you require coincidences to make your initial hypothesis fit, then you should see if an alternative hypothesis gives a better fit. Pedigrees in student examination papers should always have one 'right' answer.

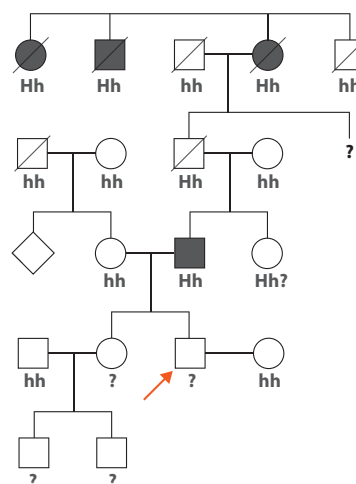


Figure 1.12 – The Huntington disease pedigree of Figure 1.7 with genotypes written in.

By convention, upper case is used for the allele determining the dominant character and lower case for the allele determining the recessive character. In this example the disease allele is dominant. Note that affected people in this family are heterozygotes. This is the normal situation with human dominant conditions. Homozygotes are usually extremely rare or unknown; when they are recorded, they are often far more seriously affected than heterozygotes. Nevertheless, such a condition is correctly described as dominant (not semi-dominant). Dominance and recessiveness are properties of phenotypes (characters, diseases...) not genes. A character is dominant if it is manifest in the heterozygote, and this is the case for Huntington disease. As it happens, Huntington disease is one of the rare human cases where homozygotes are known to exist and to be phenotypically indistinguishable from heterozygotes.

Real life is not always so kind: many real pedigrees are either too small to allow clear interpretation or do not fit any of the standard mendelian patterns. Often, as in the cases of **Joanne Brown – Case 2** and **Karol Kowalski – Case 3**, the affected individual is the only case in the family and there is no pattern to explore, although such pedigrees are consistent with autosomal recessive inheritance or a *de novo* mutation to an autosomal dominant or X-linked condition. The condition may be multifactorial, or other evidence may suggest a chromosomal abnormality or a non-genetic cause. Possible confirmatory investigations include:

- clinical identification of a syndrome – but bear in mind that some clinically defined syndromes can show more than one mode of inheritance
- karyotyping or molecular cytogenetic analysis (see *Chapters 2 and 4*)
- sequencing of candidate genes or the whole exome for mutations (see *Chapter 5*)
- checking for biochemical abnormalities, including abnormalities of mitochondrial function (see *Chapter 10*)
- checking for skewed X-inactivation (see *Chapter 11*)

Summary of modes of inheritance of monogenic characters

A mendelian pattern will be seen whenever a phenotype is caused by something at a single fixed chromosomal location – whether that 'something' is a classical gene or a chromosomal abnormality.

Autosomal dominant:

- a vertical pedigree pattern, with multiple generations affected
- each affected person normally has one affected parent (note, however, that people with severe dominant conditions often have fresh *de novo* mutations and there is no previous family history)
- each child of an affected person has a 1 in 2 chance of being affected
- males and females are equally affected and equally likely to pass the condition on

Autosomal recessive:

- a horizontal pedigree pattern, with one or more sibs affected; often only a single affected case
- parents and children of affected people are normally unaffected
- each subsequent sib of an affected child has a 1 in 4 chance of being affected
- males and females are equally affected
- affected children are sometimes the product of consanguineous marriages. In families with multiple consanguineous marriages, affected individuals may be seen in several generations

X-linked recessive:

- a 'knight's move' pedigree pattern – affected boys may have affected maternal uncles
- parents and children of affected people are normally unaffected. Never transmitted from father to son
- affects mainly males: females can be carriers, and affected males in a pedigree are linked through females, not through unaffected males
- subsequent brothers of affected boys have a 1 in 2 risk of being affected; sisters are not affected but have a 1 in 2 risk of being carriers
(As with autosomal dominant conditions, *de novo* mutations are frequent)

X-linked dominant:

- features very similar to autosomal dominant pedigrees, except that all daughters and no sons of an affected father are affected
- condition is often milder and more variable in females than in males

Y-linked:

- a vertical pedigree pattern
- all sons of an affected father are affected
- affects only males

Mitochondrial:

- a vertical pedigree pattern
- children of affected men are never affected
- all children of an affected woman may be affected, but mitochondrial conditions are typically extremely variable even within a family

Penetrance and expressivity – pitfalls in interpretation and counseling

Many human conditions show a mainly autosomal dominant pedigree pattern, but occasionally skip a generation: that is, an unaffected person who has an affected parent produces one or more affected children. Thus occasionally a person can carry the relevant gene but not manifest the condition. Such occurrences are described as non-penetrance. The **penetrance** of a character is defined as the proportion of people with the relevant genotype who show the character.

Penetrance is a property of both the gene and the character. Different syndromes show characteristically different penetrances, but different features of a syndrome can also have different penetrances. Penetrance can also be age-related, as with Huntington disease.

Non-penetrance is a pitfall in both pedigree interpretation and counseling. *Figure 1.13* shows an example. Individual III-11 must carry the disease gene despite being unaffected. Before starting a family he might have requested counseling to find out his risk of having a child with the family disease. The counselor would have had to know the penetrance of that particular condition in order to be able to calculate the risk that, despite being phenotypically normal, he might be a non-penetrant gene carrier. Fortunately in many cases a molecular test would be available that could give a definitive answer.

Reduced penetrance is a nuisance for counselors, but there is no mystery about why it happens. The surprise is really that some conditions show 100% penetrance. For such a condition, if you have the relevant gene then you will inevitably manifest the condition, absolutely regardless of all your other genes, your environment and your lifestyle. For many conditions, having the gene means that you are highly likely to manifest the condition, but occasionally a lucky combination of other genes and/or environmental factors will save you.

Genotype-phenotype correlations form a complete spectrum, from 100% penetrance down to very low penetrance (*Figure 1.14*). Characters at the high-penetrance end of the spectrum are usefully described as mendelian, while those at the low-penetrance end would be called multifactorial (see *Chapter 13*). There is no hard and fast rule where the changeover occurs. Reduced penetrance can occur with characters showing recessive inheritance – hemochromatosis (OMIM 235200) is an example – but it would not be obvious from the pedigree, so in practice incomplete penetrance is largely a problem of dominant conditions.

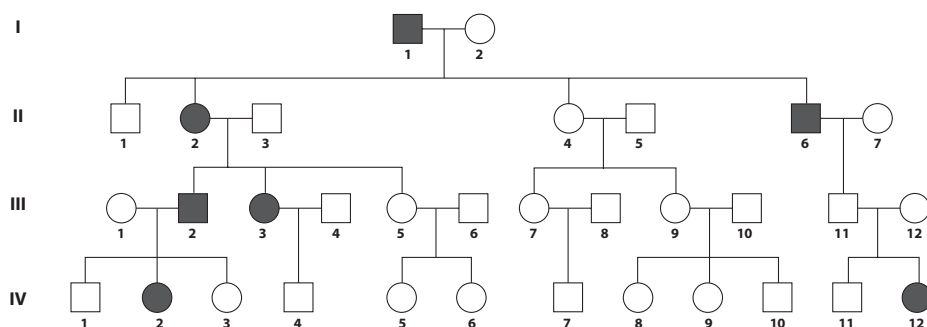


Figure 1.13 – Pedigree of an autosomal dominant condition with reduced penetrance.

The condition is non-penetrant in individual III-11. Other unaffected individuals in generation IV who have an affected parent might also be non-penetrant gene carriers.

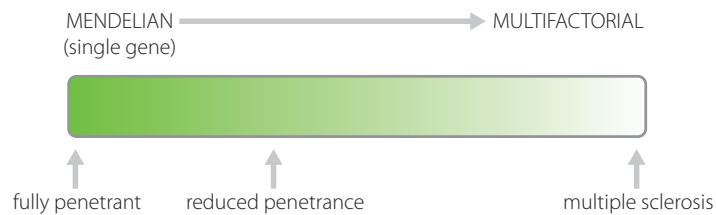


Figure 1.14 – Continuum of penetrance.

There is a continuum of penetrance from fully penetrant conditions, where other genes and environmental factors have no effect, through to low-penetrance genes that simply play a small part, along with other genetic and environmental factors, in determining a person's susceptibility to a disease. Multiple sclerosis is used as an example of a multifactorial condition where genetic factors play a major part in determining susceptibility, but current research suggests that each individual factor has very low penetrance (see *Chapter 13*).

Genetic conditions often show **variable expression**. That is, the full-blown condition involves a number of features, but many affected people show only some of those features, or may show a certain feature to differing degrees. Type 1 neurofibromatosis (NF1; see *Disease box 1*) is an example of a common autosomal dominant condition that is very variable. We could say that each feature of the syndrome has its own characteristic penetrance, or we could say that the syndrome as a whole shows variable expressivity. Either way, this is a reflection of the fact that genes do not act in isolation, but against a background of innumerable other genes and a variable environment.

Rarer modes of inheritance

X-linked dominant inheritance

For males, X-linked diseases are neither dominant nor recessive, because dominance and recessiveness are properties of heterozygotes, and males have only a single X chromosome. As explained further in *Chapter 11*, the phenomenon of X-inactivation means that even for heterozygous females, dominance and recessiveness are not as clear-cut for X-linked as for autosomal conditions. Most X-linked diseases seldom affect heterozygous females significantly, and so are described as recessive. Occasional X-linked conditions do commonly affect heterozygotes badly, and so are dominant. An example is X-linked hypophosphatemia (OMIM 307800; an inability of the kidneys to retain phosphate, leading to vitamin D-resistant rickets). At first sight the pedigree pattern (*Figure 1.15*) might well be interpreted as autosomal dominant. Consistent with this interpretation, there is a vertical pedigree pattern and on average 50% of the children of an affected person are affected. However, a male passes his X chromosome to all his daughters but none of his sons. Thus all the daughters but none of the sons of an affected father are affected. Because of X-inactivation, X-linked dominant conditions tend to be milder and more variable in females than in males.

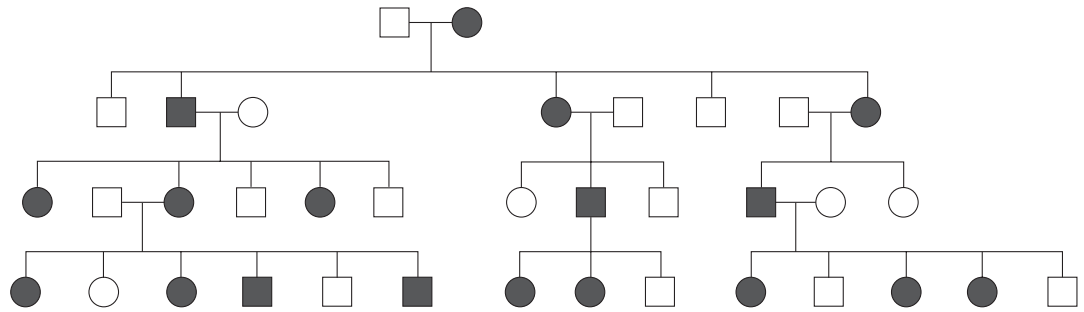


Figure 1.15 – Pedigree of an X-linked dominant condition.

Although heterozygous females are affected, such conditions are usually milder and more variable in females than in males.

Other causes of sex bias

A sex bias is not a reliable indicator of X-linkage. An autosomal condition may affect just one sex for anatomical or physiological reasons. For example, ovarian cancer may be caused by mutations in the *BRCA1* gene on chromosome 17, but obviously affects only females. Such conditions are called **sex-limited**. Sometimes a condition may be lethal in one sex but not the other. If this happens before birth, the result is a condition seen in only one sex. This is most likely with X-linked dominant conditions where heterozygous females survive but affected males die *in utero*. An example is Rett syndrome (OMIM 312750 – see *Disease box 11*): affected males normally miscarry early in pregnancy, so the classical syndrome is only seen in girls. A few affected males do in fact survive to birth, but their phenotype is so different that its cause was not recognized until molecular tests demonstrated that they had mutations in *MECP2*, which is the gene mutated in most cases of Rett syndrome.

Y-linked inheritance

For a condition determined by a gene on the Y chromosome the inheritance pattern would be simple and striking. It would affect only males, and all the sons of an affected man would be affected. However, the Y chromosome carries only about 50 genes, and since females manage perfectly well without any of them, none of these genes can be essential to life or general health. Y-linked genes are important for male sexual function, and abnormalities of the Y chromosome are a common cause of male infertility. Such abnormalities are clinically important, but because the resulting phenotype is infertility, they do not segregate in multi-generation pedigrees.

Mitochondrial inheritance

The problem in the **Fletcher family – Case 6** (Figure 1.11) illustrates an unusual mode of inheritance. All the previous cases have involved the chromosomal DNA in the cell nucleus. But, as we shall see in *Chapter 3*, mitochondria also contain DNA. Mutations in the mitochondrial DNA (mtDNA) are responsible for a few diseases, including the Fletcher case. Note four important points:

- (1) All the mitochondria of an embryo are derived from the egg, and none from the sperm. This gives rise to a matrilineal pattern of inheritance: mitochondrial mutations are passed on by the mother, but never by the father. Most conditions caused by mutations in mtDNA affect both sexes equally. Leber hereditary optic neuropathy is unusual in that, for unknown reasons, it affects mostly males.
- (2) Most of the components and functions of mitochondria are controlled by genes located on the chromosomes in the nucleus (see *Chapter 3*). Most diseases caused by malfunction of mitochondria therefore follow typical mendelian inheritance patterns. Only a very few mitochondrial diseases are caused by mutations in the mtDNA and follow the pattern of inheritance we see in the Fletcher case.
- (3) Because each cell contains many mitochondria, people can have a mix of normal and mutant mitochondria in each cell (**heteroplasmy**). The proportion can vary between tissues and in the same tissue over time. This helps explain why diseases caused by mitochondrial mutations typically have low penetrance and extremely variable expressivity.
- (4) Ova contain many mitochondria; therefore a heteroplasmic mother can have heteroplasmic children. This is in stark contrast to mosaicism for nuclear abnormalities (chromosomal or single gene), where either the normal or abnormal form can be passed on to a child, but not both (see below and *Chapter 2*).

Some further problems in pedigree interpretation

A condition is not necessarily genetic just because it is **congenital** (present at birth) or **familial** (tending to run in families). Huntington disease is an example of a condition that is genetic but not congenital, while many birth defects (congenital by definition) are caused by environmental teratogens (rubella or thalidomide, for example). Unless a familial condition shows a clear-cut mendelian pedigree pattern or a demonstrable chromosomal abnormality, it can be very difficult to decide what role genes play in its causation. Innumerable behavioral and occasional physical characters can be the result of shared family environment rather than genes. *Chapter 13* describes how these 'nature – nurture' problems can be approached. The converse is also true: a negative family history does not rule out a genetic cause for a problem. This is especially true of autosomal recessive conditions, as illustrated here by the cystic fibrosis case of the **Brown family – (Case 2)**, but it may also be seen with an autosomal dominant or X-linked condition where cases are often the result of new mutations, as with **Karol Kowalski – Case 3**. *Chapter 9* explains why new mutations are frequent with some genetic conditions and rare with others (see also *Disease box 1* in this chapter).

Mosaicism

A person whose body contains two or more genetically different cell lines is called a **mosaic**. As explained in *Chapter 2*, the process of mitosis should ensure that every cell in the body carries a complete and identical set of genes. Our discussion so far has tacitly assumed that if a person has a mutation, it is present in every cell of his body. This is true of inherited mutations, but what about freshly arising mutations? Mutations can happen to any cell at any time. Only the descendants of that cell will carry the mutation. Given the number of cells in the human body, and typical mutation rates of genes, it is clear that everybody will have the odd cell or small clone of cells carrying a mutation in almost any

gene you care to think of. Normally these odd rogue clones will have absolutely no effect. There are three circumstances in which mosaicism may be clinically important:

- (1) if the mutant cells have a tendency to grow and take over – see *Chapter 7*
- (2) if the mutation arose sufficiently early in embryonic development that the mutant line makes up a significant part of the whole body. The person may show features of the disease, maybe with a milder phenotype (if the product of the mutated gene is diffusible) or with a patchy distribution reflecting the distribution of mutant cells (if the gene product remains in the cell where it is produced). See *Disease box 6* for examples.
- (3) if the mutation affects the **germ line** (sperm or egg cells or their progenitors).

Germ-line mosaicism is a major source of uncertainty in genetic counseling. A person who by all clinical and genetic tests is entirely normal may produce several children with the same dominant or X-linked disease if he or she has a germ-line clone of mutant cells. When a normal couple have a child with a dominant condition, with no previous family history, this is evidently a new mutation. The counselor must remember that one or other parent may be a germ-line mosaic. The risk of recurrence can seldom be quantified, because we have no idea what proportion of germ-line cells carry the mutation, but it is not negligible. If a normal couple have two or more affected children, the pedigree pattern (*Figure 1.16*) looks recessive because the parents are unaffected. Misinterpreting the condition as recessive would cause serious errors when the affected children ask about their risk of passing it on. The risk would be very low for a rare recessive condition, but 50% for a dominant condition.

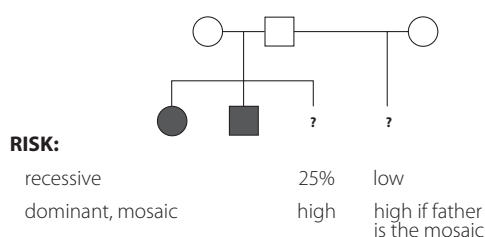


Figure 1.16 – A problem in counseling.

A couple with no previous family history of this condition have two affected children. Is this an autosomal recessive condition, or is it autosomal dominant with one of the parents being a germinal mosaic? The risks are very different depending which theory is correct.

Type 1 Neurofibromatosis (OMIM 162200)

NF1, also known as von Recklinghausen disease, is an autosomal dominant condition caused by mutations in the *NF1* gene on chromosome 17. This condition exemplifies the issues of penetrance, expressivity and *de novo* mutations described above.

The function of the *NF1* gene product, neurofibromin, is described in *Disease box 3*. The disease affects about 1 person in 3500 and is seen in both sexes and all ethnic groups. Penetrance is complete, in that some feature of the condition can be found in every affected person, but the disease can manifest in many ways and varies greatly in severity (see *Box figure 1.3*). Each child of somebody with NF1 has a 50% chance of inheriting the disease gene, and this might be checked with a DNA test, but the test cannot tell us how severely a child would be affected. Approximately

half of all cases represent new mutations. This is a frequent observation with clinically severe dominant conditions where affected individuals are less likely to have children: if the condition nevertheless persists in the population, this must be because of recurrent mutations. The situation is explored in more detail in *Chapter 9*.



Box figure 1.3 – NF1.

Mildly affected patients show only café-au-lait skin macules (a), Lisch nodules (hamartomas in the iris) and/or axillary freckling. Dermal neurofibromas (b) are frequent, and they can be numerous and grossly disfiguring (c). NF1 patients are also liable to develop a variety of benign or malignant tumors, including nerve sheath tumors and gliomas and other tumors in the central nervous system. Occasionally NF1 patients have learning difficulties, short stature or seizures.

Mutation testing is available, but not simple: the *NF1* gene is large (59 exons, see *Chapter 3*), and the functional gene sequence must be distinguished from several closely similar but non-functional 'pseudogene' sequences in the genome. Counseling in NF1 can be difficult. Although DNA or RNA analysis can characterize the mutation in a family, it cannot predict how severely affected a mutation carrier will be. Mildly affected people may not understand that their children could be severely affected, while patients with no family history may not appreciate the genetic risks.

1.5. References

Fischbeck KH and Wexler NS (2019) Oligonucleotide treatment for Huntington's disease. *New Engl. J. Med.* **380**: 2373–2374.

Useful websites

Eurogems (www.eurogems.org) is a one-stop gateway to educational resources run by the European Society of Human Genetics. It gives links to a variety of online resources covering all aspects of genetics and divided according to the target audience, from schoolchildren to professionals. All resources have been carefully checked for accuracy, reliability and relevance.

Gene reviews (www.ncbi.nlm.nih.gov/books/NBK1116/) has chapter-length clinically oriented reviews of a large number (762 in October 2019) of genetic conditions.

OMIM (On-line Mendelian Inheritance in Man; <https://omim.org>) is the first port of call for finding out about mendelian diseases or other phenotypes and the genes that determine them. Each entry has systematic links to more detailed resources and very useful lists of references for further reading. Note that OMIM is a database of mendelian conditions and so it does not contain information on chromosomal abnormalities.

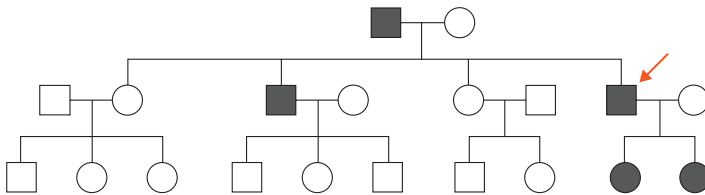
1.6. Self-assessment questions

Each of the following 10 pedigrees shows a rare disease.

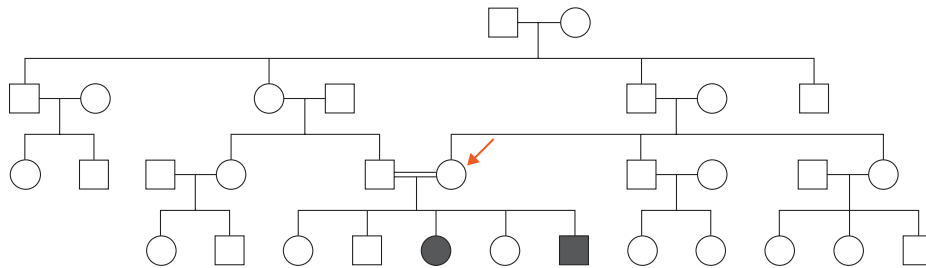
- Identify the most likely mode of inheritance, chosen from autosomal dominant (fully penetrant), autosomal dominant with about 90% penetrance, autosomal recessive or X-linked recessive
- Define the risk that the next child of the arrowed person or people will be affected.

[Hints on pedigrees 1–4 are provided in the *Guidance* section at the back of the book.]

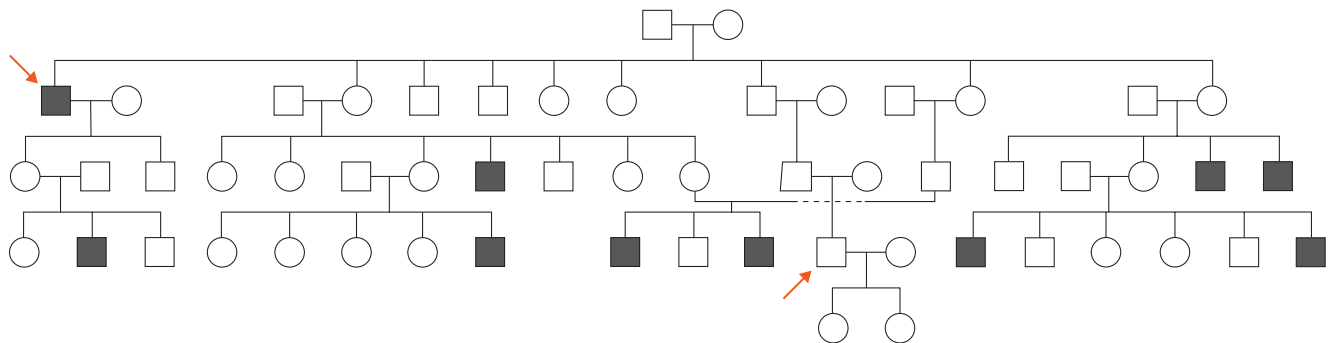
SAQ 1



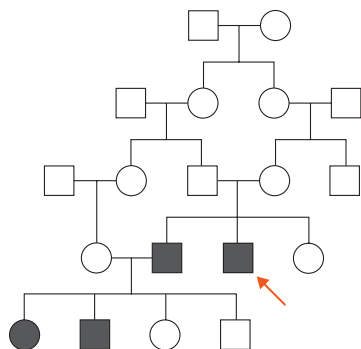
SAQ 2



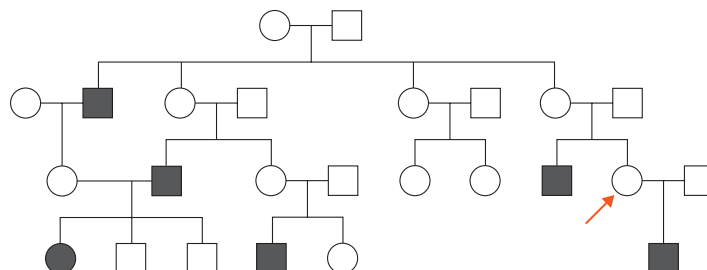
SAQ 3



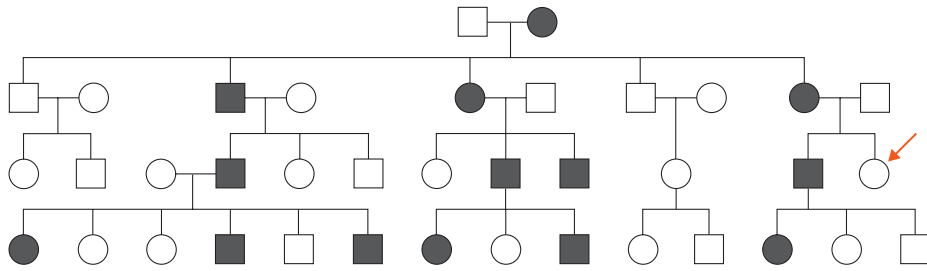
SAQ 4



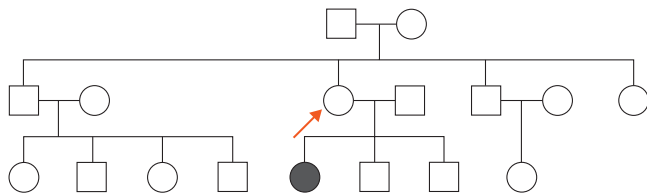
SAQ 5



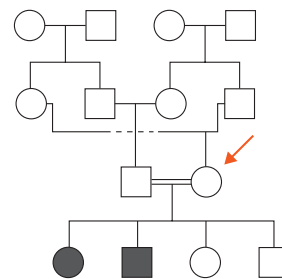
SAQ 6



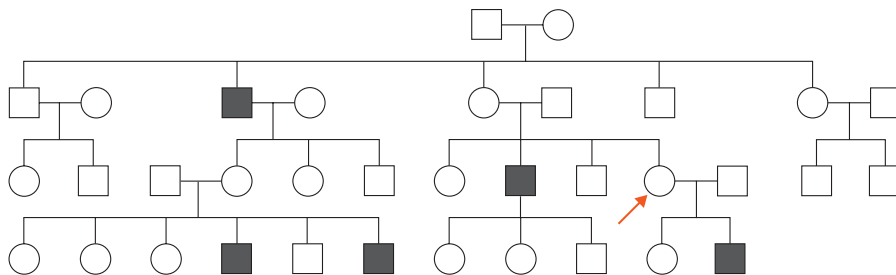
SAQ 7



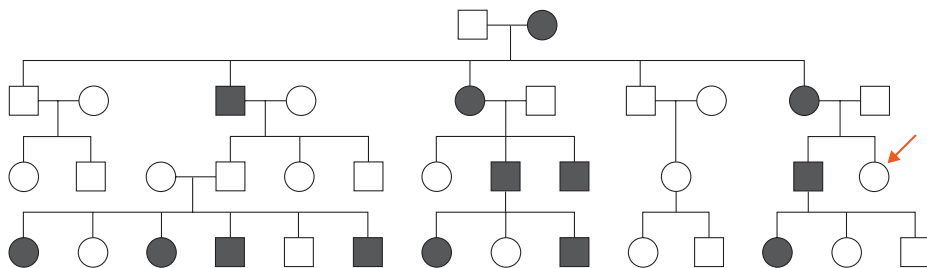
SAQ 8



SAQ 9



SAQ 10



02

How can a patient's chromosomes be studied?

Learning points for this chapter

After working through this chapter you should be able to:

- Describe the results of mitosis and meiosis, and how these are achieved
- Identify normal and simple abnormal karyotypes
- Describe human sex determination and the effects of errors
- Describe triploidy, trisomy, monosomy, reciprocal translocations, Robertsonian translocations, paracentric inversions, pericentric inversions, deletions and copy number variants
- Explain the origins of the main types of numerical and structural chromosome abnormalities
- Work out the main possible reproductive outcomes for carriers of translocations or inversions

2.1. Case studies

CASE 7 GREEN FAMILY

- George, aged 3 years
- Developmental delay, mildly dysmorphic



25 39 70 97 395

George is the second child born to healthy parents. Everything seemed fine at birth although his birth weight was less than his sister's. At his routine neonatal examination the doctor heard a heart murmur and referred him for an echocardiogram. This showed that he had a small hole between the two lower chambers of his heart (a ventricular septal defect, VSD). It was not considered serious and follow-up was recommended. By 3 years of age the murmur had disappeared and a further echocardiogram showed the VSD had closed. However George's parents were still worried about him since all his developmental milestones had been slower than his sister's. His speech development had been particularly slow and he was difficult to understand. The pediatrician referred him for speech studies and he was found to have poor movements of his palate. He then asked the opinion of a clinical geneticist. The geneticist noted that George had a narrow nose, prominent ears with over-folding of the upper part of the helices, a small mouth and long fingers. He organized a chromosome test and in addition to the routine karyotype requested a molecular test to see if there was a deletion of part of chromosome 22 (22q11).

Figure 2.1 – Child with 22q11 deletion.

Note small mouth, narrow nose and upward slant of his eyes.

CASE 8 HOWARD FAMILY

- Helen, newborn daughter of young parents
- Down syndrome confirmed



Figure 2.2 – A child with Down syndrome.

26 39 70 315 395

When Anne Howard attended the antenatal clinic in her first pregnancy she was offered screening tests for Down syndrome. She and her husband Henry decided against these because they knew that Down syndrome was particularly a risk for older women, and she and Henry were both in their early 20s; in any case, they both agreed they wouldn't really consider a termination of pregnancy. Everything went well in pregnancy and labor started at full term. When Helen was born, Anne thought she looked like her sister's baby, but the midwife was worried because she was very hypotonic (floppy), had loose skin at the back of her neck and single creases across her palms. The pediatrician was called and gently told Anne and Henry that he was concerned that Helen might have Down syndrome. He arranged to send a blood sample to the cytogenetics laboratory and requested that the result be sent as soon as possible since by now the parents and grandparents were extremely anxious. The pediatrician and the midwife saw the family several times over the next two days until the result was available. This confirmed that Helen did indeed have Down syndrome. The parents had lots of questions about Helen's future which the pediatrician answered, but he suggested he refer them to the genetic clinic in a few weeks' time to respond to their questions about how Down syndrome occurs and whether, if they were to have more children, there would be a high risk of recurrence.

CASE 9 INGRAM FAMILY

- Isabel, 10 years old with small stature and possibly delayed puberty
- ? Turner syndrome
- 45,X karyotype

26 42 70 103 285 395

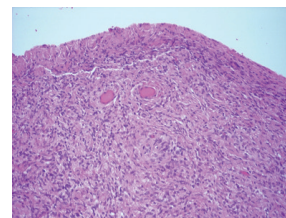
Isabel was the first baby born to Irene and Ian who were both quite tall. She had some swelling of her feet for the first few months, and was quite petite, but in general she was a healthy little baby. She developed normally in childhood but was always the smallest in the class. By 10 years of age, her classmates' growth rate had increased and a couple of her friends had started puberty. Although Irene and Ian were not really worried, the school nurse suggested that Isabel be referred to a pediatrician since her small stature seemed unusual given her parents' height. As part of her initial investigations of short stature and possibly delayed puberty, the pediatrician requested a chromosome analysis. This showed that Isabel had Turner syndrome, karyotype 45,X.



(a)



(b)



(c)

Figure 2.3 – Turner syndrome

(a) Puffy feet, (b) redundant skin at back of neck. (c) Histology of gonads: ovarian cortical stroma devoid of germ cell elements. Photo (c) courtesy of Dr Godfrey Wilson, Manchester Royal Infirmary.

2.2. Science toolkit

Why clinicians need to know about chromosomes

Chromosome abnormalities are important in a number of different clinical situations:

- they are a cause of infertility and recurrent miscarriages
- over 50% of embryos that abort spontaneously in the first trimester have a chromosomal abnormality
- about 1 newborn in 200 has multiple congenital abnormalities because of a chromosomal abnormality. Prenatal screening can detect most such abnormalities in time for the pregnancy to be terminated, if the parents so wish
- most chromosomally abnormal babies are born to parents who are entirely normal, but about 1% of people have a subtle chromosomal change that has no effect on their own health, but puts them at high risk of having miscarriages or abnormal babies
- cancer cells typically acquire extensive chromosome abnormalities not present in the normal cells of the patient, and many particular abnormalities have diagnostic and prognostic significance

How are chromosomes studied?

Although chromosomes in some organisms had been studied since the 1880s, it was not until 1956 that some novel technical tricks allowed human chromosomes to be counted and distinguished for the first time – until then, preparations of human chromosomes had always formed hopeless tangles that could not be resolved under the microscope. For the next 50 years clinical cytogenetics used essentially the same methods, albeit with progressive refinements. Highly skilled technicians examined spreads of chromosomes under the microscope to look for numerical or structural abnormalities. In recent years these techniques have been increasingly supplanted by analysis of DNA on microarrays, as described in *Chapter 4*. In the near future these methods may in turn be overtaken by advances in DNA sequencing. Nevertheless, traditional techniques are still used for specific purposes, and from an educational point of view the resulting karyotypes allow a much clearer appreciation of the chromosomal mechanisms that are so important in clinical genetics. For this reason we have illustrated all the cases in this chapter with traditional karyotypes, even though increasingly laboratories would actually have used DNA-based methods rather than microscopy to make the diagnoses discussed here.

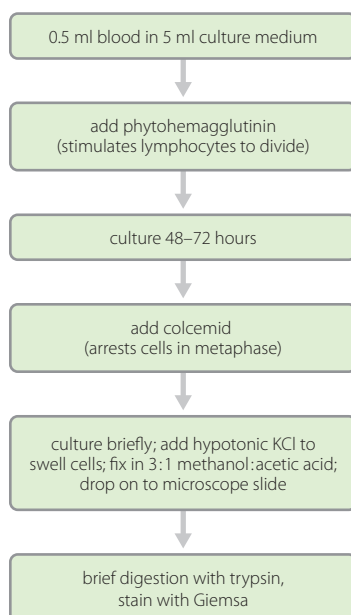


Figure 2.4 – Flowchart for processing a blood sample to obtain a standard G-banded chromosome preparation.

Karyotyping by traditional microscopic analysis

If you want to study chromosomes under the microscope you need dividing cells. As described below, chromosomes are only visible in cells that are in the act of dividing. Taking blood, skin or other samples from a person will yield plenty of cells, but few if any will be dividing. Thus traditional chromosome analysis usually involves taking a sample of non-dividing cells, which are then cultured in the laboratory and persuaded to divide (*Figure 2.4*; *Figure 2.9* shows one such cell). Suitable sources are shown in *Box 2.1*. The cytogeneticist counts the chromosomes, identifies each one and checks that its structure appears normal. The analysis is typically based on 10 cells, so that if a particular feature is not clear in one cell it can be checked in others. For record-keeping, the

Material for chromosome analysis

DNA-based analysis, as described in *Chapter 4*, can be performed on any source of cells (or indeed on cell-free DNA present in the bloodstream). Traditional microscopic analysis uses particular cells.

- **Peripheral blood lymphocytes** are the commonest material used. 0.5–10 ml of blood is treated with phytohemagglutinin, which stimulates lymphocytes to divide. Cultures are harvested after 48 hours. The protocol is summarized in *Figure 2.4*.
- **Chorionic villi** are used for early prenatal diagnosis. They are usually collected at 11–14 weeks of gestation by transvaginal or transabdominal routes (see *Chapter 14*). The procedure carries around a 1% risk of causing a miscarriage (additional to the background risk). Spontaneously dividing cells are present and can be used for rapid analysis, but results are best confirmed on cultured cells. It is important to isolate fetal material by careful dissection in order to avoid inadvertently culturing maternal cells.
- **Amniotic fluid** collected from around 16 weeks of gestation contains shed fetal cells. These are slow to grow and require around 2 weeks in culture to provide enough dividing cells for analysis. Amniocentesis (see *Chapter 14*) carries an increased risk of miscarriage of around 0.5–1% over the background.
- **Skin biopsies** are used to look for chromosomal abnormalities that may not be present in blood.
- **Testicular biopsies** are the only way to study male meiosis. In females, meiosis takes place in the fetal ovaries before birth, so there is no way of investigating meiosis clinically in a female patient.
- **Cells from a tumor biopsy** are studied in cancer patients to identify acquired somatic changes not present in the normal cells of the patient (see *Chapter 7*).

chromosomes in one cell are arranged (usually by manipulation of a digital image) into a standard karyogram (often loosely called a karyotype) as in *Figures 2.8* or *2.10*. By this means, the entire genome can be surveyed for any numerical or structural chromosomal abnormality.

Chromosomes are always prepared from dividing cells, in which the DNA has already been replicated, and therefore under the microscope they are always seen to consist of two identical **sister chromatids** joined at the **centromere**. In former times the standard preparation protocols made this structure very clear (see inset to *Figure 2.9*). The centromere and sister chromatids are much less obvious in chromosomes prepared according to current protocols, which leave the sister chromatids pressed tightly together. Cytogeneticists rely more on the banding pattern for identifying each individual chromosome and detecting structural aberrations. Banding is produced by laboratory manipulations that make each chromosome stain in a reproducible and characteristic pattern of dark and light bands. The usual method is **G-banding**. This involves subjecting the chromosomes, spread out on a microscope slide, to a brief digestion with trypsin, followed by staining with Giemsa stain. Other banding methods are sometimes used for particular analyses. R-banding produces a reversed pattern of dark and light bands, useful for checking chromosome ends. Other specialized procedures stain centromeres (C-banding) or the short arms of the acrocentric chromosomes (silver staining).

Chromosomal locations are named according to the Paris Convention as described in *Box 2.2*. Bands are numbered counting outwards from the centromere. *Figure 2.5* shows



The ideograms show ideal G-banding patterns at 550 band resolution. Major bands are labeled 1, 2, 3, etc., going from centromere to telomere. Major band 11q1 (11q means the long arm of chromosome 11, 11p the short arm) is divided into sub-bands 11q11 – 11q14, and at the highest resolution 11q14 splits into 11q14.1 – 11q14.3. Redrawn from Shafer and Tommerup (2005) with permission from S. Karger AG, Basel.

the standard ideograms and nomenclature of G-banded chromosomes at a resolution of 550 bands. Higher resolutions can be obtained by harvesting the cells well before the chromosomes become maximally contracted at metaphase of cell division (see *Figure 2.6a*). In these highly extended chromosomes, bands split into sub-bands and sub-sub-bands, allowing more precise localizations such as 7q11.23 (pronounced '7q one one point two three'). However, longer chromosomes are more likely to be tangled up, making analysis under the microscope at the highest resolutions (1500–2000 bands) very difficult. Abnormalities too small to be seen by standard 550-band analysis are best detected by switching to molecular methods. We will see this process in action as we follow the case of **George Green (Case 7)**.

Chromosomes and their abnormalities: nomenclature and glossary

Karyotypes are described by the number of chromosomes, the sex chromosome constitution, and any abnormality. Locations on chromosomes are described in relation to the Paris Convention of nomenclature shown in *Figure 2.5*. p means the short arm, q the long arm, t signifies a translocation, del a deletion, dup a duplication, inv an inversion, and der a derivative chromosome, whose structure would then be specified.

- 46,XX – a normal female
- 47,XY,+21 – a male with trisomy 21
- 46,XX,t(1;22)(q25;q13) – a female with a translocation between chromosomes 1 and 22 with breakpoints at 1q25 and 22q13 (the abnormality in **Case 5**, *Figure 2.14*)
- 46,XY,del(2)(q34;q36.2) – a male with a deletion on the long arm of one copy of chromosome 2 taking out the material between 2q34 and 2q36.2.

This is enough detail for present purposes. For the full nomenclature, covering every possible abnormality, see Shaffer, Slovak and Campbell (2009).

Acrocentric – a chromosome that has its centromere close to one end – chromosomes 13, 14, 15, 21 and 22 in humans.

Autosome – any chromosome that is not the X or Y sex chromosome.

Centromere – the position on a chromosome where the sister chromatids are joined; the location of the **kinetochore** where the spindle fibers attach during cell division to pull the chromatids apart.

Chromatid – in a dividing cell, a chromosome consists of two identical sister chromatids joined at the centromere. After cell division, and until the DNA is next replicated, a chromosome consists of a single chromatid.

Chromatin – a general term for the DNA–protein complex that makes up chromosomes.

Euchromatin – chromatin with a relatively open structure in which genes can be active; the opposite of heterochromatin.

G-banding – a standard procedure in which chromosomes are treated so that they stain in a characteristic and reproducible pattern of dark and pale bands, as shown in *Figure 2.5*.

Heterochromatin – chromatin that is highly condensed and genetically inactive. Found mainly at centromeres.

Homologous chromosomes – the two no.1s or the two no. 2s etc. in a person. Note that unlike sister chromatids, homologous chromosomes are not copies of one another, and they may differ in small ways (minor DNA sequence differences) or sometimes in large ways (because of translocations, etc.).

Inversion – a structural abnormality in which part of a chromosome is in the wrong orientation compared to the rest (see *Figure 2.19*).

Karyotype – a person's chromosome constitution – also used loosely to describe a display of a person's chromosomes, as in *Figure 2.8*, etc.

Metacentric – a chromosome that has its centromere in the middle (e.g. human chromosomes 3 and 20).

Monosomy – having one copy of one particular chromosome, but two of all the others (i.e. 45 in total for an autosomal monosomy).

Robertsonian translocation – a special type of translocation in which two acrocentric chromosomes are joined close to their centromeres as in *Figure 2.20*.

Sister chromatids – the two chromatids of a chromosome as seen in a dividing cell. Sister chromatids are copies of each other, made during the preceding round of DNA replication.

Submetacentric – a chromosome that has a long arm and a short arm, e.g. most human chromosomes.

Telomere – the special structure at the end of each chromosome arm.

Translocation – a structural abnormality in which two chromosomes swap non-homologous segments.

Triploidy – having three complete sets of chromosomes (i.e. 69 in total).

Trisomy – having three copies of one particular chromosome, but two of all the others (i.e. 47 in total).

Chromosome abnormalities

Chromosomal abnormalities can involve having the wrong number of chromosomes, or having one or more structurally abnormal chromosomes. The nature and origin of chromosome abnormalities are discussed in more detail in *Section 2.4*. Several numerical chromosome abnormalities occur sufficiently frequently to produce syndromes that are recognizable clinically (*Box 2.3*). So, for example, the midwife recognized the characteristic features of Down syndrome in **Helen Howard (Case 8)**.

Abnormalities of chromosome structure are usually the result of chromosome breakages, or sometimes errors in DNA replication or recombination. In some cases there is a loss or gain of material, in others merely a rearrangement. Most structural abnormalities are one-off events caused by random chromosome breaks. Although these do not produce specific named syndromes, clinicians have learned to suspect a chromosomal abnormality in babies who have multiple congenital abnormalities that are not ascribable to failure of one specific developmental event, or in patients with a combination of intellectual disability and dysmorphism. For these reasons it was appropriate to request chromosome analysis in the case of **Elizabeth Elliot (Case 5)**.

Loss or gain of large amounts of material is usually lethal, but smaller imbalances may result in a surviving but abnormal child. Those that are compatible with life are often too small to be seen under the microscope, but are readily detected by the molecular methods described in *Chapter 4*. They are described as **microdeletions** or **microduplications**. Many such microdeletions have been described, some one-off, others recurrent. Some give rise to recognizable recurrent syndromes (*Box 2.4*). An alert clinician would suspect one such syndrome in the case of **George Green (Case 7)**.

Of those listed in the table in *Box 2.4*, Wolf–Hirschhorn, cri du chat and Miller–Dieker syndromes involve deletion of the end of a chromosome arm. These are usually the

Syndromes due to numerical chromosome abnormalities

Triploidy (69,XXX, XXY or XYY)

Triploidy is common at conception but triploid embryos and fetuses almost never survive to term, and those that do so do not live for long.

Autosomal trisomies

All possible autosomal trisomies can be found among early miscarriages, but only trisomies 13, 18 and 21 generally survive to term. Chromosomes 13, 18 and 21 have the lowest density of genes of any chromosomes in our genome, so that fewer genes are present in abnormal numbers in these trisomies, compared to trisomies of other similar sized chromosomes. How trisomies originate is discussed below.

- +21 Down syndrome – see **Case 8**. The only autosomal trisomy compatible with survival into adult life.
- +18 Edwards syndrome – affected babies normally die in the first year of life. Although they can be externally relatively normal, they have subtle signs, are growth retarded and have many internal malformations. Rare long-term survivors show very little developmental progress.
- +13 Patau syndrome – 50% of affected babies die in the first month, and the rest within the first year. They can have mid-line malformations of the head and face, ranging from mild (closely-spaced eyes, central cleft lip) to very severe (gross malformations of the face with a single central eye and holoprosencephaly – failure of the brain to develop two hemispheres). Polydactyly is another common feature.

Autosomal monosomies

All autosomal monosomies are lethal at the earliest stages of pregnancy.

Sex chromosome abnormalities

Having wrong numbers of sex chromosomes is much less deleterious than having wrong numbers of autosomes – unsurprisingly, because normal development takes place in people with 1 or 2 X chromosomes and 0 or 1 Y chromosomes (46,XX, 46,XY). For the Y chromosome this is because it carries very few genes, and none are essential to life. The X chromosome is different: it carries about 1000 genes including many that are essential to life, but the mechanism of X-inactivation (see *Chapter 11*) greatly reduces the effect of having differing numbers of X chromosomes.

As the examples below show, a person is male if he has a Y chromosome, regardless of the number of X chromosomes, and female if not. The *SRY* gene on the Y chromosome is believed to be the master switch in sexual differentiation.

- 45,X Turner syndrome – females, pubertal failure, infertile, often short stature, normal intelligence. May have neck webbing, heart defects (coarctation of the aorta) and horseshoe kidneys. See **Case 9 (Isabel Ingram)**.
- 47,XXY Klinefelter syndrome – males, pubertal failure, infertile and often tall with female distribution of body fat. There may be a slightly lowered IQ compared to siblings.
- 47,XXX females, mostly undiagnosed because they are relatively normal. There may be a slightly lowered IQ compared to siblings.
- 47,XYY males, tall, possibly mildly reduced intelligence but within the normal range. The great majority of XYY men are living normal lives and are not diagnosed, but there may be a slightly increased risk of behavior problems.

result of random breakages and the proximal breakpoint varies between different patients, but the deletion always encompasses a critical region specific to the syndrome. The other syndromes in the table involve interstitial breaks with normally exactly the same breakpoints in every patient. These usually arise through recombination between misaligned low-copy DNA repeats, as explained in *Disease box 2*. For each of these microdeletion syndromes, there is a corresponding microduplication syndrome, although the clinical features of the duplication syndrome may be less specific.

Recurrent microdeletion and microduplication syndromes

A number of well-recognized clinical syndromes are caused by **microdeletions** or **microduplications**. These involve a change in a chromosomal segment that is too small to be noticed on a standard chromosome preparation like the one shown in *Figure 2.8*. When an alert clinician suspects such a change on clinical grounds it can be detected and characterized using the molecular cytogenetic techniques described in *Chapter 4*. Some well-defined microdeletion or microduplication syndromes are listed in the table below.

Syndrome	Location of change	Comments
α -thalassemia	Del at 16p13	Major East Asian form
Wolf–Hirschhorn	Del at 4p16	Low birthweight, ID, fits, typical face
Cri du chat	Del at 5p15	ID, typical face, high-pitched cry
Williams–Beuren	Del at 7q11.23	See <i>Disease box 2</i>
Angelman	Del at 15q11–q13	See <i>Chapter 11</i>
Prader–Willi	Del at 15q11–q13	See <i>Chapter 11</i>
Miller–Dieker	Del at 17p13	Lissencephaly, ID, typical face
Smith–Magenis	Del at 17p11.2	ID, behavioral problems, abnormal sleep patterns
Potocki–Lupski	Dup at 17p11.2	Developmental delay, autistic spectrum disorder
Di George–VCFS	Del at 22q11	See Case 7, George Green .

ID = intellectual disability.

In some cases similar events have produced non-pathogenic variations in the numbers of certain genes. For example, normal healthy individuals can differ in their number of copies of the green color vision pigment gene on the X chromosome, or the gene for salivary amylase on chromosome 1p21. In *Chapter 10* we will discuss the implications of having different numbers of copies of the *CYP2D6* gene on chromosome 22q13. Variations in the number of copies of the α -globin gene at 16p13 give rise to the various forms of α -thalassemia in Asia:

- Most people have two tandemly repeated copies of the gene on each copy of chromosome 16.
- People heterozygous for deletion of a single copy (three copies overall) are healthy carriers of α -thalassemia.
- Those with two normal alpha genes have heterozygous α -thalassemia with microcytosis.
- People with only one normal alpha gene in total have Hb H disease with microcytosis and hemolysis.
- A baby with no normal alpha gene has fatal hydrops fetalis (OMIM 236750).

Why do we have chromosomes?

A diploid human cell contains 2 m of DNA. Imagine a typical cell nucleus, 10 μm across, magnified one million fold, to the size of a lecture room 10 m across. The DNA would be represented by 2000 km of thin string, occupying much of the space in the room. Now replicate the DNA, turning each single strand of string into a double strand like a twin electric flex. And now you must divide the cell. In about an hour a cell succeeds in precisely dividing its replicated DNA, so that each daughter cell gets exactly one copy of every piece of DNA. If you are going to avoid hopeless tangles and confusion in your lecture room full of twin flex, you need to organize it in some very precise way. That is what chromosomes do in a cell.

Centromeres and telomeres

Chromosomes are not simply passive packages of DNA. They are functional cellular organelles, and parts of their function depend on two special structures, centromeres and telomeres.

- A functional chromosome must have one, and only one, centromere. As mentioned previously, the two sister chromatids are joined at the centromere. Importantly, the centromere (or strictly, the kinetochore, a structure located at the centromere) is the attachment point for the spindle fibers that pull the chromosomes apart during cell division (see below).
- Telomeres are special structures at each end of a chromosome. Telomeres contain long arrays of tandemly repeated DNA sequences, $(\text{TTAGGG})_n$. Because of the detailed enzymology of DNA replication, each chromosome end loses around 10–20 repeat units each time a cell divides (see Box 7.2). If the telomere is totally lost the chromosome becomes unstable, normally leading to cell death. Some theories link this process with ordinary aging or with responses to lifestyle stresses, but this is controversial. Telomeres have enough repeats to survive the cell divisions that occur within the lifetime of a person, but between generations they need to be renewed. Germ-line cells, and also cancer cells, produce a special enzyme, **telomerase**, that is able to restore telomeres to full length, helping to make such cells immortal.

The behavior of chromosomes during cell division

Preparing to divide

As discussed above, the primary function of chromosomes is to allow cells to distribute their DNA to daughter cells in an orderly fashion.

- Replicating the DNA.* Each chromosome initially contains a single immensely long DNA double helix. When a cell is preparing to divide, during S phase of the cell cycle (see Figure 7.6), the DNA is replicated, but the two copies remain attached to each other. Each chromosome then consists of two identical sister chromatids, each containing a full copy of the DNA double helix and joined together at the centromere. When visible under the microscope, chromosomes always have this structure (even though, as mentioned above, the two sister chromatids are rarely distinct in standard preparations) – but it is important to remember that the normal state of a chromosome in a non-dividing cell is as a single chromatid.
- Condensing the chromosomes.* During the early part of cell division (**prophase**) the chromosomes become much more compact, until they become visible under the microscope.

What is seen next depends on what daughter cell is to be produced. There are two sorts of cell division:

- **Mitosis** is the normal form of cell division, and the form in which chromosomes are almost always studied for clinical purposes. In mitosis the replicated DNA is divided exactly equally between the two daughter cells, so that they are genetically identical.
- **Meiosis** is a specialized form of cell division that is used only to produce gametes (sperm or eggs). Meiosis has two purposes. First, the number of chromosomes must be reduced from 46 to 23, so that when the sperm and egg fuse, the result is a 46 chromosome zygote. Second, meiosis uses two mechanisms, described below, to ensure that every gamete carries a novel and unique combination of the parental genes. As mentioned in *Box 2.1*, clinically meiosis can be studied in males through a testicular biopsy, and this may be part of investigations of male infertility. Female meiosis is virtually impossible to study in humans because most stages take place in the fetus before birth. However, the consequences of errors in meiosis are central to clinical cytogenetics.

Mitosis

In mitosis (*Figure 2.6a*), when each chromosome becomes visible it consists of two highly condensed sister chromatids held together at the centromere. At the end of prophase the nuclear membrane dissolves and the chromosomes move to the center of the cell. The positions of the nuclei of the two daughter cells are already marked by radiating arrays of microtubules. These attach to the kinetochore at the centromere of each individual chromosome. Each chromosome is held by microtubules radiating from both poles of the cell. The microtubules contract, pulling the chromosomes to lie on the equatorial plane of the cell (**metaphase**). Eventually the centromere of each chromosome splits, so that as the microtubules continue to contract, one chromatid of each chromosome is pulled to each pole of the cell (**anaphase**). Once all the chromatids have arrived at the poles they decondense, nuclear membranes are formed round them, and the cell divides into two daughter cells.

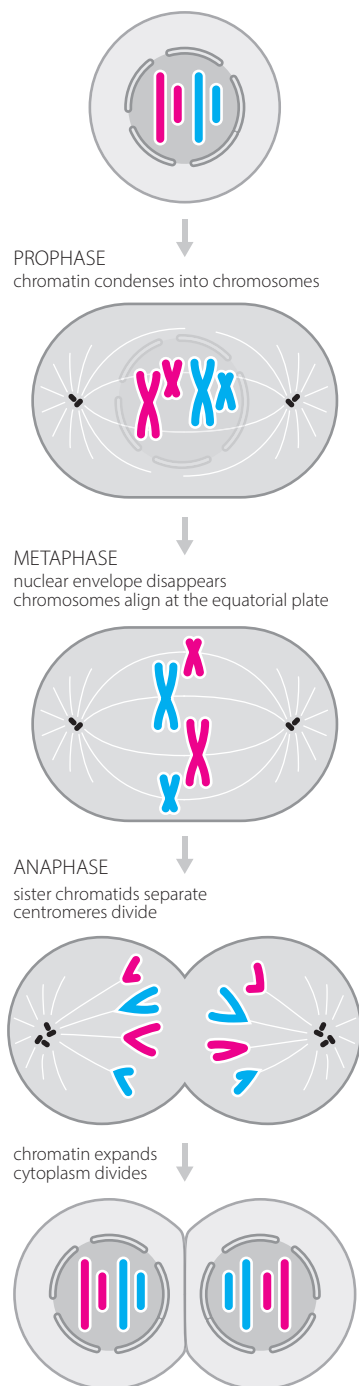
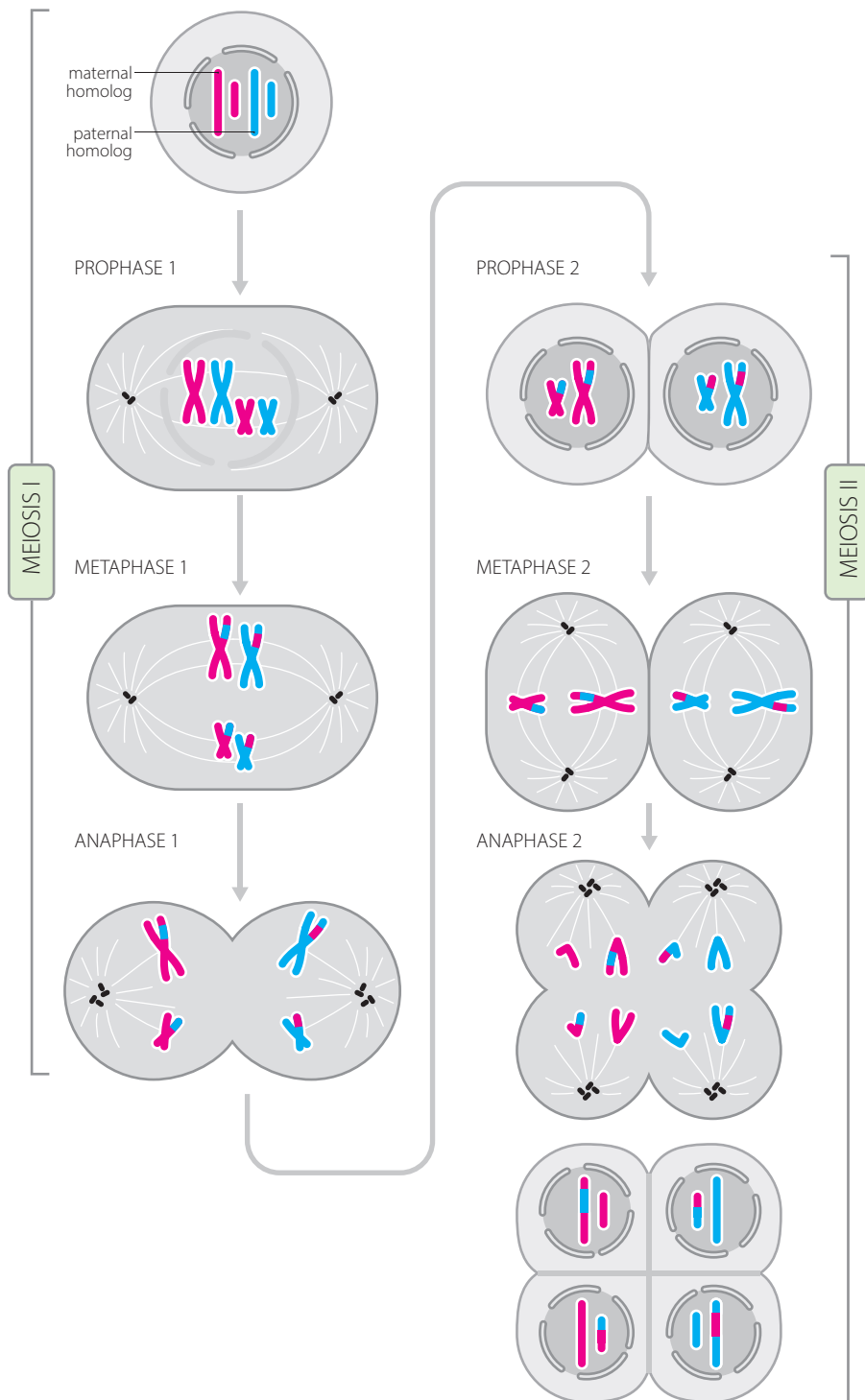
Note the essential feature of mitosis:

- the sister chromatids are copies of each other, and each daughter cell receives one chromatid of each chromosome
- each chromosome behaves independently. Although there are two copies (homologs) of each chromosome, the two do not interact in any way. Look at *Figure 2.9* and note the random arrangement of the chromosomes in the spread. This is a key difference between mitosis and meiosis.

Meiosis

As mentioned above, meiosis is the highly specialized form of cell division that is used only to produce gametes (*Figure 2.6b*). Gametes have 23 chromosomes and each gamete is genetically unique. Meiosis consists of two successive cell divisions; meiosis II is similar to mitosis, but meiosis I has special features.

During prophase I the chromosomes condense and become visible, as in mitosis, but in meiosis I the bodies that appear are not 46 separate chromosomes but 23 bivalents (*Figure 2.7*). Each bivalent is a four-stranded structure, consisting of two homologous chromosomes (the two no. 1s, etc.), each of which consists of two sister chromatids. The two sister chromatids of a chromosome are identical, because they are copies of each other, but the

(a) MITOSIS**(b) MEIOSIS****Figure 2.6 – Two types of cell division.**

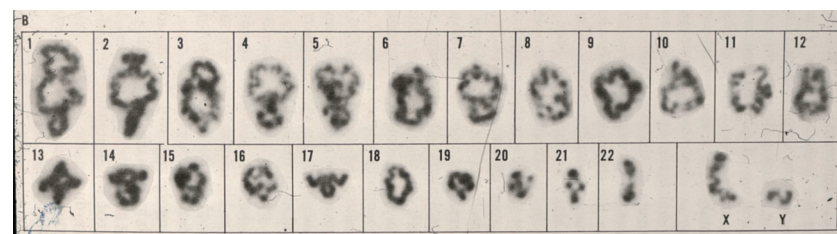
(a) Mitosis produces two genetically identical diploid (46 chromosome) daughter cells. (b) Meiosis produces four genetically different haploid (23 chromosome) cells (although in oogenesis only one of these develops into a mature oocyte; the others form the polar bodies).

two homologs are not identical. You might think of each as a long line of pigeon-holes. Each will have the same set of pigeon-holes, but the content of corresponding pigeon-holes on the two may be different. For example, near the bottom of the long arm of each copy of chromosome 9 there is a locus (a pigeon-hole) for ABO blood group. But one homolog may carry the A gene and the other the O gene. They are not copies of each other.

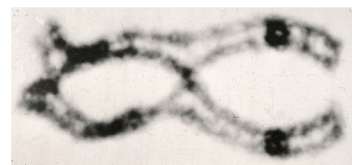
The pairing of homologs is extraordinarily accurate. When there is a structural abnormality, so that the homologs do not completely match, matching segments pair (unless this involves tying the chromosomes in impossible knots). *Figure 2.17* shows an example of this effect when there is a translocation, and *Figure 2.22* shows what happens when one homolog has an inversion. In male meiosis the X and Y chromosomes also pair. Although most of their sequence is completely different, there is a short region of homology at the tips of the short arms (the **pseudoautosomal region**, see *Section 11.2*) and the X and Y use this to pair end-to-end.

At anaphase I the spindle fibers pull the chromosomes apart. In mitosis the centromere of each individual chromosome splits and the two sister chromatids are pulled apart; but in meiosis I the two homologous chromosomes pull apart, each still consisting of two sister chromatids joined at the centromere. Thus at the end of meiosis I each daughter cell has 23 chromosomes, each of which consists of two sister chromatids. In meiosis II the sister chromatids are pulled apart, just as in mitosis, so that the final product of meiosis is four cells, each containing 23 single-chromatid chromosomes. **Case 8 (Howard family)** and **Case 5 (Elliot family)** show examples of what can go wrong in meiosis.

Whatever its other merits and demerits, sex has one sole purpose in biology: to produce novel combinations of genes. Partly this is achieved simply by involving two different people in the process of sexual reproduction. But there are two additional mechanisms at work generating yet more novelty, and both of these depend on the way chromosomes behave during meiosis (*Figure 2.6b*).



(a)



(b)

Figure 2.7 – Examples of chromosomes during meiosis.

(a) A cell from a testicular biopsy showing chromosomes during prophase I of male meiosis. Each of the 23 structures is a bivalent, consisting of two homologous chromosomes, each having two chromatids. Note the end-to-end pairing of the X and Y chromosomes. (b) A bivalent seen in meiosis in an amphibian, which has large chromosomes that make the four-stranded structure clear.

- (1) When a person forms a gamete, only one of their two no.1 chromosomes will be picked to go into the gamete. It might be the one they got from their mother or it might be the one they got from their father. With each chromosome there are two choices – either the maternal or the paternal one. Over all 23 chromosomes we have $2 \times 2 \times 2 \times 2 \times \dots = 2^{23}$ different ways of picking the one no. 1, one no. 2, etc. that go into a particular gamete; 2^{23} is 8 388 608.
- (2) A second mechanism increases the number of possible variations from 8 million to effectively infinity. As previously mentioned, in the early stages of meiosis homologous chromosomes (the two no. 1s, etc.) pair up. But they don't simply stick together (synapse); they exchange segments. This is **genetic recombination**. The DNA of one homolog is physically cut and joined to the DNA of the homologous chromosome. Not surprisingly, the mechanism of recombination is complicated. It looks as though the DNA of the two chromosomes has been cut at precisely the same position and the cut ends joined together the other way round – as has apparently happened in *Figure 2.7b*. Actually that is not exactly how it is done – in reality, after the first cut, one of the cut strands invades the other chromosome. A complicated sequence of further strand invasions, DNA breakdown and resynthesis and DNA cuts ensues (see *Box 10.4* for more detail). This process requires a number of enzymes to unwind, strip back, resynthesize and cut and join the DNA – but for present purposes we can focus just on the end result, a precise exchange of segments.

The effect of recombination is that homologous chromosomes swap segments. Normally there is at least one crossover (point of recombination between synapsed chromosomes) in each arm of each chromosome pair. On average there are about 60 crossovers in each cell in spermatogenesis, and about 90 in oogenesis, though the actual number varies considerably between people and between cells. To a first approximation, crossovers are distributed at random along each of the 23 pairs of chromosomes (although closer investigation reveals some interesting non-random features). Thus each chromosome in every sperm or egg that a person produces carries a unique combination of genes that came from his or her father and mother. This topic is discussed further in *Chapter 8* where we consider how disease genes can be mapped to a particular chromosomal location, because the techniques used to do that depend on genetic recombination.

As a result of these mechanisms, every conceptus is a unique combination of a unique sperm with a unique egg. The only people who are not genetically unique are monozygotic twins, whose cells are all derived by mitosis from a single original conceptus. The fact that monozygotic twins are unique individuals despite being clones reminds us that genetics is not everything in life.

2.3. Investigations of patients

A note on karyotypes. As mentioned above, most laboratories no longer use karyotyping under the microscope as the default technique to look for chromosome abnormalities; instead they use microarrays as described in *Chapter 4*. Karyotyping under the microscope still has its place, mainly for checking for balanced abnormalities – see *Figure 2.14* for an example. However, for educational purposes, for understanding the nature and origins of chromosome abnormalities, traditional karyotypes are far more intuitive than microarray data. Therefore we illustrate cases with old-fashioned karyotypes, even though in reality most of the investigations would probably have used microarrays.

CASE 7 GREEN FAMILY

- George, aged 3 years
- Developmental delay, mildly dysmorphic
- Normal 46,XY karyotype but suspect microdeletion

25 39 70 97 395

The combination of slow development, a heart defect, palatal problems and mildly dysmorphic features in George suggested a chromosomal abnormality. Blood was taken for cytogenetic analysis, but the result (*Figure 2.8*) was a normal male karyotype, 46,XY. Because of his particular combination of clinical features the geneticist suspected that George might have the Di George – velo-cardio-facial syndrome (VCFS) which is caused by a microdeletion at chromosome 22q11. As described in *Box 2.4*, the Di George – VCFS deletion is too small to be visible under the microscope. He therefore requested a molecular test. The test and its result are described in *Chapter 4*.



Figure 2.8 – George Green's karyotype.

The result (at this resolution) shows a normal male karyotype, 46,XY. These karyotype diagrams are produced by electronic manipulation of a digital image of the original chromosome spread. Sometimes in the spread two chromosomes overlapped. The apparent crosses seen here, for example, in the right-hand copies of chromosomes 6 and 16 are artefactual, the result of the digital manipulation needed to separate two overlapping chromosomes.

CASE 8 HOWARD FAMILY

- Helen, newborn daughter of young parents
- Down syndrome confirmed
- 47,XX,+21 karyotype

26 39 70 315 395

The clinical diagnosis of Down syndrome in Helen was confirmed by karyotyping. A 2 ml blood sample was taken and, as shown in *Figure 2.4*, the cells were cultured, harvested after 48 hours, spread on a microscope slide and stained by G-banding. The cytogeneticist analyzed 10 cells by eye down the microscope (one of the cells is shown in *Figure 2.9*), and for record purposes she used an image analyzer program to arrange chromosomes from one cell into a standard karyotype (*Figure 2.10*). Her report gave the karyotype as 47,XX,+21, confirming that Helen had typical Down syndrome, trisomy 21. An additional reason for checking Helen's karyotype is discussed below in *Section 2.4*.

As expected, the pedigree showed nothing noteworthy (no previous abnormal babies or recurrent miscarriages) in either Anne's or Henry's family, and it is not illustrated here. Their first question was why it had happened, and why to them? The counselor explained that

it was a one-off accident in meiosis (see *Figure 2.11*). Either the two paired chromosomes 21 (in the first meiotic division), or the two sister chromatids of the one copy of 21 in the second division, had failed to disjoin at anaphase and segregate into separate daughter cells. They had both ended up in the same daughter cell, producing an egg or sperm with 24 chromosomes, including two copies of no. 21.

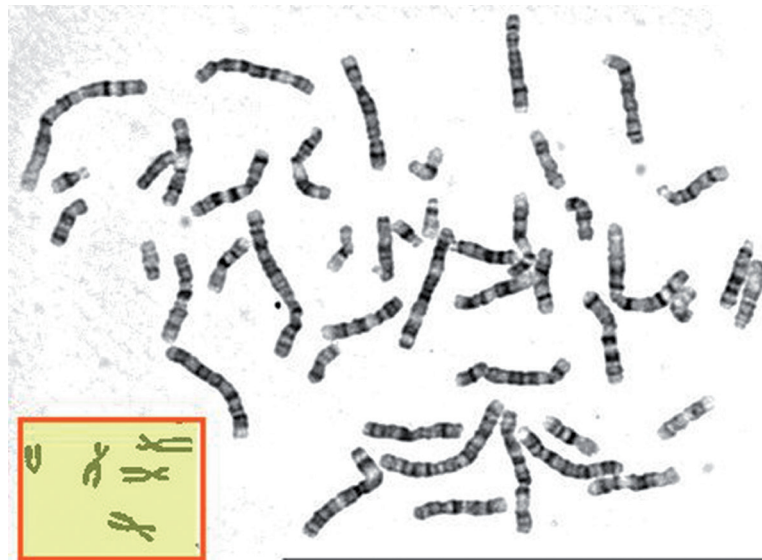


Figure 2.9 – 47,XX+21 spread.

There are three copies of chromosome 21. Note that homologous chromosomes (the two no. 1s etc.) behave entirely independently in mitosis. The preparation method used here leaves the two sister chromatids of each chromosome tightly pressed together, so that the position of the centromere is not obvious to the eye. This makes the banding pattern easier for the cytogeneticist to recognize. The inset shows some chromosomes that have been handled differently, to make the structure of sister chromatids joined at the centromere more obvious.

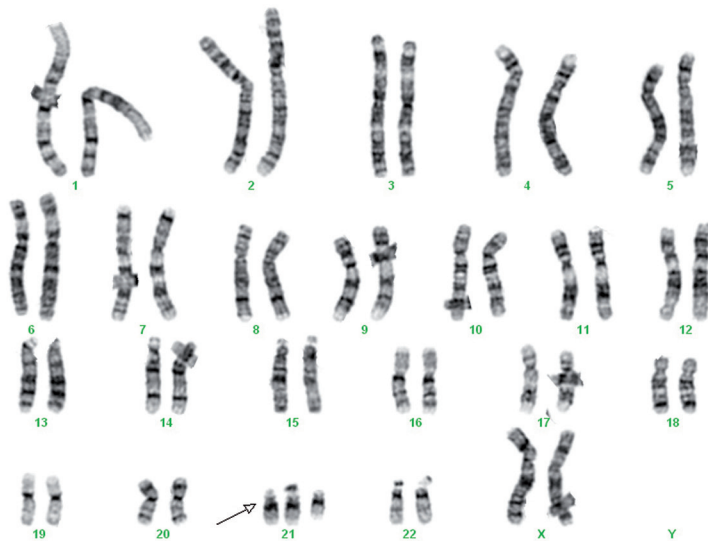


Figure 2.10 – Karyotype showing trisomy 21 (47,XX+21).

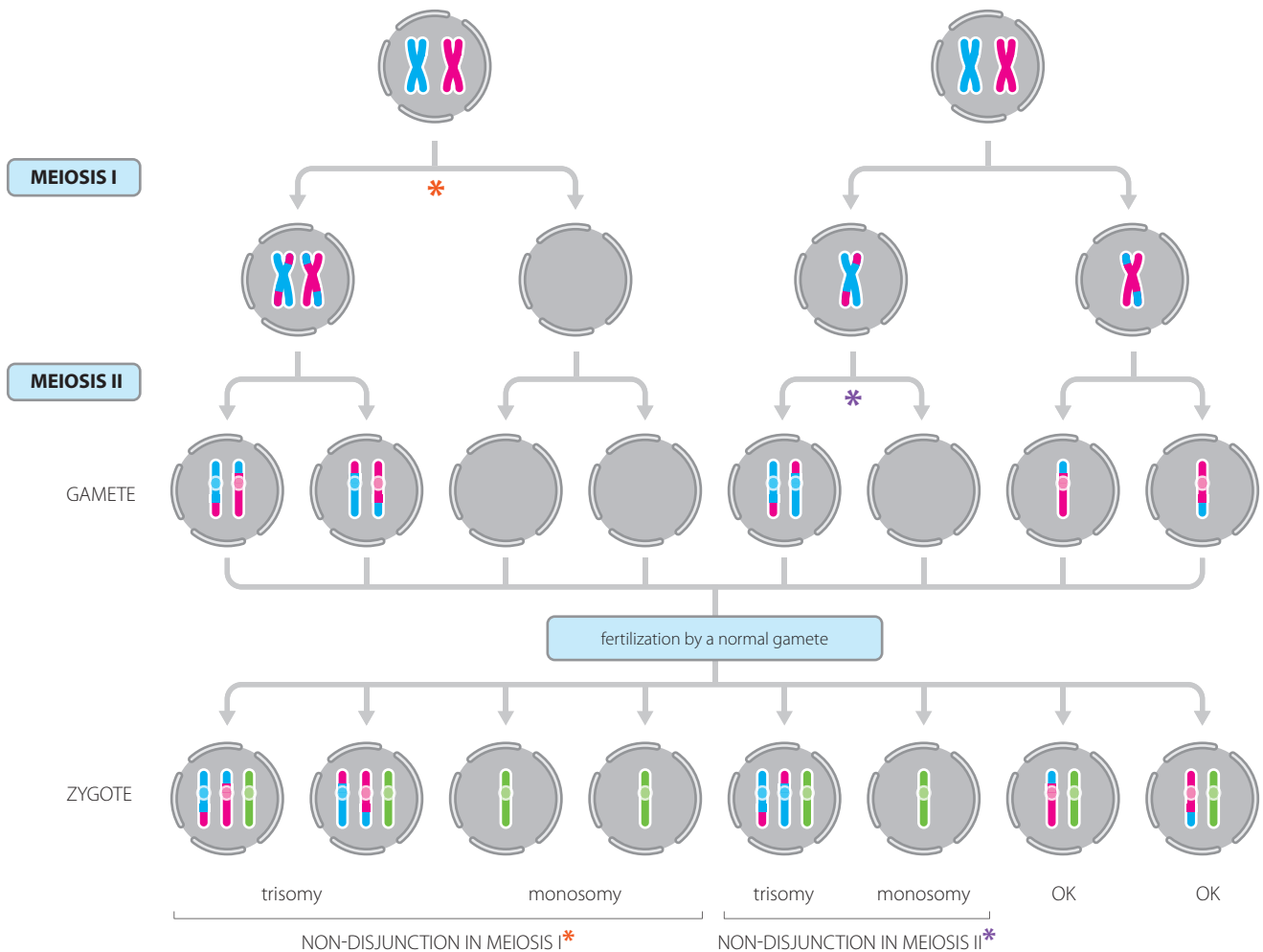


Figure 2.11 – The effects of non-disjunction in meiosis.

The non-disjunction involves only the single pair of chromosomes (meiosis I) or the single chromosome (meiosis II) shown; all the other chromosomes (not shown) disjoin and segregate normally.

Anne asked if it was always the woman's fault, as she had heard that said. No; in the first place these things are nobody's fault, and secondly, in principle the non-disjunction could happen at either division of meiosis in either parent, though DNA marker studies showed that 70% of cases were due to non-disjunction in the first meiotic division in the mother. This might possibly be a reflection of the extremely long duration of that stage in women, from before birth until whenever the relevant egg ovulated. In men, meiosis goes on continuously in the testes from puberty to old age. The individual risk rises sharply with the age of the mother (*Figure 2.12*) but Anne and Henry had been mistaken in their reason for declining screening. Though the individual risk is higher for an older woman, because most babies are born to younger women, most Down syndrome babies are also born to younger mothers.

Later, Anne and Henry requested an appointment to discuss options for prenatal diagnosis in any subsequent pregnancy. This led to a discussion of screening and the intervention described in *Chapter 12*.

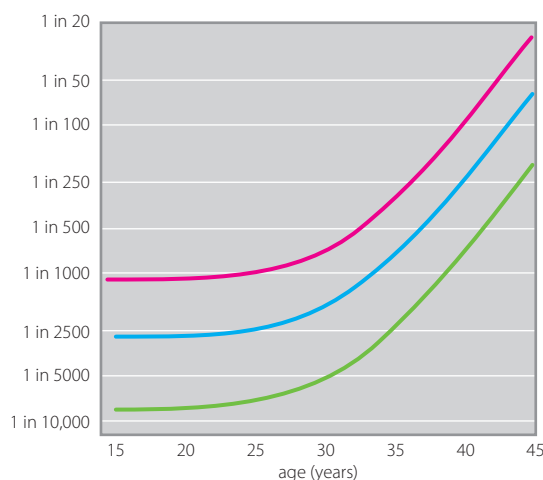


Figure 2.12 – Risk of having a baby with trisomy 13 (green curve), trisomy 18 (blue curve) or trisomy 21 (pink curve) depending on the age of the mother. Reproduced from the NHS Fetal Anomaly Screening Programme handbook for laboratories, with permission, under an Open Government Licence v3.0.

CASE 9 INGRAM FAMILY

- Isabel, 10 years old with small stature and possibly delayed puberty
- ? Turner syndrome
- 45,X karyotype

26

42

70

103

285

395

As described earlier, although Isabel had a few mildly abnormal features at birth, she only came to medical attention because she was small compared to her parents' heights and her peer group. Her phenotype and history of swollen hands and feet were strongly suggestive of Turner syndrome. Chromosome analysis confirmed this (*Figure 2.13*). This is the only human monosomy that is not lethal early in development. Because males survive with only one X chromosome, maybe it is not surprising that Turner syndrome is not always lethal. But in fact it is lethal in over 90% of cases. Fetuses with Turner syndrome can be grossly distended with fluid and the great majority abort spontaneously. The survivors are often born with puffy (edematous) hands and feet and with redundant skin on the neck, which represent the remains of presumably milder fetal edema.

Unlike all the trisomies, the risk of Turner syndrome does not increase with maternal age. The mechanism is different. Rather than non-disjunction, Turner syndrome is the result of anaphase lag, in which one of the sex chromosomes moves too slowly to the pole of a daughter cell during cell division, and ends up outside the nucleus, whereupon it is broken down. It can arise during gametogenesis or after conception during an early mitotic division. Many Turner women are mosaics (see *Section 1.4*, and below), and they can be either 45,X / 46,XX or 45,X / 46,XY mosaics. Where there is XY tissue in the gonad, it has a propensity to become malignant, and therefore the gonads are best surgically removed.

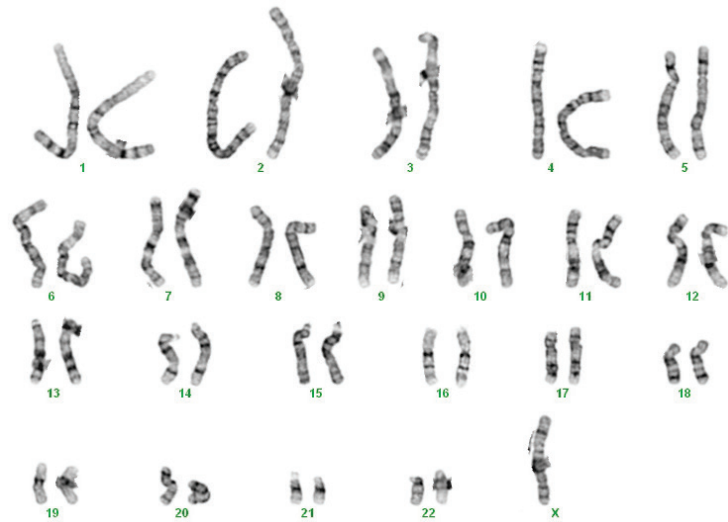


Figure 2.13 – Karyotype of Isabel Ingram.

Although Isabel will never be able to have children normally, treatment with estrogens can allow her to develop normal secondary sex characteristics and greatly assist her personal and social life. Modern reproductive technology has allowed some Turner syndrome patients to bear children using donor eggs. Treatment with growth hormone can result in improved growth and final height.

CASE 5 ELLIOT FAMILY

- Baby girl Elizabeth, parents Elmer and Ellen
- Multiple congenital abnormalities
- Family history of reproductive problems
- ? Chromosome abnormality
- Ellen – balanced 1:22 translocation
- Elizabeth – unbalanced segregation product

4

12

43

68

100

395

Previously it had been found that baby Elizabeth suffered from multiple congenital abnormalities (*Figure 1.5*). This suggested she had a chromosomal imbalance, but did not allow the geneticist to guess which chromosomes might be involved. Ellen had had one previous miscarriage, not in itself remarkable, but it was noted that her sister had also had two miscarriages. Further enquiries revealed a family history of reproductive problems – the pedigree is shown in *Figure 1.10*. Blood was taken from the baby and both parents for chromosome analysis. The analysis of Elizabeth was actually performed using the technique of array-comparative genomic hybridization, as described in *Chapter 4*. This technique can detect abnormalities too small to be seen on a conventional karyotype. However, to make the chromosomal events clear, we show a standard karyotype here. The results showed:

- Elmer – normal male karyotype, 46,XY
- Ellen – a balanced translocation between chromosomes 1 and 22 (*Figure 2.14*)
- Elizabeth – an unbalanced segregation product (*Figure 2.15*)

When they learned these results, the parents, as well as being very distressed and wanting to know the implications for Elizabeth, were keen to understand exactly what had happened and why. An explanation of the cytogenetics was given using language familiar to the family and explaining technical terms.

Ellen was a constitutional carrier of a balanced translocation – that is, it was present in every cell of her body (see below for a discussion of balanced vs. unbalanced abnormalities). The translocation had been present in the fertilized egg from which Ellen developed, and the pedigree suggested it was already present in one of Ellen's maternal

grandparents. At some time in that person or a more distant ancestor, chromosomes 1 and 22 had undergone breakages. Chromosome breaks are common events, but cells have machinery to repair them so that most go unnoticed. In this case, two simultaneous breaks had generated four broken ends, and by bad luck the repair machinery had

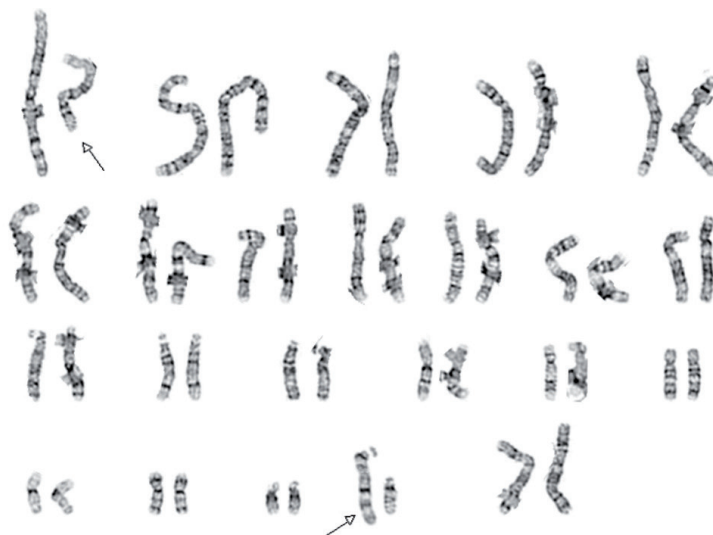


Figure 2.14 – G-banded karyotype of Ellen's chromosomes.

There is a balanced translocation. Chromosomes 1 and 22 have exchanged segments (arrows). The translocation is described as 46,XX, t(1:22)(q25;q13). Note that this is a case where traditional karyotyping may still be the method of choice. The microarray-based techniques described in *Chapter 4*, which are now the preferred method in many genetics centers, cannot detect balanced abnormalities, although DNA sequencing may do so, as explained in *Chapter 5*.



Figure 2.15 – G-banded karyotype of baby Elizabeth.

She has inherited Ellen's normal chromosome 1 but her translocated chromosome 22 (arrow). She is therefore trisomic for the portion of chromosome 1 distal to 1q25, the translocation breakpoint, and monosomic for chromosome 22 distal to 22q13.

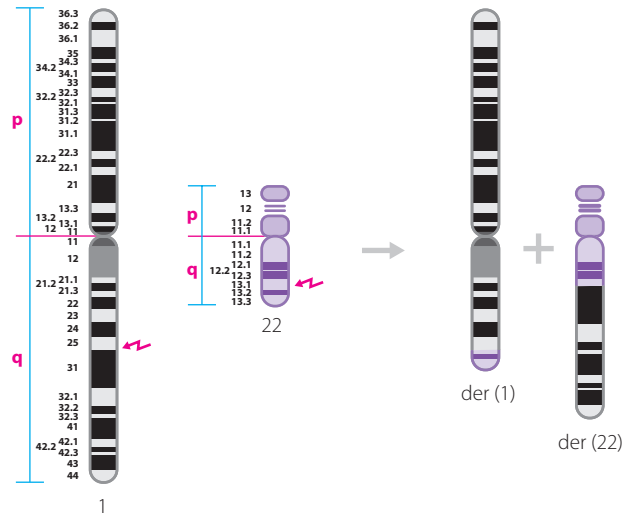


Figure 2.16 – How the 1;22 translocation in Ellen Elliot originated.

Chromosomes 1 and 22 broke at the positions indicated by the arrows, and the cell's DNA repair machinery rejoined the ends to form the two derivative chromosomes as shown. The derivative chromosomes are labeled der(1) and der(22).

joined those up the wrong way round (*Figure 2.16*). Alternatively, maybe the translocation had been produced by inappropriate action of the machinery responsible for genetic recombination during meiosis, cutting and joining non-matching chromosomes in a germ-line cell. Because each derivative chromosome nevertheless had a single centromere, mitosis proceeded with no problems, and because no genetic material was extra or missing, there was no phenotypic effect.

Although Ellen's cells could go through mitosis with no problems, meiosis was a different matter. In the first division of meiosis, homologous (matching) chromosomal segments pair (*Figure 2.6b*). In this case, pairing would produce a cross-shaped structure containing four whole chromosomes – a quadrivalent. When spindle fibers attached to the four centromeres and pulled them apart, they could segregate in various ways (*Figure 2.17*). She could produce entirely normal gametes, gametes carrying the balanced translocation, or various unbalanced forms. One of the latter had produced Elizabeth: she had partial trisomy of chromosome 1 and simultaneously partial monosomy of chromosome 22. The resulting genetic imbalance was the cause of her abnormalities.

There was a risk of problems in future pregnancies, which could take various forms depending on the way the translocated chromosomes segregated. The result could be another child with the same problems as baby Elizabeth, or a different segregation pattern could result in a child with a different set of multiple abnormalities; maybe the result would be severe enough to cause a miscarriage; or of course they could hope to be lucky and have a normal baby. All these various outcomes could be seen in other family members on the pedigree (*Figure 1.10*). The risk was substantial, but it was not easily quantifiable for several reasons:

- we can't predict the exact probability that the translocated chromosomes would segregate in each of the various possible ways

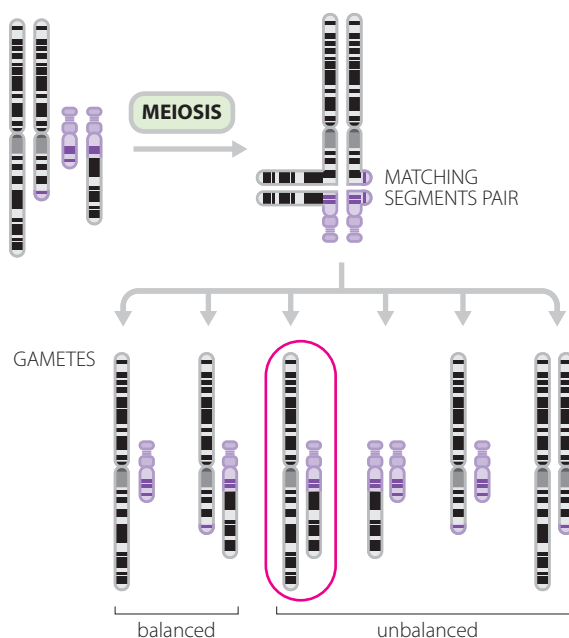


Figure 2.17 – Possible ways the chromosomes in Ellen could segregate in the first meiotic division.

During prophase I matching chromosome segments pair, resulting in a cross-shaped quadrivalent containing the normal and translocated copies of chromosomes 1 and 22. At anaphase I they pull apart, and the diagram shows various ways this could happen. The gamete that gave rise to baby Elizabeth is circled. Other more complex segregation patterns (3:1 segregation) are also possible. Note that the events shown took place in the first division of meiosis, when each chromosome in fact consisted of paired sister chromatids. For clarity the individual chromatids have not been shown.

- with some possible unbalanced outcomes a conceptus might abort so early that it would not be recognized as a problem
- it is uncertain whether other less fatally unbalanced karyotypes would result in a miscarriage or a live-born abnormal baby

If it had been Elmer who carried the translocation the risk would be lower, because abnormal sperm are less likely to win the race to fertilize the egg. If Elmer and Ellen wished, it would be possible in future pregnancies to check prenatally and offer termination where the fetus had one of the unbalanced karyotypes. This led eventually to the actions described in *Chapter 4*.

Months later, once Elmer and Ellen had begun to cope with Elizabeth's health problems, and had absorbed the information about the translocation and their own risks for future children, the question of family risk was discussed. The family history (see *Figure 1.10*) suggested that Ellen's aunt and sister could well be carrying the same balanced translocation, as might her younger sister, who was not yet married. The counselor explained that they needed to be made aware of the risk, preferably by Elmer and Ellen having a word with them to raise the subject. The counselor offered to see the relatives to explain the situation and the options available, including genetic testing. In the genetic clinic it is important to deal with each branch of a family separately, unless family members indicate otherwise. Confidentiality should be maintained and information about one family member only given to another with permission.

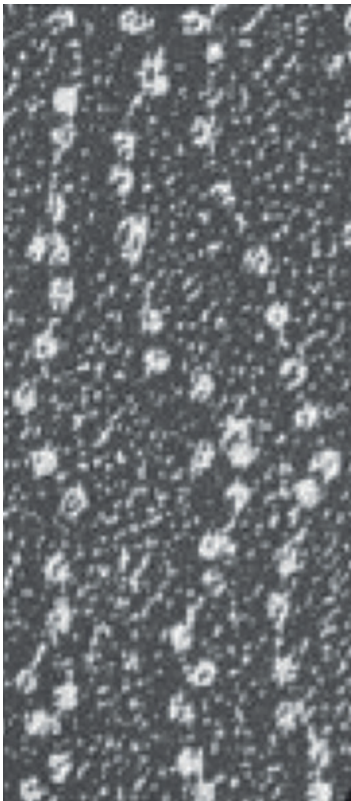
2.4. Going deeper...

What are chromosomes?

Chromosomes are packages of DNA. Each chromatid consists of a single immensely long DNA double helix, packaged by a diverse set of proteins and some RNA molecules. **Chromatin** is the generic name for the resulting DNA–protein complex. If you stretched out the DNA in a normal cell nucleus it would extend to an astonishing 2 meters, all squashed into a nucleus typically 10 micrometers (10 millionths of a meter) in diameter. When a cell divides the chromosomes must be accurately partitioned into the daughter cells, as described above, and in order to avoid impossible tangles the immensely long chromatin threads are bundled up into compact packages. These are the chromosomes that can be seen under the light microscope (*Figures 2.9 and 2.10*).

Chromosomes are equally present in non-dividing (interphase) cells, but the more extended chromatin fibers are too thin to be seen under the microscope. In order for our genome to function, the fibers need to be highly organized within the cell nucleus. The basic structure of chromatin is a string of beads (*Figure 2.18a*). The beads are called **nucleosomes**. A nucleosome has a roughly spherical core made up of eight molecules of special small proteins called histones, with 147 bp of DNA wrapped round it. Above this level, the chromatin is organized into a series of loops (*Figure 2.18b*) whose structure and function is the focus of much current research. These structures are crucial in regulating the way genes work, and will be considered in a bit more detail in *Chapter 11*.

(a)



(b)

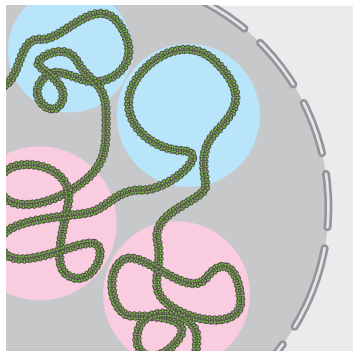


Figure 2.18 – Packaging DNA.

The DNA in a cell nucleus exists as a DNA–protein complex called chromatin. (a) The most basic component of chromatin is the nucleosome, consisting of approximately 147 bp of double-stranded DNA wrapped round an octamer of histone proteins. Under the electron microscope the appearance is of a string of beads. (b) This 'string of beads' is then organized into a series of loops that bring distant stretches of DNA into close proximity, and thereby allow interactions that are crucial to controlling how genes work.

Numerical and structural chromosome abnormalities

Chromosome abnormalities can be **numerical** (wrong number of chromosomes, e.g. **Helen Howard** (Figure 2.10), **Isabel Ingram** (Figure 2.13) and the examples in Box 2.3) or **structural**, when one or more chromosomes contain the wrong DNA, as with **Ellen Elliot** (Figure 2.14). The book by Gardner, Sutherland and Shaffer (2011) gives an in-depth treatment of the origins and implications of chromosomal abnormalities. Here we discuss the essentials.

Numerical abnormalities

These are of two types:

- **Errors of ploidy** are errors where there are the wrong number of complete sets of chromosomes. Normal cells are **diploid** ($2n = 46$ chromosomes). Gametes are **haploid** ($n = 23$). Occasionally two sperm fertilize one egg, producing a triploid ($3n = 69$). Triploids can also be produced if the whole meiotic process fails, resulting in a diploid gamete that then fertilizes a normal haploid gamete. As mentioned in Box 2.3, triploidy in humans is a common error at conception, but triploids virtually never survive to term. Tetraploidy results when a cell replicates its DNA but then does not divide. Tetraploidy and higher degrees of ploidy (polyploidy) may be seen in individual cells, but not in a whole person.
- **Aneuploidy.** All the above abnormalities involve cells with complete sets of chromosomes, which are called **euploid**. **Aneuploid** cells have just one or more single chromosomes extra or missing. Cells or people with one chromosome extra or missing are trisomic or monosomic for that chromosome. Tetrasomy and nullisomy are also possible. The reasons why aneuploidy causes clinical problems are discussed in Chapter 6.

Structural abnormalities

These (see Figures 2.16, 2.19 and 2.20) include:

- **Reciprocal translocations.** These arise when any two chromosomes swap non-homologous segments (see Figure 2.16). A carrier of a balanced reciprocal translocation is at risk of producing offspring with trisomy of one of the translocated segments and, at the same time, monosomy of the other (Figure 2.17). This was the problem in **Case 5 (Elliot family)**.
- **Robertsonian translocations.** These involve a translocation between two acrocentric chromosomes (13, 14, 15, 21 or 22) with the breakpoints in the proximal short arms, just above the centromere (Figure 2.20). A carrier of a Robertsonian translocation is themselves entirely normal, but is at risk of producing a conceptus with either complete trisomy or complete monosomy for one of the chromosomes involved. For example, somebody carrying a Robertsonian translocation involving chromosome 21 is at risk of producing a child with trisomy 21. About 3–4% of Down syndrome cases are due to such translocations. Phenotypically such a child will be indistinguishable from any other child with Down syndrome, but the recurrence risk is much higher. That is an additional reason why **Helen Howard (Case 8)** was karyotyped, even though there was little doubt about the diagnosis of Down syndrome.

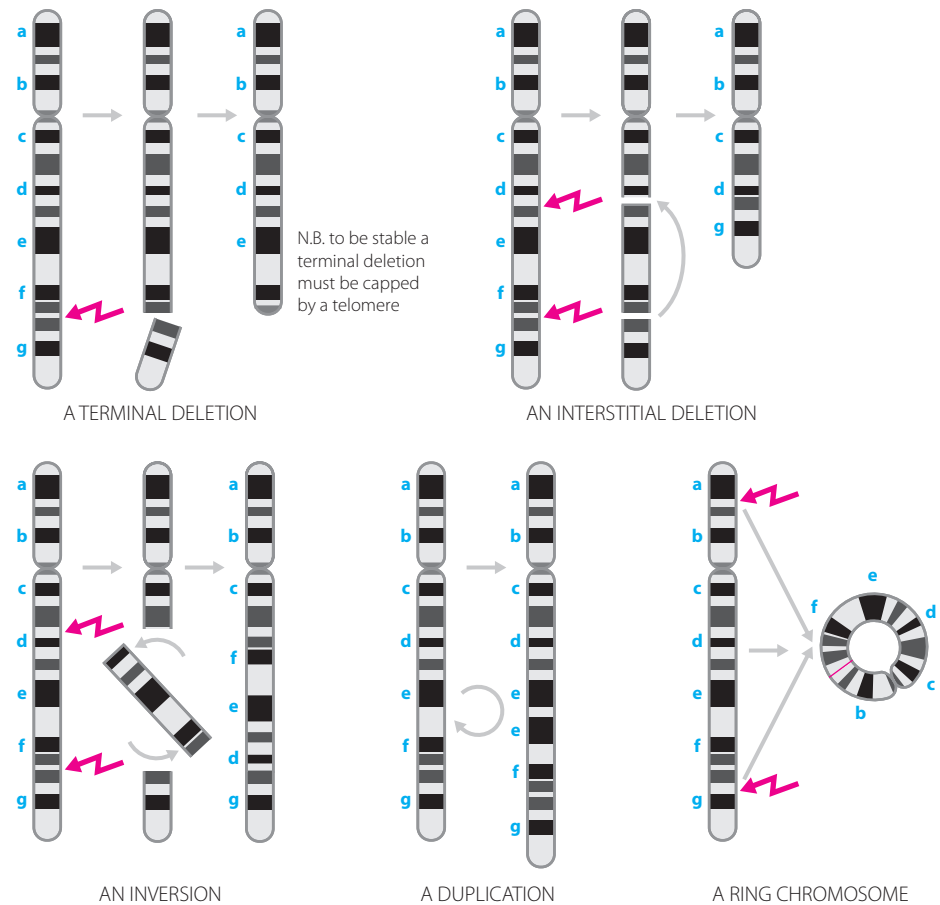


Figure 2.19 – Structural chromosome abnormalities.

These can arise either through mis-repair of chromosome breaks or through inappropriate function of the genetic recombination machinery. See *Figures 2.16* and *2.20* for translocations.

- **Deletions** can be interstitial or terminal, though a chromosome must always have telomeres, so any stable terminal deletion must have somehow acquired a telomere. Ring chromosomes (see *Figure 2.19*) are a special type of terminal deletion. Deletions generally have severe effects, and large deletions are lethal.
- **Duplications** normally have less severe effects than the corresponding deletions.
- **Inversions** can be of any size. If they involve the centromere (pericentric inversions) they change the overall chromosome shape; if not (paracentric inversions) they can only be detected by careful examination of the banding pattern.

Copy number variants (CNVs)

Duplications or deletions of chromosomal segments large enough to be visible under the microscope have severe, often lethal, effects. For many years it was therefore assumed that most of the genetic variation between normal healthy people would consist of changes of single nucleotides in the DNA. The new technique of comparative genomic

hybridization (described in *Section 4.2*) showed that this view was wrong. Deletions and duplications ranging from a few nucleotides up to one megabase (1 million nucleotides) are common in normal healthy people. In 2009 Itsara and colleagues used this then novel technique to catalog copy number variation in 2493 healthy individuals. The results (*Figure 2.21*) show a remarkable diversity, quite unsuspected previously. Only some parts of the genome can vary in this way without causing problems. Clinical geneticists rely on databases of pathogenic and non-pathogenic CNVs to interpret the significance of a variant found in a patient.

Balanced and unbalanced abnormalities

A chromosomal abnormality is described as **balanced** if there is no material extra or missing: the DNA is just divided into packages incorrectly. Provided each chromosome has a single centromere and proper telomeres, a cell can proceed normally through mitosis. Thus a fertilized egg with a balanced abnormality should be able to develop into a normal adult. **Ellen Elliot** (*Figure 2.14*) is an example. However, if there is pathogenic extra or missing material the abnormality is described as **unbalanced**. Microarray-based techniques (see *Chapter 4*) cannot detect balanced abnormalities.

The distinction between balanced and unbalanced abnormalities is a useful tool for thinking about the consequences of chromosomal variants, but it becomes fuzzy when pushed too far. Robertsonian translocations are regarded as balanced even though two acrocentric short arms have been lost (*Figure 2.20*). The short arms of all the acrocentric chromosomes carry similar genes (sequences encoding ribosomal RNA), and losing two

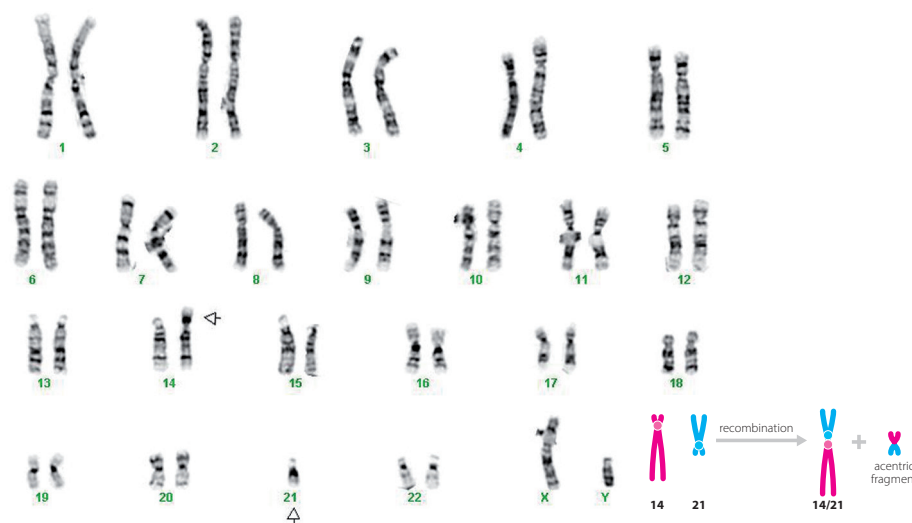


Figure 2.20 – A Robertsonian translocation.

The inset shows how this common type of chromosome abnormality arises. The short arms of all the acrocentric chromosomes (13, 14, 15, 21, 22) contain similar DNA. Inappropriate recombination between two non-homologous chromosomes produces the fusion chromosome, which functions as a normal single chromosome in mitosis. The breakpoints are just above the centromere, so that the fusion chromosome actually has two centromeres, but they are very close together and function as one. The small acentric fragment comprising the two distal short arms is lost.

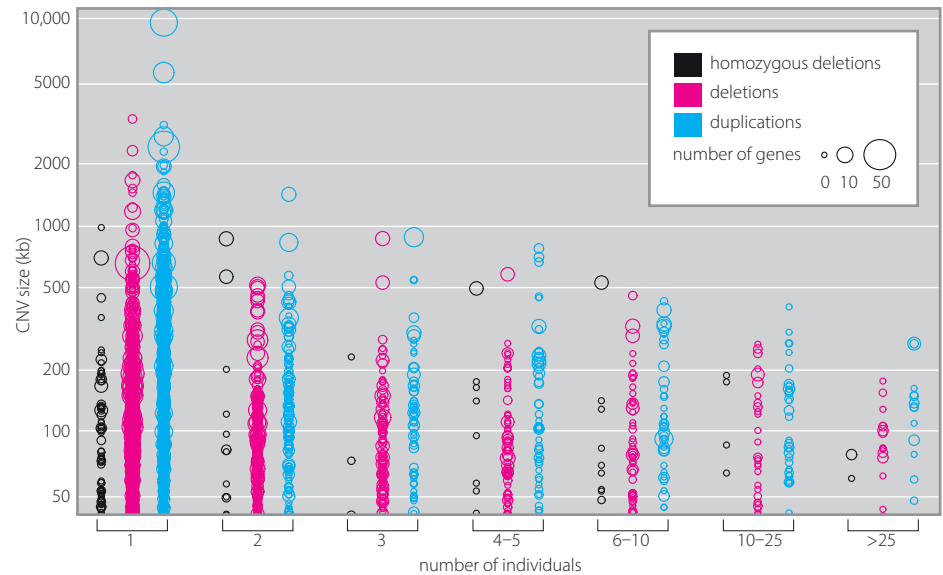


Figure 2.21 – Non-pathogenic copy number variants found in a survey of 2493 healthy individuals.

Variants are classified according to type (duplications, heterozygous deletions, homozygous deletions), size (y axis), frequency (x axis), and number of genes involved (size of circles). Common variants were less than 500 kb in size, but much larger variants were seen in rare individuals. Adapted from Itsara A *et al.* (2009) *Am J Hum Genet* **84**: 148–161, with permission from Elsevier.

of the ten short arms has no phenotypic effect. The many CNVs that can be found in normal healthy people (Figure 2.21) stretch the concept of balance still further. One would not describe these variants as unbalanced. Nevertheless, much of our DNA does have to be present in the correct quantity and **Helen Howard** (Figure 2.10) and **Elizabeth Elliot** (Figure 2.15) exemplify the problems of unbalanced chromosomal abnormalities.

Balanced abnormalities become important in meiosis when homologous chromosomes pair up. Things can go wrong in two ways.

- Translocated chromosomes have segments derived from more than one original chromosome. During prophase of meiosis I they will pair with multiple partners, forming trivalents or quadrivalents (associations of three or four chromosomes) instead of bivalents (associations of two homologs). These are liable to segregate incorrectly at anaphase (Figure 2.17)
- Chromosomes with inversions pair with a single partner forming a looped structure (Figure 2.22), but if a crossover occurs within the loop the recombinant chromosomes are abnormal (see also SAQ 4).

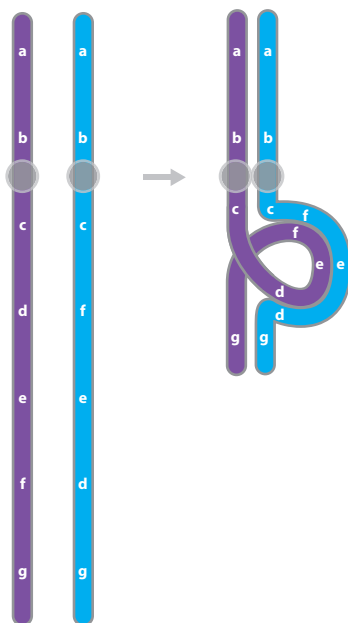


Figure 2.22 – Chromosome pairing at meiosis I in a person who has an inversion in one of a pair of homologous chromosomes.

As far as possible matching sequences (indicated by the letters) pair. In itself this causes no problems, but if there is a crossover within the loop the resulting gamete will be unbalanced.

Although balanced abnormalities do not normally affect a person's phenotype, there are exceptions.

- One or more of the breakpoints may slice through a gene. This will prevent that gene from working, which may or may not matter to the individual.
- Sometimes, as described in *Chapter 3*, a breakpoint does not disrupt a gene itself, but separates a gene from a control element located some way away on the same DNA strand. Again, this may prevent the gene from working.
- Occasionally, when a misplaced recombination or DNA repair joins together segments from different chromosomes, the join creates a novel gene out of parts of two genes that were located near the breakpoints on the different chromosomes. Such chimeric genes are important in cancer (*Chapter 7*).
- Finally, balanced X;autosome translocations cause special problems because of X-inactivation. This is discussed in *Chapter 11*.

Constitutional and mosaic abnormalities

Any genetic variant can be present either in constitutional form (that is, present in every cell of a person) or in mosaic form (present only in some cells). Mosaicism has already been discussed briefly in *Chapter 1* and in this chapter in connection with **Case 9 (Ingram family, Turner syndrome)**. It is always the result of post-zygotic events: one cell in a person or embryo consisting of many cells undergoes some change (note, however, that mitochondrial heteroplasmy can be transmitted through the egg, because each egg cell contains a large number of mitochondria, see *Chapter 1*). Any chromosomal abnormality (except triploidy) that can arise in constitutional form through a meiotic error can also arise in mosaic form through a mitotic error. Many abnormalities that would be lethal if present in constitutional form can survive in mosaics. For example, a patient may have mosaic trisomy 8, but is unlikely to have full constitutional trisomy 8.

Mosaicism is not restricted to chromosomal abnormalities. Small-scale DNA sequence changes can also be present in mosaic form – they are just harder to spot. Mosaicism is most easily detected by techniques that look individually at each of a large number of cells. Thus cytogeneticists, scanning a spread of dividing cells on a microscope slide, have long been familiar with chromosomal mosaicism. The DNA-based techniques described in *Chapters 4* and *5* normally use the pooled DNA from thousands of cells. Here, mosaicism shows as a faint extra band on a gel or a minor variant feature in the sequencer output, and is hard to detect unless present in a substantial fraction of the cells used. Special techniques can detect extremely low-level mosaicism in DNA, but they depend on knowing what precise variant you are looking for. Sometimes a pedigree will include features that hint at mosaicism – we saw an example in *Chapter 1* (*Figure 1.16*). With the increasing sophistication of genetic analysis, mosaicism is being seen in more and more clinical situations. The whole topic is considered in more detail in *Disease box 6*. Mosaicism is also fundamental to the whole of cancer (see *Chapter 7*): a tumor is made up of cells that have acquired growth-promoting genetic variants that were not present in the normal cells of the patient.

A microdeletion syndrome: Williams–Beuren syndrome

Children with Williams–Beuren syndrome (WBS, OMIM 194050) are recognized by their combination of learning difficulties, small size and characteristic face with full lips and cheeks and short nose (*Box figure 2.1*). Many have a heart problem, supraventricular aortic stenosis (SVAS), which may require surgery, and some suffer from infantile hypercalcemia, which worsens their failure to thrive and may lead to kidney damage, but which tends to resolve spontaneously. WBS occurs sporadically; there is almost never any family history. About 1 in 20 000 births is affected in both sexes and all ethnic groups (Pober, 2010).

All WBS patients have a microdeletion on one of their copies of chromosome 7 that takes out 1.55 (or occasionally 1.84) Mb of DNA. The deletion is a recurrent one, almost always arising *de novo* in each new case. The reason why the same pathogenic deletion occurs over and over again is that the deleted region is flanked by almost identical repeated sequences (*Box figure 2.2*). When the two copies of chromosome 7 pair during meiosis, occasionally these repeats mispair. If there is then a crossover between the mispaired repeats, the result is to generate one chromosome carrying the WBS deletion and another carrying a duplication of the region.

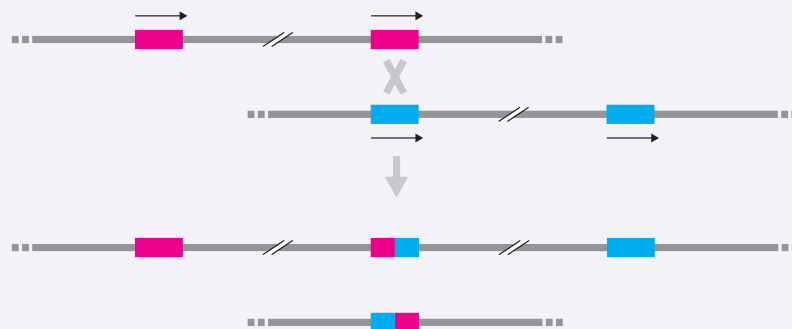
This mechanism, **non-allelic homologous recombination** (the recombination is between sequences that are homologous, i.e. they have the same sequence but are not alleles), is responsible for many other recurrent microdeletion or microduplication syndromes.

The human genome is rich in low-copy repeats, and whenever these occur in a tandem orientation on the same chromosome, they predispose that region to deletion or duplication.

Although the WBS deletion is too small to be seen by conventional cytogenetics, it includes 25 genes. The question then arises, which genes are responsible for which features of the syndrome? For SVAS, the answer is the elastin gene. People with point mutations affecting just the elastin gene show SVAS similar to that seen in Williams syndrome. For other features of the syndrome the answer is less clear. A particular fascination of WBS comes from the cognitive and behavioral phenotype that affected individuals display. Although WBS children have a global IQ similar to children with Down syndrome (usually in the range 40–85), their pattern of abilities and deficiencies is strikingly different. They have relatively good verbal skills which contrast with their poor spatial skills. Asked to draw an object or copy a diagram, a WBS child will crudely reproduce the details but fail to integrate them into any overall picture (*Box figure 2.3*). On the other hand, they can be very eloquent verbally. They also have characteristic behavior and personality traits. They manifest anxiety in unfamiliar situations but may be inappropriately friendly to strangers.

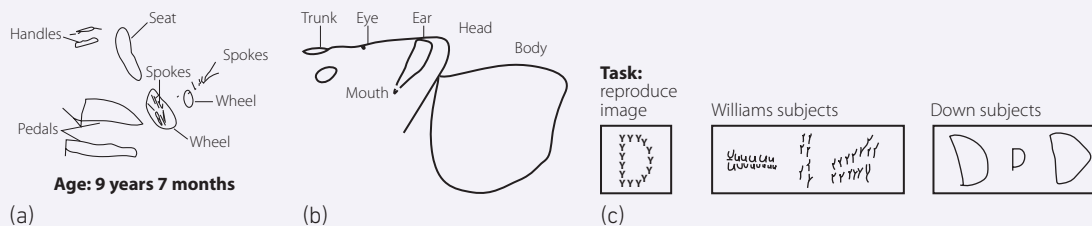


Box figure 2.1 – Williams–Beuren syndrome.



Box figure 2.2 – Generation of a deletion and the reciprocal duplication by non-allelic homologous recombination.

The WBS region is flanked by 300–500 kb stretches of DNA that show greater than 98% sequence homology (boxes). These occasionally mispair during meiosis, allowing recombination between them.



Box figure 2.3 – Drawings by people with WBS.

(a) A bicycle and (b) an elephant. In each case the drawing consists of disconnected parts (labeled by the tester), not integrated into a whole. (c) Children with WBS and children with Down syndrome matched for age and IQ were asked to copy the diagram on the left. Again, note how the WBS children see the details but not the overall design. Illustrations reproduced with permission from Dr Ursula Bellugi, The Salk Institute for Biological Sciences.

Many researchers hope that this syndrome will give us clues to the more general genetic determinants of normal speech acquisition, cognition and behavior.

2.5. References

- Gardner RJM, Sutherland GR and Shaffer LG** (2011) *Chromosome Abnormalities and Genetic Counseling*, 4th edition. Oxford University Press, Oxford. Gives an in-depth treatment of the material covered in this chapter.
- Itsara A, Cooper GM, Baker C, et al.** (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**: 148–161.
- Pober BR** (2010) Williams–Beuren syndrome. *New Engl. J. Med.* **362**: 239–252.
- Shaffer LG, Slovak ML and Campbell LJ**, eds (2009) *ISCN 2005, An International System for Human Cytogenetic Nomenclature*. S. Karger AG, Basel. The definitive reference for cytogenetic nomenclature.
- Sudmant PH, Kitzman JO, Antonacci F, et al.** (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**: 641–646.
- Trask BJ** (2002) Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Rev. Genet.* **3**: 769–778. A broad review of developments in human cytogenetics since 1956.

More detail on mitosis and meiosis can be found in any cell biology textbook.

Useful websites

The Online Biology Book by MJ Farabee has clear and straightforward descriptions and diagrams of mitosis and meiosis (although some of the links to external resources are no longer working):

www2.estrellamountain.edu/faculty/farabee/BIOBK/BioBookTOC.html

The website of Hironao Numabe at Tokyo Medical University contains many (English language) graphics and animations relevant to this chapter:

www.tokyo-med.ac.jp/genet/index-e.htm

A website from the University College Dublin on 'Understanding genetics and rare diseases' includes a set of basic animations explaining the nature and consequences of a variety of chromosome abnormalities:

www.ucd.ie/medicine/rarediseases/understandinggeneticdisorders/

2.6. Self-assessment questions

- (1) Which meiotic divisions in which parent could, by non-disjunction, potentially produce a child with
 - (a) 45,X [Guidance provided for this one at the back of the book.]
 - (b) 47,XXY
 - (c) 47,XYY?
- (2) Draw the quadrivalent and possible gametes and conceptuses from the following translocations.
 - (a) t(2;4)(q22;q32) [Guidance provided for this one at the back of the book.]
 - (b) t(5;10)(p14;p13)
 - (c) t(7;9)(q32;p21)
- (3) Using the diagrams of quadrivalents that you drew for the previous question, put in a crossover between the translocated and normal chromosomes in each of the paired segments. Work out the consequences. (This is not done just to give you a headache – there is normally at least one crossover per chromosome arm during meiosis.)
- (4) Consider the carrier of a balanced Robertsonian 14:21 translocation whose karyotype is shown in *Figure 2.20*. During meiosis the translocation chromosome will form a trivalent with the normal chromosomes 14 and 21. Draw the possible ways this can segregate when a gamete is formed, and the consequences of each of these for any conceptus.
- (5) Work out the possible gametes, conceptuses and live-born babies that a carrier of a balanced Robertsonian 21:21 translocation married to a chromosomally normal person could have.
- (6) Considering the inversion heterozygote shown in *Figure 2.22*, work out the consequences of a crossover occurring either within the inversion loop or outside it. Is it different if the centromere lies within the inverted segment (a pericentric inversion) rather than outside it as in the figure (a paracentric inversion)?

03

How do genes work?

Learning points for this chapter

After working through this chapter you should be able to:

- Name the bases, sugars and nucleosides that form normal DNA and RNA, and sketch a DNA double helix, marking in base pairs and the 5' and 3' ends
- Draw diagrams showing the principles (but not the detailed enzymology) of DNA replication, of transcription, of splicing of the transcript and of the way an mRNA sequence specifies the amino acid sequence of a polypeptide
- Describe the general features of the nuclear and mitochondrial genomes
- Sketch a typical human gene structure, showing exons, introns, the promoter, the start and stop codons, the 5' and 3' untranslated regions and splice sites
- Describe in outline the role of the promoter, enhancers, transcription factors and chromatin structure in determining gene expression

3.1. Case studies

CASE 10 O'REILLY FAMILY

- Orla has severe myopia, short stature and hip problems
- Family history of similar problems
- ? Stickler syndrome

57 70 134 158 395

Orla O'Reilly, who is married to Raymond, is only 4' 11" (150 cm) tall and has worn glasses for severe myopia since she was a young child. Her brother Oliver is also short and myopic, and he was born with a cleft palate. He also wears hearing aids. They take after their father who is short and stocky and has recently needed bilateral hip replacements; when he was 35 years old he underwent surgery for a retinal detachment in one eye, and laser treatment to prevent a detachment in the other eye. Orla went for a medical check for an insurance policy and mentioned that she had been getting some hip pain. The doctor doing the medical had just been on a genetics course and, after taking a family history, thought there might be a connection between Orla's and Oliver's medical problems and those of their father, so he referred her to the genetics clinic. There Orla had a detailed examination including an eye test which, in addition to the myopia, showed she had paravascular lattice retinopathy. The doctor also noted she had a short nose with a flat nasal bridge and rather knobby joints. He told Orla he suspected she might have a condition called Stickler syndrome. This condition is due to mutations in genes encoding components of either type II or type XI collagen (see Snead and Yates (1999) for a review).

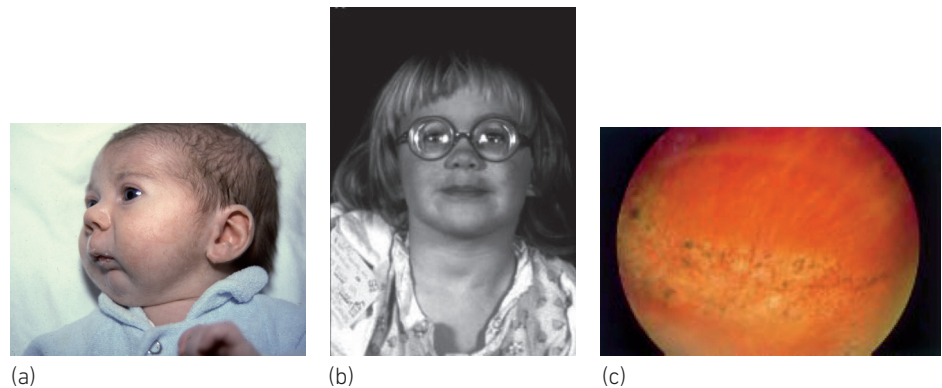


Figure 3.1 – (a) A baby with Stickler syndrome. Note the small jaw (often associated with cleft palate) and rather flat face with prominent eyes. (b) Facial features of a 4-year-old child with Stickler syndrome. (c) Typical pigmented paravascular retinal lattice degeneration. (b) and (c) reproduced from Snead and Yates (1999) with permission from the BMJ Publishing Group.

3.2. Science toolkit

How genes work is the subject of two famous hypotheses. Neither of them is wholly true, but both are nevertheless useful tools for thinking about genes.

- In the 1940s Beadle and Tatum proposed that the job of each gene was to specify one particular enzyme (the **one gene – one enzyme hypothesis**). It is not completely true because many genes specify non-enzymic proteins, and others specify functional RNA molecules rather than proteins – but in the form of ‘one gene – one polypeptide’ it remains a useful first tool for thinking about what genes do.
- Some years later, Francis Crick defined the essential function of DNA in the ‘Central Dogma’ of molecular biology (*Figure 3.2*). Genes are functional units of the DNA, and Crick’s Central Dogma states that the function of a gene is to specify the structure of a protein. The Human Genome Reference Sequence (GRCh38) lists 20 465 protein-coding genes (plus 2960 in sequences that could not be integrated into the Reference Sequence). The Central Dogma is useful but not absolutely true. Occasionally the flow of information from DNA to RNA is



Figure 3.2 – The Central Dogma of molecular biology.

The arrows don’t mean that DNA is turned into RNA, etc.; they mean the information contained in DNA is transferred to an RNA molecule, which in turn transmits its information to a protein. In other words, genes consist of DNA, but exert their effects through proteins. DNA also specifies the information in other DNA molecules (DNA replication).

reversed, when a special enzyme (reverse transcriptase) makes a DNA copy of an RNA molecule. This is a critical part of the life cycle of RNA viruses, but is not part of the mainstream metabolism of a human cell. More importantly, RNA has many functions apart from specifying the amino acid sequence of proteins. Ribosomal RNA and transfer RNA are the best-known examples of such functional RNAs, but there are many others. In fact, the number of genes specifying non-coding RNAs in the Reference Sequence (22 229) exceeds the 20 465 protein-coding genes listed (all figures taken from the ENSEMBL browser, October 2019). It remains true that most of the genes of concern to clinical geneticists specify proteins.

A full account of the structure and function of genes would take a whole book. All the processes involved are vastly complicated and involve innumerable proteins and other molecules to effect and control them. Fortunately, most of the detail is irrelevant to clinical practice. For clinicians, while there is no limit to the level of detail that it would be desirable to understand, the amount that is essential is much more manageable. The basic essential topics are explained below or in *Boxes 3.2 and 3.3* and are:

- the general structure of nucleic acids as chains of A, G, C and T (or U) units
- the double helix
- 3' and 5' ends of DNA
- the exon–intron structure of genes
- splicing of the primary transcript
- the genetic code

If you are interested, many excellent textbooks cover these topics in more detail. Additionally, there are several websites that provide freely accessible course material (see *Section 3.5*).

Structure of nucleic acids

Nucleic acids (DNA and RNA) are made of subunits called **nucleotides** linked together in long unbranched chains. Each nucleotide comprises three modules: a base, a sugar and a phosphate. DNA chains are made of only four types of nucleotide. The sugar is always deoxyribose and the base can be adenine (A), guanine (G), cytosine (C) or thymine (T). The chemical formulae are given in the last section of this chapter, but it is not necessary to know them in order to use this book. RNA is also made up of four types of nucleotide. Here the sugar is always ribose, the bases are A, G and C as in DNA, but instead of T RNA has uracil (U).

DNA, as famously described by Watson and Crick in 1953, normally exists as two polynucleotide chains wrapped round each other – the double helix. Its crucial feature is that the two chains will only zip together correctly if opposite every A in one chain is a T in the other, and opposite every G is a C. Base-pairing, A with T and G with C, explains how DNA is able to be replicated (Figure 3.3). This mechanism enables the genetic information in the mother cell to be copied and passed during mitosis to both daughter cells, as described in Chapter 2. RNA does not normally exist as double helices. This is not because of any feature intrinsic to the chemical structure of RNA, but because normal cells do not contain complementary strands of most RNA molecules, nor any enzyme that would construct an RNA strand using an RNA template. Cells use DNA and RNA to do different jobs, and they use the chemical differences between them as recognition signals for targeting enzymes to DNA or RNA as appropriate.

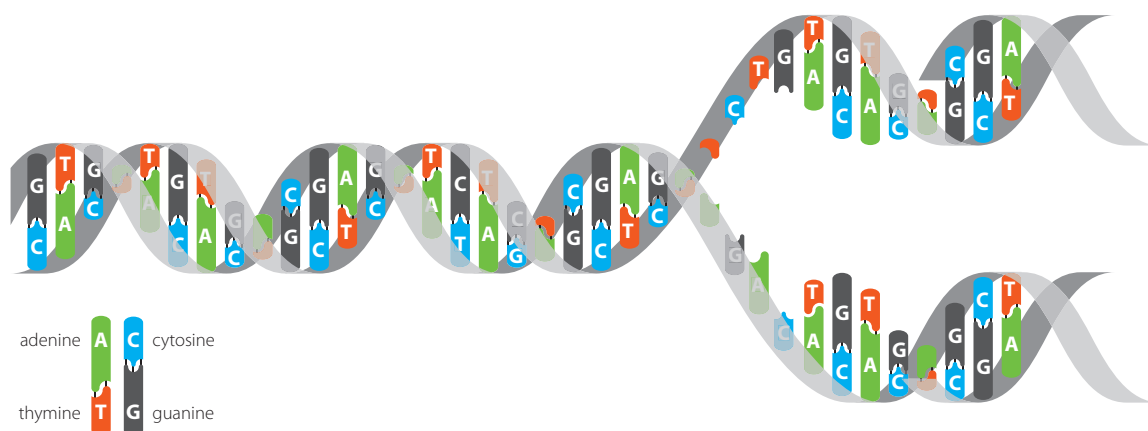


Figure 3.3 – Principle of DNA replication.

Each strand of the double helix serves as a template for synthesis of an exact replica of the other strand. A similar mechanism is used to make an RNA copy of one strand of the double helix when a gene is transcribed (U in RNA pairs with A in DNA). This diagram shows the principle and the result of DNA replication; the actual molecular events are much more complicated. In particular, this diagram ignores the fact that nucleic acid chains can only grow in the 5' → 3' direction (see *Box 3.2* and *Chapter 4*).

The human genome comprises around 3×10^9 bp of DNA (see *Box 3.1* for an explanation of the units). A normal diploid cell contains two copies of the genome (two copies of chromosome 1, two copies of chromosome 2, etc.). As described in *Chapter 2*, each chromosome contains a single immensely long double helix of DNA, packaged by histones and other proteins into chromatin. Chromosome 21, the smallest, contains 47 Mb of DNA, chromosome 1, the largest, 245 Mb. In addition to this nuclear genome, mitochondria have their own small genome, comprising a single circular DNA double helix 16 569 bp long.

A note on units

The size of a piece of DNA is measured in nucleotides (nt), base pairs (bp), kilobases (kb or kbp = 1000 bp) or megabases (Mb or Mbp = 1 000 000 bp). Because DNA is virtually always double-stranded, the distinction between bases and base pairs is often ignored when talking about the size of a piece of DNA. Thus a DNA double helix comprising one million base pairs can be described as either 1 Mbp or 1 Mb – it is not 2 Mb.

BOX 3.1

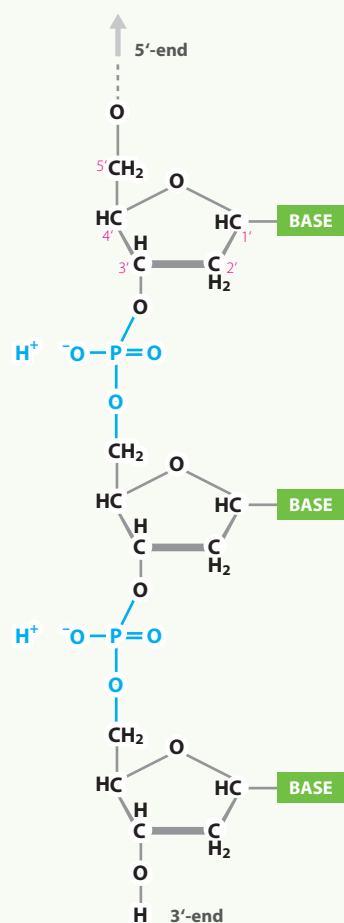
The structure of genes: exons and introns

Most genes of humans and other higher organisms are organized in a strange and unexpected way. In the continuous DNA sequence, the part that will ultimately specify the amino acid sequence of a protein is split into segments (**exons**) interrupted by non-coding sequences (**introns** or intervening sequences). The number and size of introns varies between genes without any evident logic. The average human gene has 9 exons averaging 145 bp each and the introns average 3365 bp each, but the range is very wide – see *Table 3.1*.

5' and 3' ends

To understand clinical genetics it is not crucial to know a lot about the chemical structure of DNA, but there is one non-obvious feature that does matter. Writing a DNA sequence as a string of letters like AGTTGCACG obscures one important point: the two ends are not chemically identical. Looking at the chemical formula of the sugar–phosphate backbone (*Box figure 3.1*), successive deoxyribose units are linked through the carbon atoms labeled 5' ("5 prime") and 3'. The uppermost deoxyribose has its 5' carbon free, while the lowermost has its 3' carbon free. Biochemically this difference is crucial. Enzymes that act on the 5' end of DNA will not act on the 3' end, and vice versa. The two chains in a DNA double helix are anti-parallel. If a double helix is drawn vertically one chain will have its 5' end at the top, and the other will have its 3' end at the top. The same is true of the DNA–RNA double helices that are temporary intermediates when a gene is transcribed. This has several important consequences:

- All DNA and RNA synthesis proceeds in the 5'→3' direction. That is, all the enzymes that string nucleotides together (DNA and RNA polymerases) can only add nucleotides to the 3' end of a polynucleotide. This will turn out to be crucial when we consider the polymerase chain reaction in *Chapter 4*.
- It is a universal convention that DNA or RNA sequences are written in the 5' to 3' direction. The sequence AGTTGCACG means 5'-AGTTGCACG-3'. It is just as wrong to write a sequence 3' to 5' as it is to write English from right to left. If for any reason you need to write a sequence in the 3' to 5' direction, it is essential to label the ends to make this clear.
- Only one strand of the DNA of a gene is transcribed (used as a template for synthesizing an RNA copy). This is called the **template strand**. Suppose part of this sequence reads 5'-AGTTGCACG-3'. The RNA transcript is complementary to this, with A wherever the template strand has T, G where the template has C, etc., that is, 3'-UCAACGUGC-5' (see *Figure 3.4a*; remember the strands are anti-parallel and RNA uses U in place of T. Writing it in the conventional 5'→3' direction, its sequence would be CGUGCAACU. Even with very short sequences like these, their relationship is not immediately obvious. To avoid this problem, by convention we write the DNA sequence of a gene not as the template strand but as its complementary strand – 5'-CGTGCAACT-3' in this case. This strand is called the **sense strand**. The RNA sequence is just the same as the sense strand, except for replacing T with U.
- Sequences lying 5' of a gene or sequence of interest (on the sense strand) are often referred to as **upstream**; those 3' are **downstream**.



Box figure 3.1 – DNA.

A single strand of DNA showing the 5' and 3' ends. The carbon atoms of the deoxyribose sugars are numbered 1', 2' etc. The prime (') is to distinguish them from the carbon atoms of the bases, which have their own numbering scheme (not shown here). The phosphate groups carry a negative charge – this is the basis for separating nucleic acids by electrophoresis (see *Box 4.3*).

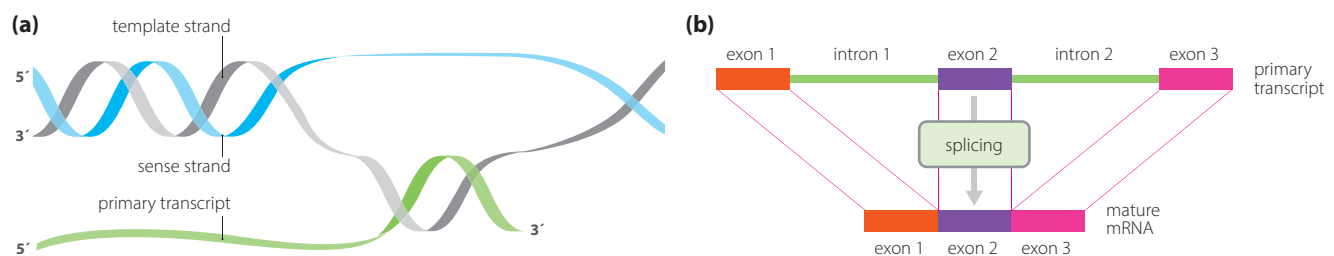
Table 3.1 – Structures of some human genes

Gene	Size in genome (kb)	No. of exons	Average exon size (bp)	Average intron size (bp)	Exons as % of primary transcript
Interferon A6 (<i>IFNA6</i>)	0.57	1	570	–	100%
Insulin (<i>INS</i>)	1.4	3	154	483	32%
Class 1 HLA (<i>HLA-A</i>)	2.7	7	160	269	41%
Collagen VII (<i>COL7A1</i>)	51	118	78	358	18%
Phenylalanine hydroxylase (<i>PAH</i>)	78	13	206	6264	3.4%
Cystic fibrosis (<i>CFTR</i>)	188	27	227	7022	3.2%
Dystrophin (<i>DMD</i>)	2090	79	178	26 615	0.7%

The number and size of introns varies very widely between genes. You can look at the size and exon–intron structure of any gene using the ENSEMBL genome browser as explained in *Box 3.7*.

Splicing of the primary transcript

When a cell needs to make a particular protein it first makes an RNA copy of one strand of the DNA of just the relevant gene (*Figure 3.4a*). Transcription is a very dynamic process, involving only scattered small segments of the genome at any one time, but varying according to the needs of the cell. The way this works is described in a little more detail in the final section of this chapter. The entire sequence, exons and introns, is transcribed to make the **primary transcript**. Within the cell nucleus this is then processed by physically cutting out the introns and splicing together the exons (*Figure 3.4b*). The RNA of the introns is broken down and appears to have no useful purpose. Splicing is accomplished within the cell nucleus by a large multimolecular machine called the spliceosome, which is a complex of proteins and small RNA molecules (see *Disease box 10*). We can safely

**Figure 3.4 – Summary of transcription.**

(a) The DNA double helix unwinds locally to allow RNA polymerase to assemble the primary transcript, and rewinds after the polymerase has moved on along the chain. The transcript is made using the template strand; its sequence is complementary to the template strand and identical to the sense strand. (b) The primary transcript is processed by cutting out introns and splicing exons together to form the mature messenger RNA (mRNA).

ignore most of the complicated molecular details, but we do need to consider how introns are recognized.

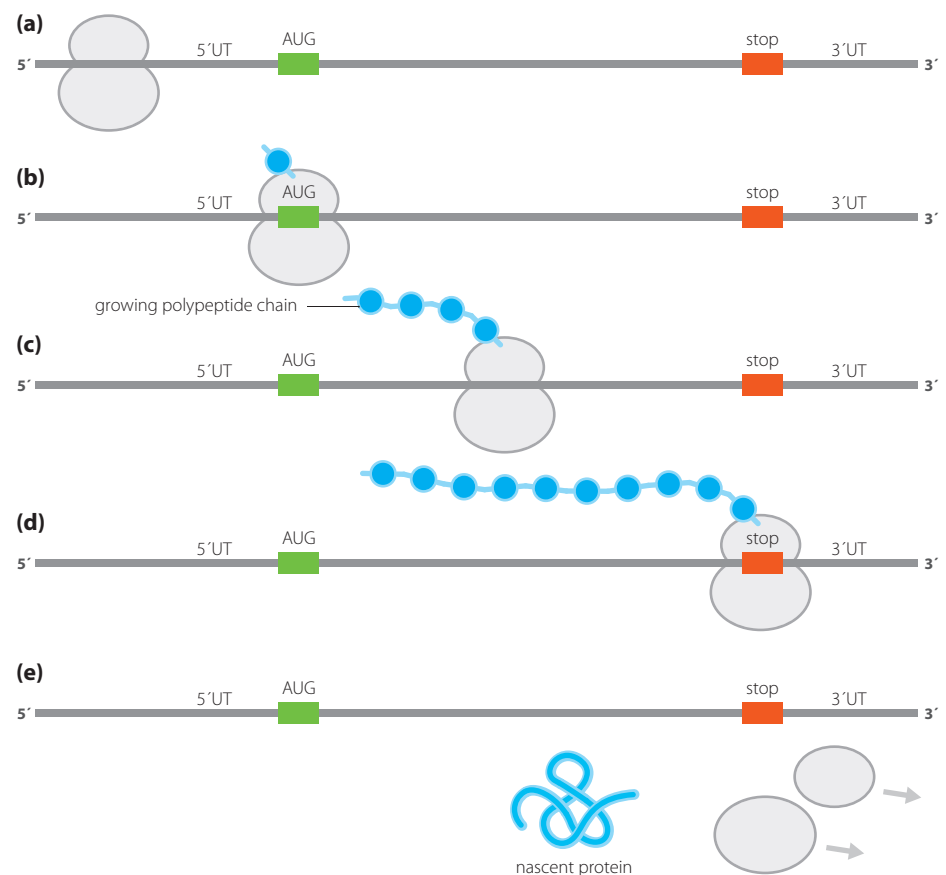
Almost all human introns start with GU (GT in the DNA of the sense strand) – this is called the donor splice site – and end with AG (the acceptor splice site). By themselves these signals would not be sufficient to define splice sites – there are innumerable GU and AG dinucleotides within exons or introns that are not used as splice sites. To be recognized as a splice site the GU or AG must be embedded in a broader **consensus sequence**. A functional splice site has a suitable combination of short sequence motifs that bind proteins or small RNA molecules in the spliceosome. These individual motifs are only loosely defined, making splice sites hard to predict by analysis of DNA or RNA sequences. This is frustrating for clinical geneticists because, as we will see later, sequence variants affecting the efficiency of splice sites are major causes of diseases.

Translation and the genetic code

The mature mRNAs, now comprising just exon sequences, are exported to the cytoplasm where they engage with a ribosome. Ribosomes are yet another large multimolecular machine including many different proteins and several species of non-coding RNA. Again, more details of the mechanism of protein synthesis can be found in textbooks or the recommended websites; for present purposes we need note only the following (see *Figure 3.5*):

Figure 3.5 – Summary diagram of translation.

(a) A ribosome attaches at the 5' end of the mRNA. (b) It moves along the 5' untranslated sequence until it encounters the AUG initiation codon. At this point it picks up the first amino acid. (c) The ribosome moves along the mRNA, picking up amino acids according to the codons of the mRNA and incorporating them into the growing polypeptide chain. (d) When it reaches a stop codon, translation of the message is complete. (e) At this point the ribosome falls off the mRNA and dissociates into its two subunits, and the polypeptide is released. The mature functional protein must be correctly folded, perhaps chemically modified and transported to the appropriate intracellular or extracellular location.



- Amino acids are specified by successive triplets of nucleotides (codons) in the mRNA. The details of the genetic code can be found in *Table 6.1*.
- Ribosomes attach to the 5' end of the mRNA and move down it in a 5'→3' direction.
- The coding sequence starts some way downstream of 5' end of the mRNA. It is inaugurated by an invariant AUG embedded in a broader consensus sequence (the Kozak sequence). The parts of the mRNA between the 5' end and the start codon are called the 5' untranslated sequence (5'UT).
- The initiating AUG sets the reading frame. The concept of the reading frame is best explained by an example (see *Box 3.3*).
- The ribosome slides along the mRNA, adding the appropriate amino acid to the growing polypeptide chain in accordance with the genetic code. Amino acids are ferried to the ribosome by a family of small RNA molecules, the **transfer RNAs** (tRNA).
- The ribosome works its way along the mRNA until it encounters a **stop codon**. There are three stop codons, UAG, UAA and UGA. When the ribosome meets a stop codon it releases the polypeptide it has made and falls off the mRNA. The parts of the mRNA downstream of the stop signal comprise the **3' untranslated sequence** (3'UT). They are important for regulating the stability of the mRNA.

The reading frame

Consider the following string of letters:

ISAWTHEBIGBADDOGEATTHECAT

We can read successive triplets of letters in three alternative reading frames:

ISA WTH EBI GBA DDO GEA TTH ECA T ... or:

I SAW THE BIG BAD DOG EAT THE CAT ... or:

IS AWT HEB IGB ADD OGE ATT HEC AT

Similarly, the string of nucleotides in an mRNA molecule could be translated by a ribosome in three different reading frames, only one of which gives a sensible message (i.e. encodes the desired protein). The reading frame of an mRNA is defined by the AUG start codon. Mutations that change the reading frame have catastrophic effects on the gene function (see *Chapter 6*).

BOX 3.3

Translation is not the end of the story

Translation finishes with the release of the newly synthesized polypeptide chain from the ribosome. However, several more processes are needed to convert the nascent polypeptide into a fully functional protein (see also *Box 3.4*).

- Folding the chain into the correct 3-dimensional structure requires no additional information – the amino acid sequence potentially dictates the folding. However, until they are correctly folded, proteins are unstable and vulnerable. It has also recently become apparent that partially folded or incorrectly folded proteins can be toxic to the cell. A number of 'chaperone' molecules assist the

folding process and protect the polypeptide during folding, while misfolded proteins are detected and degraded.

- Many proteins incorporate chemical modifications to the basic polypeptide. Often sugars are attached (glycosylation) or a variety of other small molecules. The chain may be cleaved; pairs of cysteines (see *Box 3.6* for names and formulae of the 20 amino acids, and an explanation of N- and C-termini) may be cross-linked to form S–S (disulfide) bridges that lock the structure in place; other amino acid residues may be chemically modified, for example, prolines may be hydroxylated. All polypeptides initially have an N-terminal methionine, incorporated in response to the AUG initiation codon, but very often this is cleaved off.
- Proteins must be transported to an appropriate location. The destination is often specified by a short N-terminal **signal peptide** that is removed during the process of protein sorting. In other cases the signal is a sequence of amino acids located somewhere within the chain, and is not removed. In the case of **Joanne Brown (Case 2)** it will turn out that one of her two *CFTR* genes carries a mutation that prevents the protein being correctly located in the cell membrane (her other copy of the gene carries a different mutation – that is, she is a **compound heterozygote**).
- Structural proteins may be further modified in their final location.

Collagens provide good examples of the way post-translational processing may be required to convert a nascent polypeptide into the functional protein – as described in *Box 3.4*.

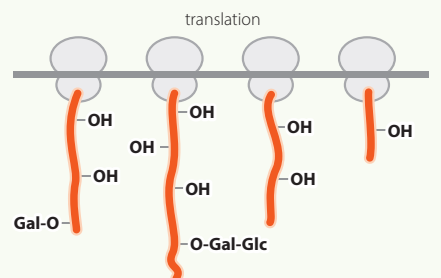
Biosynthesis of collagens

Almost one-third of the total protein mass of the human body consists of collagens. They are major proteins of the extracellular matrix, the connective and supporting tissue that holds cells together. They form cartilage, tendons, the matrix of bones and the basic support of many membranes. Humans have 30 or so collagen genes, distributed around the genome. These encode at least 27 forms of collagen, with different types being found in different tissues.

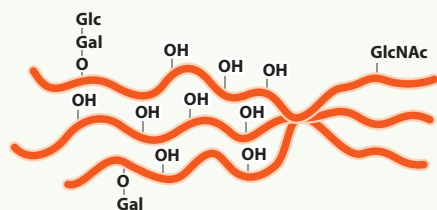
The basic collagen structure is a triple helix of three polypeptide chains wound tightly round one another. Collagens may be homotrimers (all three chains the same) or heterotrimers. For example, Type I collagen, the major collagen of skin, tendon and bone, is made of two α -1 chains, encoded by the *COL1A1* gene on chromosome 17, and one α -2 chain encoded by the *COL1A2* gene on chromosome 7. The initial product of these genes is a procollagen. The future triple helical region has a repetitive structure, Gly–X–Y, where X and Y can be any amino acid, but are often proline or lysine. This procollagen undergoes extensive post-translational modification:

- in the rough endoplasmic reticulum (see *Figure 3.9a*) a proportion of lysines and prolines are hydroxylated by special enzymes that use oxygen, Fe^{2+} and ascorbic acid as cofactors
- sugar residues are attached to some of the hydroxyl groups
- three chains are then wound together, starting at their C-terminal ends, to form the triple helix
- the resulting procollagen is secreted, after which special enzymes cleave off the C-terminal and N-terminal propeptides
- finally, the triple helical molecules are assembled into large multimers and crosslinked through lysine residues.

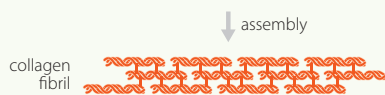
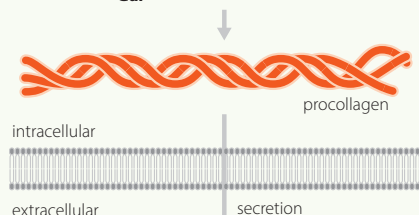
Some collagens (I, II, V, XI, XXIV, XXVII) form fibrils; collagens IV, VIII and X form meshworks that support membranes, while other collagens have various specialized functions. **Case 10 (Orla O'Reilly**, discussed below) illustrates the consequences of a mutation in one collagen gene.



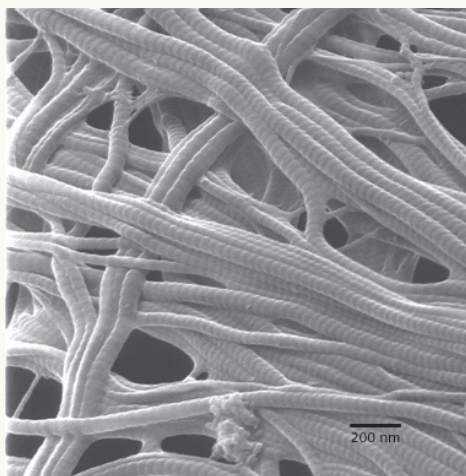
ER and Golgi modification



ER – endoplasmic reticulum
Glc – glucose
Gal – galactose
GlcNAc – N-acetyl glucosamine



(a)



(b)

Box figure 3.2 – (a) Collagen biosynthesis, from nascent polypeptide to mature fibril; (b) electron micrograph of collagen fibers (reproduced courtesy of Zeiss).

3.3. Investigations of patients

Most of the cases described so far in the book will require DNA investigations. Here we describe how each case fits into the picture of gene structure developed in the previous section.

CASE 1 ASHTON FAMILY

- John, healthy 28-year-old son of Alfred Ashton
- Family history of ? Huntington disease
- Autosomal dominant inheritance
- Need for diagnostic PCR test

1 8 **67** 103 153 395

Huntington disease is caused by a change to the coding sequence of the *HTT* gene on chromosome 4, position 4p16. The *HTT* gene has 67 exons, covers 169 kb of genomic DNA (Figure 3.6) and encodes a 3141 amino acid protein called huntingtin. Huntington disease patients always have a sequence change in exactly the same place in the gene. Part of exon 1 encodes a run of glutamines in the huntingtin protein. Every patient with Huntington disease has an expansion of this sequence, so that the encoded protein contains a much longer run of glutamines. How this makes the patient ill is discussed in *Disease box 4*. The DNA test needs to check the size of the sequence encoding this polyglutamine run.

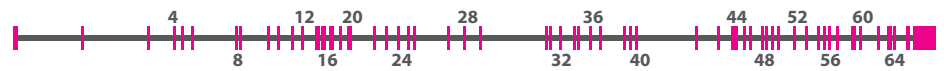


Figure 3.6 – Structure of the Huntington disease gene.

The long horizontal line represents the genomic DNA, short vertical bars represent the exons.

CASE 2 BROWN FAMILY

- Baby Joanne, recurrent infections, poor growth
- Sweat test confirms she has cystic fibrosis
- Autosomal recessive inheritance
- Need for molecular test

2 10 **67** 132 154 313 395

As mentioned previously, cystic fibrosis is always caused by mutations in the *CFTR* gene on chromosome 7. One way or another, these changes prevent the gene from encoding a functional protein. The disease is recessive, so if Joanne has cystic fibrosis both copies of her *CFTR* gene must be mutated. *CFTR* is quite a large gene, with 27 exons spread over 188 kb of chromosome 7 at position 7q31.2 (Figure 3.7). Unlike in Huntington disease, mutations may be anywhere in the gene. Different patients with cystic fibrosis may have different pathogenic variants – over 1700 such variants are described in the Human Gene Mutation Database. Almost all the variants are changes to a single nucleotide or a few adjacent nucleotides. Joanne's two variants might be the same or different. Strategies for mutation detection are discussed in *Chapter 4*.

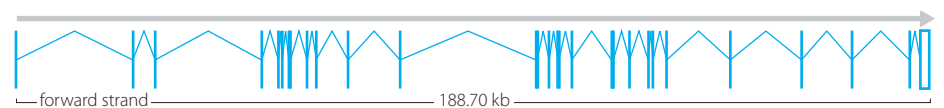


Figure 3.7 – Structure of the *CFTR* gene.

An alternative graphical display of Human Genome Project data, as displayed by the ENSEMBL genome browser. Some closely spaced exons appear as a single bar. The arrow shows the 5'→3' direction of the sense strand.

CASE 3 KOWALSKI FAMILY

- Karol, first son of Kamil and Klaudia
- Developmental delay, hypotonic, severe intellectual disability
- Difficulties of genetic testing in such cases
- Likely need for exome sequencing

3 10 **67** 102 134 155 395

The problem with intellectual disability is that, even when the cause is genetic, mutations in any one of thousands of different genes might cause it. The reason for this diversity is not hard to understand. Our brain is our most complex organ. We still understand little of how the billions of neurons and supporting cells interact to produce an organ that makes us capable of feeling and thinking – but clearly these remarkable abilities depend on the correct functioning of innumerable subsystems. If any one of these fails, the whole mechanism may fail. In most cases there is nothing in the phenotype to suggest which gene might be the cause. Without a specific candidate gene to test, there was until recently no way forward for genetic investigations. But now massively parallel DNA sequencing techniques,

CASE 4 DAVIES FAMILY

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy
- X-linked recessive inheritance
- Problems of testing dystrophin gene

collectively termed next-generation sequencing and described in *Chapter 5*, make it possible to sequence every exon of every gene in Karol's genome (his whole exome) in a single operation. Although this solves one problem it creates another: how to trawl through the massive list of variants revealed by the sequencer to home in on the one causative variant. Karol's problem could be due to homozygosity for a recessive variant, but in societies where consanguineous marriages are unusual, it is more likely due to heterozygosity for a *de novo* dominant variant. In *Chapter 5* we will follow the process of exome sequencing and there, and in *Chapter 6*, we will see how bioinformaticians tackle these challenges.



Duchenne muscular dystrophy is X-linked, so the causative gene must be located on the X chromosome. It is in the proximal short arm, at Xp21, and encodes a protein, dystrophin, that is essential for making muscle cells robust enough to withstand mechanical stresses over many years. The dystrophin gene is one of the most remarkable in the human genome. It is huge, covering over 2 million base pairs of DNA. 99.3% of this comprises introns; after they are spliced out of the primary transcript the 79 exons make a 13 kb mature mRNA. Two-thirds of cases are caused by partial deletions of the gene. Because almost all the genomic sequence is intronic, the deletion breakpoints almost always fall in introns, so their effect on the mature mRNA is to remove one or more contiguous exons (*Figure 3.8*). Searching for mutations anywhere in this vast gene would be a substantial challenge. The initial test is therefore to check the genomic DNA for missing exons. These are obvious in DNA from an affected boy; in a carrier woman they are masked by her normal X chromosome, which of course has all exons intact.

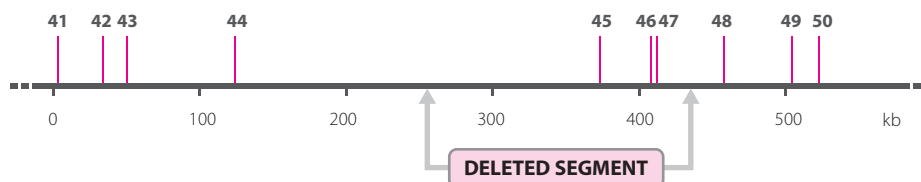


Figure 3.8 – A deletion of part of the dystrophin gene.

The figure shows a 500 kb region containing exons 41–50. These exons are all 100–200 bp long, and so if drawn to scale each exon would be represented by a line occupying less than 0.05% of the width of the figure. Random deletion breakpoints therefore almost always fall in introns. Their effect is to remove one or more complete exons from the mature mRNA. The deletion shown removes exons 45–47 from the mature mRNA, while leaving all the other exons intact.

CASE 5 ELLIOT FAMILY

- Baby girl Elizabeth, parents Elmer and Ellen
- Multiple congenital abnormalities
- Family history of reproductive problems
- ? Chromosome abnormality
- Ellen – balanced 1:22 translocation
- Elizabeth – unbalanced segregation product
- Reciprocal translocation



The cytogenetic analyses (*Sections 2.3 and 4.3*) tell us all we need to know in this family. As pointed out in *Chapter 2*, in reality the analysis of baby Elizabeth's condition would have used a DNA-based microarray test. Further DNA testing is not relevant, although if Elmer and Ellen were to request prenatal testing in a subsequent pregnancy to check whether the fetus had the same problems as baby Elizabeth, this would be done by testing the fetal DNA.

CASE 6 FLETCHER FAMILY

- Frank, aged 22, with increasingly blurred vision
- Family history of visual problems
- Possible mitochondrial inheritance
- ? Leber hereditary optic neuropathy
- Test mitochondrial genome

5

13

69

130

157

395

The pedigree and clinical phenotype suggested a diagnosis of Leber hereditary optic neuropathy (LHON). This is usually caused by mutations in the DNA of the mitochondria, rather than the DNA of the chromosomes in the cell nucleus.

The small mitochondrial genome (*Figure 3.9b*) is markedly different from the nuclear genome. In many ways it is much more similar to bacterial genomes – an observation consistent with the belief that mitochondria evolved from bacteria that lived in a symbiotic relationship inside some ancestral cell. The mitochondrial genome is circular and compact. Though only 16 569 bp long it contains 37 genes. Like bacterial genomes, the genes are tightly packed together with very little intergenic DNA and contain no introns – a marked contrast to the nuclear genome which averages only 7 multi-exon genes per megabase (20 465 genes spread over 3000 Mb of genomic DNA). However, mitochondria are very far from being independent micro-organisms. Most mitochondrial functions depend on proteins encoded by nuclear genes that are transported to the mitochondrion after being synthesized. Importantly, this means that diseases caused by mitochondrial dysfunction are not necessarily caused by mutations in the mitochondrial DNA. For example, mitochondrial DNA is replicated by DNA polymerase gamma, the gene for which (*POLG1*) is located in the nucleus on chromosome 15q25. Mutations in *POLG1* can cause Alpers syndrome (OMIM 203700), with neurodegeneration and liver failure, or progressive external ophthalmoplegia (OMIM 157640 and 258450). LHON is one of the few diseases caused by mutations in the mitochondrial DNA.

The DNA extracted from a clinical sample includes the mitochondrial as well as the nuclear DNA. The diagnosis in Frank Fletcher will be confirmed if his DNA sample can be shown to have one of the mitochondrial DNA mutations characteristic of LHON (see *Chapter 5*).

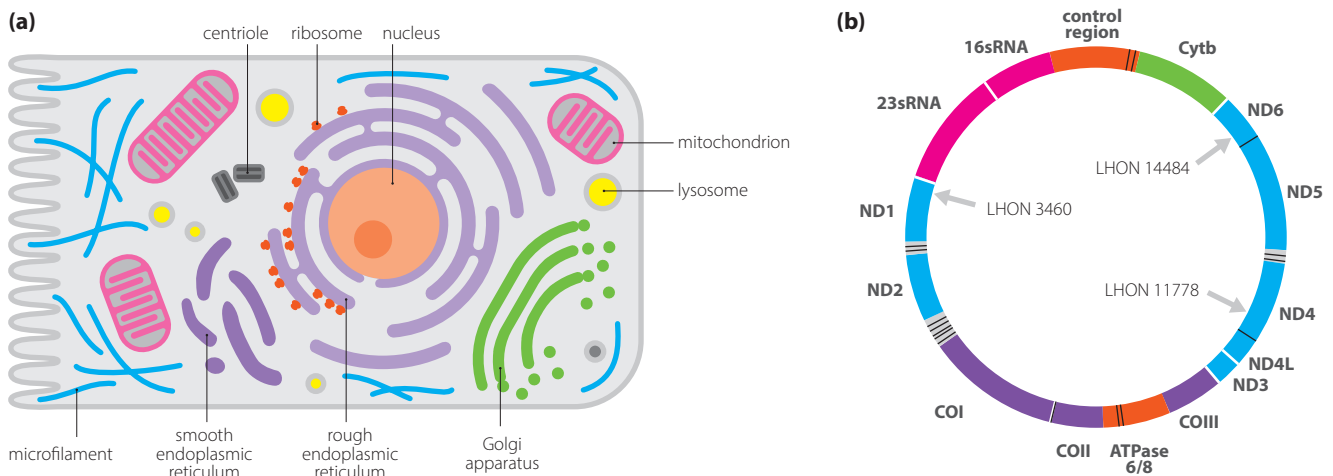


Figure 3.9 – (a) A cell showing the nucleus and mitochondria. Cells may contain 100 – 1 million mitochondria, depending on the type of cell. (b) The mitochondrial genome. The 13 protein-coding genes are labeled and the locations of the three common LHON mutations shown. The other 24 mitochondrial genes encode functional RNAs: the two ribosomal RNAs (red segments) and 22 transfer RNAs (small genes indicated by unlabeled thin lines).

CASE 7 GREEN FAMILY

- George, aged 3 years
- Developmental delay, mildly dysmorphic
- Normal 46,XY karyotype but suspect microdeletion
- Test for microdeletions

25 39 **70** 97 395

George's combination of learning difficulties and mild dysmorphism suggested a chromosomal abnormality, and his clinical features resembled those of patients with an interstitial deletion of chromosome 22 at position 22q11. Under the microscope his karyotype appeared normal (*Figure 2.8*) but a deletion of less than 3–5 Mb of DNA would not have been visible. Given an average density of protein-coding genes across the whole human genome of about 7 per Mb of DNA, a deletion that is too small to see under the microscope can still affect dozens of genes. Checking the diagnosis requires a test able to see deletion of a region of DNA that is small in cytogenetic terms but still very large in molecular terms. This will require a different technique from the previous cases, as described in *Chapter 4*.

CASE 8 HOWARD FAMILY

- Helen, newborn daughter of young parents
- Down syndrome confirmed
- 47,XX,+21 karyotype
- Options for prenatal testing

26 39 **70** 315 395

The clinical features and chromosome analysis (*Figures 2.2* and *2.10*) established the diagnosis. Many couples request prenatal testing for Down syndrome, either because, like Helen's mother, they have had an affected child, or because of the increased risk when the woman is older. Conventional cytogenetic tests can take up to 2 weeks because of the slow growth of cells in culture, forcing a very stressful wait on the parents. As explained in *Chapter 12*, new DNA-based tests allow a much faster answer.

CASE 9 INGRAM FAMILY

- Isabel, 10 years old with small stature and possibly delayed puberty
- ? Turner syndrome
- 45,X karyotype
- Risk of Y-chromosome DNA

26 42 **70** 103 285 395

Clinical features and chromosome analysis (*Figures 2.3* and *2.13*) established the diagnosis. As explained in *Chapter 2*, Isabel may have started life as a 46,XY conceptus and lost the Y chromosome in one of the early mitotic divisions. If this was the case, and if any of the cells in her streak gonads retained a Y, these cells can give rise to a malignant gonadoblastoma. Therefore it is important to check for presence of Y-chromosome DNA sequences. If any are found, then gonadectomy is usually recommended.

CASE 10 O'REILLY FAMILY

- Orla has severe myopia, short stature and hip problems
- Family history of similar problems
- ? Stickler syndrome
- Test collagen II genes

57 **70** 134 158 395

Orla's combination of high myopia and joint problems, inherited in an autosomal dominant manner, is characteristic of people who have a mutation in Type II (or occasionally Type XI) collagen. As described in *Box 3.4*, there are at least 27 different human collagens, encoded by at least 30 genes. To recapitulate, each collagen gene encodes a polypeptide, procollagen, which is subject to extensive post-translational modification. The final processed collagen molecule contains tightly wound triple helices of polypeptide chains.

Collagen II is a homotrimer, consisting of three $\alpha 1$ polypeptide chains encoded by the *COL2A1* gene on chromosome 12q13. It forms fibrils that are major structural proteins of cartilage, and are also important in the vitreous humor of the eye and in the inner ear. Confirming the diagnosis in Orla requires this gene to be sequenced. If no mutation is found in the *COL2A1* gene the next step would be to check her collagen XI. Collagen XI is a heterotrimer comprising two chains encoded by the *COL11A1* gene on chromosome 1p21 and one chain encoded by the *COL11A2* gene on chromosome 6p21.

All collagen genes have a similar structure. The ends of each gene, encoding the N-terminal and C-terminal propeptides (see *Box 3.4*) are different for each gene, but the central part, that will form the triple helix of the mature collagen, is encoded by a large number of short exons, mostly 45 or 54 bp in length. The *COL2A1* gene has 54 exons, while *COL11A1* and *COL11A2* have 67 and 66, respectively. Because Stickler syndrome is dominant we will be looking for a single mutation. In principle the causative variant might be anywhere in any of the three candidate genes but, as described in *Chapter 6*, only certain types and locations of variants are likely to produce Stickler syndrome, thus making the search somewhat easier.

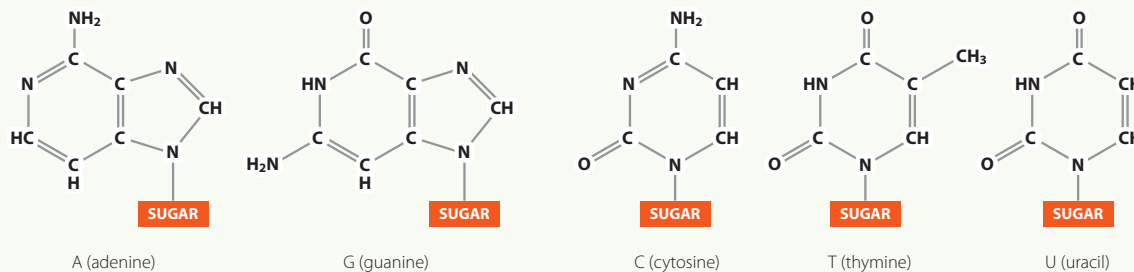
3.4. Going deeper...

Some chemistry

For reference, *Box 3.5* shows the chemical structures of the bases A, G, C, T and U, and *Box 3.6* shows the 20 amino acids that are used to make proteins.

Chemical formulae of A, G, C, T and U

The formulae and names show the bases, and indicate how each base is attached to a sugar (ribose in RNA, deoxyribose in DNA). Bases are classified into purines (A, G) and pyrimidines (C, T, U). See *Figure 5.3* for a formula of a complete nucleotide.



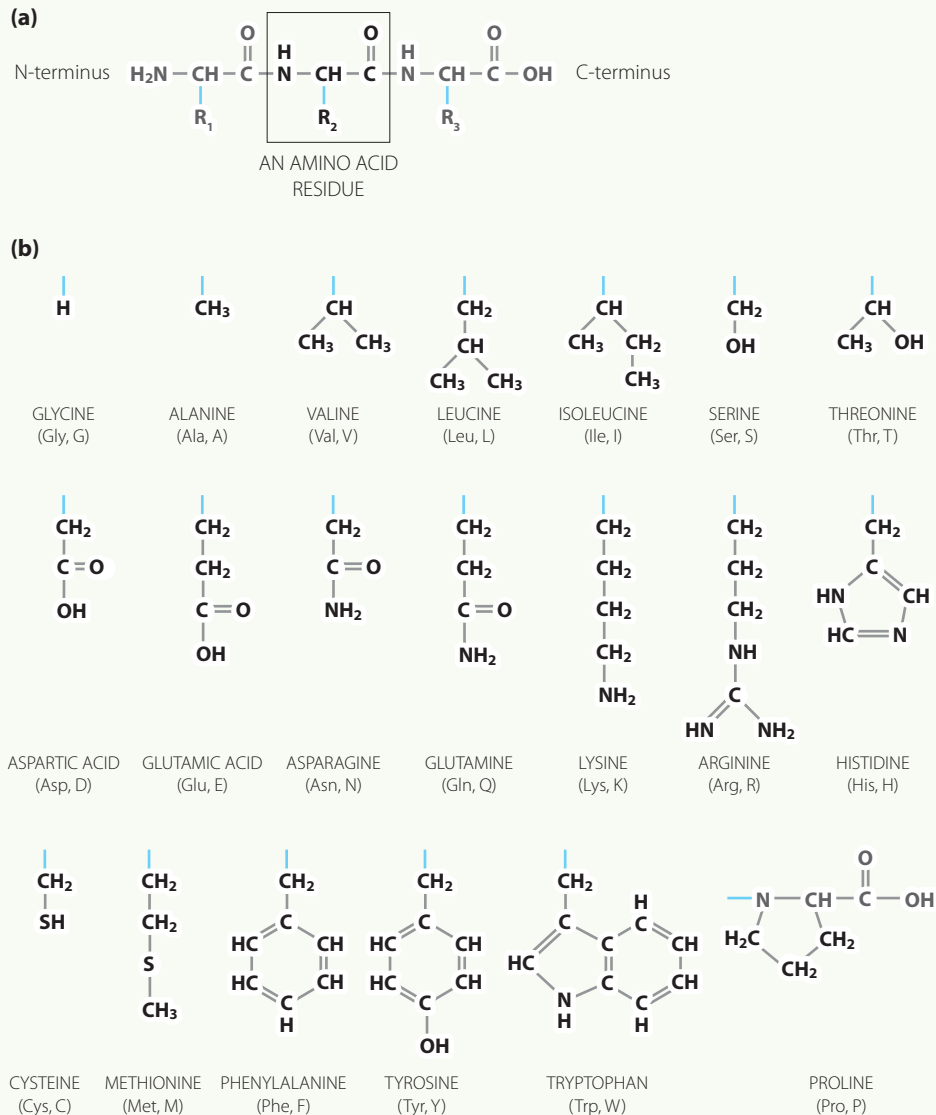
BOX 3.5

One gene often encodes more than one protein

Contrary to the one gene – one polypeptide hypothesis, a single gene can very often encode several different proteins (*Figure 3.10*). The main mechanism for this is alternative splicing. Often there is more than one way of splicing the primary transcript. Certain exons may be either incorporated or skipped in the mature mRNA, giving rise to **splice isoforms**. Sometimes in the primary transcript there are two alternative splice sites marking the beginning or end of an exon. A gene may also have several alternative promoters and first exons (the dystrophin gene has seven). All these variables mean that most genes can encode more than one protein – some can potentially encode over 1000. The average in one set of well-studied genes (from the ENCODE project, see below) is 5.4. Some of this variability may be just noise in the system, where a cell is bad at recognizing the one correct signal – but in many cases it is functional, with isoforms having distinct functions. Mutations that affect the balance of splice isoforms can be clinically significant, but hard to recognize in the DNA sequence.

Structure of proteins

(a) Chemical formula of a polypeptide. The amino acids differ according to the nature of the side chain, labeled R_1 , R_2 , R_3 here. A real protein would probably contain several hundred amino acid residues. (b) Chemical formulae of the side chains (R groups) of the 20 amino acids used to make up proteins. The three letter and single letter abbreviated names of each amino acid are shown.



Switching genes on and off – transcription and its controls

All the cells of our body contain the identical set of genes – that is the purpose of mitosis. So how do they become so very different from each other: brain cells, liver cells, skin cells, muscle cells, etc? The answer is that they express different subsets of their repertoire of genes. Differential switching of genes is crucial to development. Additionally, an

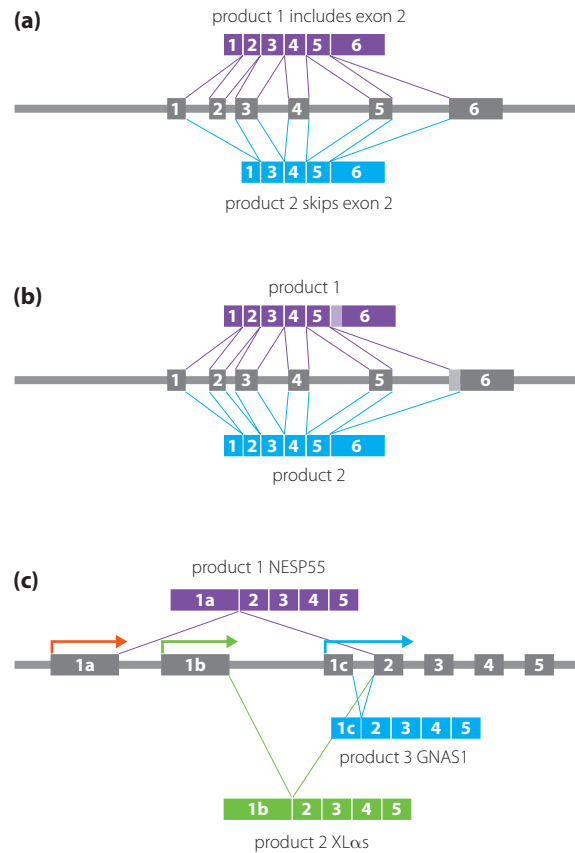


Figure 3.10 – How one gene can encode more than one protein.

(a) Exons may be variably incorporated into the mature mRNA or skipped. (b) An exon may have two alternative splice sites. (c) A gene may have two or more alternative promoters and first exons (this example shows the *GNAS1* gene, which encodes three differently named proteins). The majority of all human genes use one or more of these mechanisms to encode more than one protein.

individual cell will switch genes on or off depending on its current needs. Switching can occur at the level of transcription or translation. The major on–off controls operate at the level of transcription. Translational controls operate mainly to fine-tune gene expression.

Selectivity is the key to transcription. Although the DNA of a chromosome is a single huge molecule, the RNAs produced by transcription are a diverse collection of much smaller molecules, made by transcribing selected small segments of the DNA. Choosing appropriate segments to transcribe is crucial to the life of the cell. This is largely controlled by the way the DNA is packaged. As mentioned in *Section 2.4*, DNA is packaged by histones and other proteins into chromatin. Chemical modifications of the DNA and histones affect the local packaging and determine the accessibility of a DNA sequence to the proteins needed for transcription (see *Chapter 11* for a fuller account).

Transcription of a gene starts when a large multiprotein initiation complex has been assembled at the promoter, upstream of the 5′ end of the gene. Active promoters are defined by their ‘open’ chromatin structure and by specific modifications to histones

in nearby nucleosomes. Short sequence motifs nearby bind specific proteins – general and gene-specific **transcription factors**. These in turn bind other proteins, and so build up a complex including the RNA polymerase that does the actual transcription. Each individual DNA–protein interaction may be weak, but once several different proteins are loosely bound, protein–protein interactions between them glue the complex together. Proteins bound to distant regulatory sequences (enhancers) can be brought into the complex by DNA looping. Some of the proteins of the initiation complex are present all the time in every cell, but others (themselves the products of genes that are highly regulated) occur only in certain cells, or only when the cell responds to certain signals. By acting in different combinations, the 1600 or so human transcription factors can achieve flexible control over the expression of a much greater number of genes.

Once started, transcription proceeds until some loosely defined stop signal is reached. The end result (the primary transcript) is a single-stranded RNA molecule, typically 1–100 kb long, corresponding precisely in sequence to the sense strand of the DNA (*Figure 3.4*).

From gene to genome

The Human Genome Project was launched in 1990 and culminated triumphantly with publication of the ‘finished’ human genome sequence in 2004 – on time and under budget. The key publications (International Human Genome Sequencing Consortium, 2001, 2004) do not of course print out all 3 000 000 000 As, Cs, Gs and Ts, but they describe the methods used and the key features of our genome. The 2004 paper is quite brief and technical, but the blockbuster 2001 paper ranks with *On the Origin of Species* and Watson and Crick’s 1953 description of the double helix as one of the seminal milestones in biology.

The raw sequence is held in freely accessible public databases. To access it you use a genome browser program. Several of these are freely available on the internet as a public service. Widely used ones include the SANTA CRUZ and ENSEMBL browsers. *Box 3.7* shows how to use ENSEMBL to view the intron–exon structure, DNA sequence and chromosomal environment of any gene.

The ‘finished’ human Reference Sequence was obtained by piecing together literally millions of short runs of sequence, derived from several different anonymous donors. Thus it does not represent the sequence of any one person. Indeed, probably nobody has exactly this sequence. Instead, it functions as a stable reference, against which the millions of variants in individuals can be catalogued. Rapid progress in DNA sequencing technology has brought the cost of sequencing the entire genome of an individual down to around \$1000. As a result, a rapidly increasing number of individual genome sequences – already (2020) over a million – are available on the internet, giving us an unprecedented insight into the way the genomes of normal healthy individuals can vary. For clinicians, this information provides an essential background for interpreting the likely effect of variants found in patients.

The human genome sequence is a stunning achievement – but it is important to understand its limitations. In fact, by itself, the raw sequence does not tell us very much. It needs to be **annotated** to identify the genes and other functional elements contained in the sequence. Annotation is based on a combination of laboratory experiments to

How to use the ENSEMBL genome browser

One of the most frequent uses of a genome browser is to find information about the intron–exon structure and sequence of a gene. Here is how the ENSEMBL browser does this for the *CFTR* gene (based on access on 6 May 2019). The SANTA CRUZ browser (www.genome.ucsc.edu) offers similar tools.

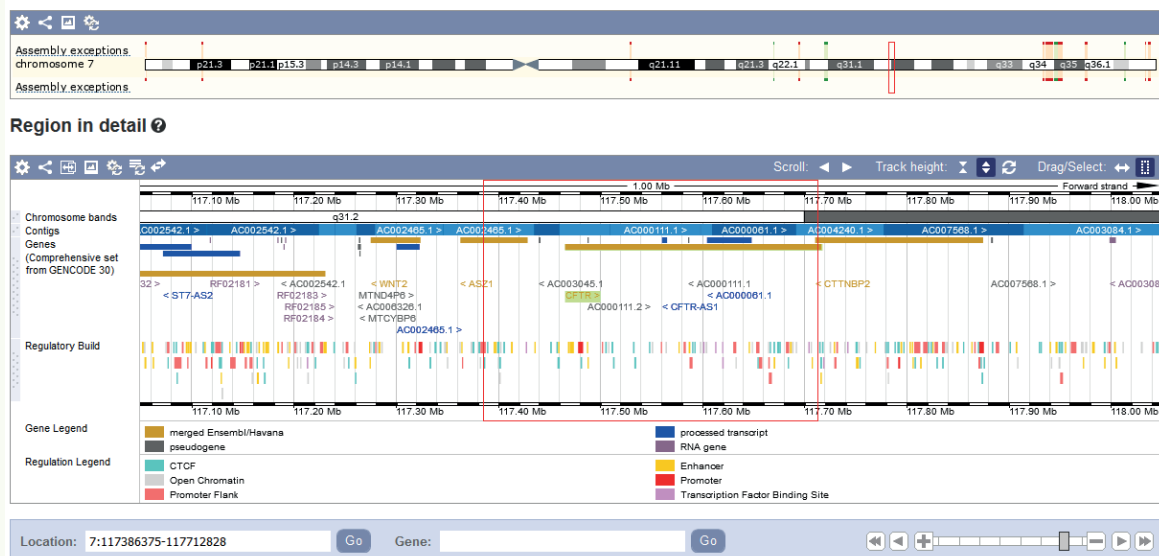
Go to www.ensembl.org/Homo_sapiens/Info/Index

1. Fill in the search box to get ENSEMBL to look for a gene with the appropriate name – in this case, *CFTR*.
2. If you have got the name right, this will produce a long list of possible entries, including (probably at the top of the list) '*CFTR* (Human Gene) ENSG0000001626'
3. Clicking on this brings up a page including a diagram showing all the different *CFTR* gene transcripts (currently 18) that are present in databases. Like most human genes, the *CFTR* gene can give rise to a number of different transcripts, many of which are non-functional. The diagram shows the way the exons of each transcript are distributed across the genomic DNA.
4. Clicking on a transcript opens a small box giving some extra information and links to exons, cDNA sequence and the protein.
5. Clicking on the 'exons' entry in the box brings up a list showing the position in the Reference Sequence of each exon, its sequence and (depending on the options chosen) extensive documentation of all the reported variants in each exon.
6. Side menus on the 'exons' page give you access to a great variety of further information, including the amino acid sequence of the encoded protein and reported variants. The 'Summary' item links to a diagram of the exon–intron structure of the chosen transcript (Figure 3.7).

All the various displays are customizable, allowing you to select the information you wish to be displayed.

An alternative common use of the browser is to view genes and other features of a particular chromosomal region. Staying with the example of the *CFTR* gene, when you have selected your gene (stage 3 above), a tab at the top of the display reads 'Location 7:117,465,784 – 117,715,971'. Clicking on this produces an extensive display including the diagram in Box figure 3.3.

Chromosome 7: 117,386,375–117,712,828



Box figure 3.3 – Screenshot of the chromosomal region around the *CFTR* gene (red box) as shown by ENSEMBL.

The display, currently of 1 Mb, can be zoomed in or out, and many more user-selectable tracks of information can be displayed. Clicking on items in the display brings up further information and links.



Figure 3.11 – Domain architecture of proteins.

Each line represents one family of proteins and each colored shape represents a structural and functional protein domain. The homeodomain (green) is found in all the proteins shown here, together with a variety of other functional domains. These in turn can be found in different combinations in other proteins.

identify transcripts and computer analysis of the sequence to identify features such as protein-coding exons, RNA genes and regulatory elements. Genome browsers like ENSEMBL use huge amounts of annotation data to turn the raw sequence into useful information.

Even when we have a full catalog of genes, this will not in itself tell us what any gene does. **Functional genomics** tries to provide this information. Again, the analysis uses both laboratory data and computer analysis of gene sequences. The human **proteome** (the complete set of all proteins) includes maybe 1 million different proteins, many more than the number of protein-coding genes because of a combination of alternative transcription, alternative splicing, alternative post-transcriptional modification and natural amino acid sequence variation – see *Figure 3.10*. Most proteins obtain much of their function through combinations of a limited number – maybe 1000 – of functional modules (*Figure 3.11*). Thus one protein might have a DNA binding module, a protein–protein interaction module and a steroid hormone receptor. Recognizing the modules encoded by a gene can help identify the function of its protein product.

An overview of the human genome

When the focus moves from individual protein-coding genes to the whole genome, a number of puzzling features emerge. According to ENSEMBL, our genome contains about 20 465 protein-coding genes. The average number of exons is about 9, though the actual number varies widely: some genes consist of just a single exon with no introns, while the gene encoding the muscle protein titin has over 300 exons; *Table 3.1* gave a few examples. Exons average around 145 bp in size. Considering all these numbers leads to two surprising conclusions. First, if you add up the sizes of every exon of every protein-coding gene, the total (the **‘exome’**) comes to only about 30 Mb – hardly more than 1% of the DNA in our genome (*Figure 3.12*). This raises the question, what does the other 99% do? Secondly, it appears that we have hardly more protein-coding genes than the 1 mm long nematode worm *Caenorhabditis elegans*, a creature much studied as an example of an extremely simple animal. According to ENSEMBL it takes 20 222 protein-coding genes to build this worm. Surely it is not just anthropocentric arrogance to think that we are far more complex creatures? If we don't have more genes than the worm, maybe we use the same number of genes in a smarter way? Maybe the fact that we have so much non-coding DNA is at least partly because much of it is involved in regulating the expression of genes? We need to take a closer look at the non-coding DNA.

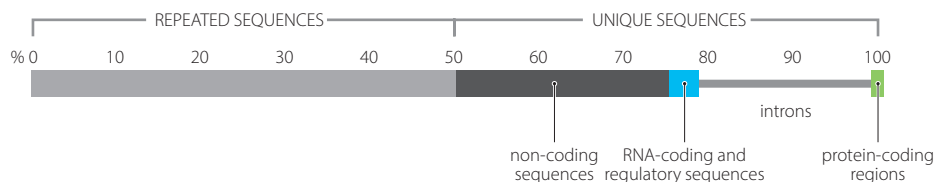


Figure 3.12 – What does all our DNA do?

About 1% does what the Central Dogma says DNA does, encoding protein. A further approximately 5% is conserved between humans and mice, implying it has some sequence-dependent function. Some of the heterochromatin functions as centromeres of chromosomes. About half of the rest of our DNA consists of thousands of copies of transposable elements that have spread within the genome like an infection. Data from the Draft Human Genome Sequence (2001). Diagram courtesy of Dr Jamie Ellingford, Manchester.

Looking at our non-coding DNA

How much of the non-coding DNA is functional, and how much is 'junk DNA' is a hotly disputed question. In addition to the 20 465 protein-coding genes in the Human Reference Sequence, ENSEMBL lists 22 229 genes for non-coding RNAs. These include both long non-coding RNAs, which are often spliced from multi-exon genes, and a whole panoply of short (16–30 nucleotide) RNAs. Some of these have known functions, many do not. One important functional class are the microRNAs, of which 1917 are currently catalogued in a database (www.mirbase.org/summary.shtml?org=hsa) These 21–22 nucleotide RNAs regulate translation by binding to the 3' untranslated region of mRNAs. A single miRNA can affect the expression of dozens, maybe hundreds, of genes.

Around 5% of our genome is conserved between humans and mice, implying some sequence-dependent function. Correct expression of many genes depends on **enhancers**. As mentioned above, these are relatively short sequences that bind transcription factors and other proteins. Enhancers can be located up to a megabase away from the coding sequence they regulate; the DNA loops round to bring the enhancer, with its bound proteins, close to the promoter of the target gene. While sequence conservation is a pointer to function, the converse is not necessarily true. Many enhancers show quite rapid sequence evolution across species. An ambitious international project set out to catalog every functional element in the human genome, using a wide variety of techniques (ENCODE Project Consortium, 2012). ENCODE reported that over 80% of all the non-repetitive DNA in the genome is transcribed, at least in some cells and at some times. They identified 70 292 promoter-like sequences and 399 124 with the chromatin signature of enhancers. How much of the pervasive transcription is functional, and how much is random noise, remains unclear.

How much of the non-conserved DNA is functionless junk? Supporting the 'junk DNA' argument, the onion genome is 5 times the size of the human genome, and that of an amoeba 30 times – surely all that extra DNA is not functional? Moreover, the *Fugu* pufferfish has eliminated most of the intergenic non-coding DNA, with no apparent ill-effects. Almost half of our DNA consists of short sequences present in huge numbers of copies scattered across the genome (see *Figure 3.12*). There are 1.5 million copies of 100–300 bp sequences called SINEs (short interspersed nuclear elements) and 850 000 copies of sequences called LINEs (long interspersed nuclear elements). These comprise about 13% and 21% of our genome, respectively. Complete LINE elements are 6–8 kb long, but the majority are truncated. Other large families of repeats make up a further 11% of our genome. All these repetitive elements are believed to have multiplied within our genome like an infection. They can be seen as a sort of genomic parasite. Originally they had the ability to jump from one chromosomal location to others, hence their name, transposons. Most have lost this ability, but a few retain it. Whether they have any function useful to us, their hosts, is controversial. At least they are for the most part harmless.

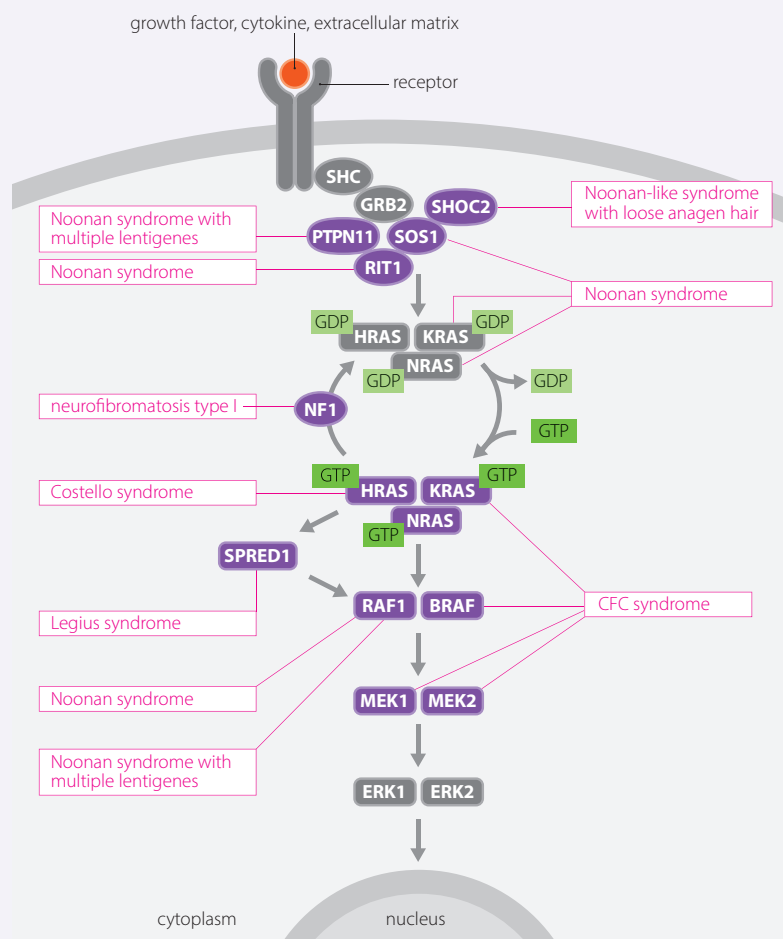
In addition to these high copy-number repeats, there are many low copy-number repeats that have arisen by recent duplication events. We saw an example in *Disease box 2*. There are many traces of sequences, including gene sequences, being duplicated, and then the copies diverging by accumulation of mutations. Often when genes duplicate only one copy remains functional, the other becoming a non-functional **pseudogene**. ENSEMBL lists 15 171 human pseudogenes.

Overall, the human genome appears chaotic. Evidently there has been no selective pressure for a tidy genome. The contrast with the elegance and efficiency of most of our anatomy and physiology is very striking. It is fascinating to speculate on how and why all this has evolved. It is certain that far more of our genome has clear functions, whether encoding functional RNA molecules or regulating other genes, than the 1.2% of protein-coding sequence. Moreover, cells are awash with innumerable RNA species, few of which have known functions. Maybe there is a whole undiscovered world of RNA functions, and we should consider revising our view of cells as essentially protein-driven machines in favor of seeing them as primarily RNA machines. Alternatively, maybe cells are just not very good at distinguishing junk from functional DNA when making transcripts. Much remains to be discovered!

From genes to diseases: the RASopathies

The relation between genes and diseases is complex. There is no one gene – one disease hypothesis to correspond to the one gene – one enzyme hypothesis. Although patients with some diseases do always have mutations in the same gene (e.g. cystic fibrosis, Huntington disease), often the relationship is less clear. A particular disease can be caused by mutations in several different genes, while different mutations in the same gene may cause different diseases. Much of the confusion can be resolved by thinking in terms of how genes function in pathways. The diseases collectively known as RASopathies illustrate this.

Initially these presented as a confusing series of syndromes with overlapping features and no simple correlation between genes and diseases. The fog cleared when it was realized that all the various mutations were causing either over-activation or under-activation of a single cell signaling system, the RAS–MAPK pathway. This pathway relays signals from cell surface receptors to the nucleus, where they control transcription of numerous genes involved in cell proliferation and differentiation. Transcription of the target genes is ultimately controlled by two transcription factors, ERK1 and ERK2, members of the MAPK (mitogen-activated protein kinase) family. However, the link between the cell surface receptors and ERK1/2 is not direct, but involves a series of proteins (see Box figure 3.4). There are also proteins that antagonize



Box figure 3.4 – The RAS–MAPK pathway.

the action of others in the cascade. This complexity, so typical of cell biology, allows multiple points of control, including interactions with other signaling pathways.

Receptors for several different signals trigger conversion of small intracellular proteins of the RAS family (HRAS, KRAS, NRAS) from inactive GDP-bound forms to the active GTP-bound forms. Accessory proteins assist the conversion (guanine exchange factors such as SOS1) or reverse it (GTPase activating proteins such as neurofibromin, the product of the *NF1* gene). RAS–GTP triggers phosphorylation and activation of BRAF. BRAF then phosphorylates MEK1 and MEK2, activating them to phosphorylate and activate ERK1 and ERK2. The role of protein phosphorylation and specifically tyrosine kinases is explored in *Chapter 7* (see *Figure 7.4*).

Mutations that reduce the ability of a protein to transmit the signal (loss of function mutations) make cells under-responsive to external signals. However, similar mutations affecting the inhibitory proteins make the system over-react. A similar effect is seen when proteins in the main pathway suffer activating (gain of function) mutations, which make them activate the pathway even in the absence of an external signal. Excessive RAS–MAPK signaling causes hyperproliferation. Strong gain of function mutations, especially of the three RAS genes (*HRAS*, *KRAS* and *NRAS*) and of *BRAF* are very common in tumors (see *Chapter 7*) and in mosaic form are the cause of some non-heritable syndromes (see *Disease box 6*). Presumably such mutations would be incompatible with life in constitutional form. Less strongly activating mutations can be inherited and cause RASopathy diseases. Several of the syndromes can be caused by mutations in any one of several genes, while different mutations in the same gene can cause different syndromes.

Thinking in terms of pathways, rather than individual genes, has enabled geneticists to make sense of a previously confusing set of overlapping syndromes.

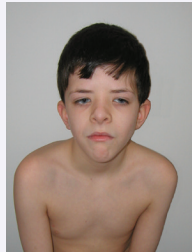
Syndrome	OMIM number	Genes involved	Molecular pathology and features
CFC (cardiofacio-cutaneous syndrome)	115150	<i>KRAS</i> <i>BRAF</i> <i>MEK1</i> <i>MEK2</i>	Mild gain of function mutations, resulting in increased RAS–MAPK signaling. Coarse facial features, heart defects, developmental disability.
Costello	218040* (actually a dominant condition)	<i>HRAS</i>	Gain of function mutations resulting in increased RAS–MAPK signaling. Coarse facial features, severe feeding problems, failure to thrive, short stature, heart defects, papillomata. Increased risk of tumors.
Noonan	163950	<i>PTPN11</i> <i>KRAS</i> <i>NRAS</i> <i>SOS1</i> <i>RAF1</i> <i>SHOC2</i>	Mild gain of function mutations, resulting in increased RAS–MAPK signaling. Short stature, pectus excavatum, heart defects, downslanting palpebral fissures and ptosis. Mild learning disability.
NF1 (neurofibromatosis Type 1)	162200	<i>NF1</i>	Loss of function mutations in an inhibitor of RAS action, resulting in increased RAS–MAPK signaling. Café-au-lait patches, neurofibromas of skin and nerves, Lisch nodules in the eye, macrocephaly – see <i>Disease box 1</i> . May have similar facies to Noonan syndrome.
Legius	611431	<i>SPRED1</i>	Loss of function mutations in a negative regulator of RAS–RAF interaction, resulting in increased RAS–MAPK signaling. Café-au-lait patches without the other complications of NF1.

Syndrome	OMIM number	Genes involved	Molecular pathology and features
Noonan syndrome with multiple lentigenes	151100	<i>PTPN11</i> <i>BRAF</i>	Loss of function mutations, resulting in reduced RAS–MAPK signaling. Lentigenes, EKG conduction abnormalities, ocular hypertelorism, pulmonary stenosis, abnormal genitalia, retardation of growth and deafness.

* OMIM numbers beginning with 2, such as 218040, are used for recessive conditions. When Costello syndrome was given this number it was mistakenly thought to be recessive.



(a)



(b)



(c)



(d)



(e)

Box figure 3.5 – RAS–MAPK syndromes.

(a) NFI, (b) Noonan, (c) Noonan with multiple lentigenes, (d) CFC, and (e) Costello.

3.5. References

- Cramer P** (2019) Organization and regulation of gene transcription. *Nature*, **573**: 45–54. *A very detailed review.*
- ENCODE Project Consortium** (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**: 57–74.
- International Human Genome Sequencing Consortium** (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**: 860–921.
- International Human Genome Sequencing Consortium** (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**: 931–945.
- Snead MP and Yates JRW** (1999) Clinical and molecular genetics of Stickler syndrome. *J. Med. Genet.* **36**: 353–359.
- Strachan T and Read AP** (2019) *Human Molecular Genetics*, 5th edn. CRC Press. *For fuller discussion of all the topics covered in this chapter.*
- Youle RJ** (2019) Mitochondria—striking a balance between host and endosymbiont. *Science* **365**: eaaw9855. *A detailed look at how the nuclear and mitochondrial genomes co-operate.*

Useful websites

ENSEMBL genome browser – www.ensembl.org/Homo_sapiens/Info/Index

ENCODE project explorer – www.encodeproject.org (gives access to ENCODE data spread across many different publications).

MicroRNA database – <http://mirbase.org/>

PFAM protein structure database – <http://pfam.xfam.org>

SANTA CRUZ genome browser – <http://genome.ucsc.edu>

For general background, many different free online resources cover the general material of this chapter. The Eurogems website (www.eurogems.org) links to a range of useful resources.

Cold Spring Harbor Laboratory DNA Learning Center <https://dnlc.cshl.edu/>

OpenStax College *Biology* <http://cnx.org/content/col11448/latest/>

University of Utah Health Sciences Genetic Science Learning Center
<http://learn.genetics.utah.edu/>

3.6. Self-assessment questions

- (1) Consider the DNA sequence

CCAGCTTCGCAAGTC

Which base is immediately downstream of a CG dinucleotide, (a) in the strand shown, and (b) in the complementary strand?

- (2) A partial gene sequence from the database reads

CAGCTGGAGGAACTGGAGCGTGCTTTTGAG

Write out the sequence of the template strand and the mRNA.

- (3) The sequence of a 150-nucleotide section of chromosome 7 is shown. It is written in groups of 10 to make counting easier. The sequence in upper case is exon 1 of a gene. Flanking nucleotides shown in lower case are not part of the exon. The initiation codon is double-underlined.

1 gcagccaatg gaggggtggtg ttgcgcggggg ctgggattag ggccggggcg
a

51 aaatgGGATC CTCCAAGGCG ACCATGGCCT TGCTGGGTAA GCGCTGTGAC
b c

101 GTCCCCACCA ACGGCgttag acctcagtag tgaatcagga cctcactcct
d e

- (a) What number nucleotide in the sequence is the first one to be transcribed into mRNA?
- (b) Name the parts of the sequence underlined and labeled a, b, c, d, e, choosing from the following list:
- 3' untranslated region
 - 3rd codon
 - 5' untranslated region
 - 9th codon
 - acceptor splice site (the 3' end of an intron)
 - donor splice site (the 5' end of an intron)

part of exon 2
 part of intron 1
 part of intron 2
 part of promoter

- (4) Here are two nucleic acid sequences. Sequence (A) is a genomic sequence and (B) is the corresponding mRNA.

(A)

```
ATGACCACGCTGGCCGGCGCTGTGCCCAGGATGATGCGGCCGGGCCCGGGGCAGAACTACCCGCGTAGC
GGGTTCCCGCTGGAAGGTAAGGGAGGGCCTCAGCGCGCCGCGCTTCTCTTTTTTACCTTCCCACAGTGT
CCACTCCCCTCGGCCAGGGCCGCGTCAACCAGCTCGGCGGTGTTTTTATCAACGGCAGGTACCAGGAGA
CTGGCTCCATACGTCCTGGTGCCATCGGCGGCAGCAAGCCCAAGGTGAGCGGGCGGGCCTTGCCCTCCT
CGCCTGCCCGCCTGTTCTCTTAAAGCAGGTGACAACGCCTGACGTGGAGAAGAAAATTGAGGAATACAA
AAGAGAGAACCCGGGCGTGCCGTCAGGTACTAGGCCCATTAACCTCTCCCCGCTTCCTTCCTCCTCCCG
CCCCCAGTGAGTTCCATCAGCCGCATCCTGAGAAGTAAATTCGGGAAAGGTGAAGAGGAGGAGGCCGTC
CTGAGCGAGCGAGGTAAGCGGTGGCGCCTTGGGCGGCGGTTGAAGTAGCTTTTTATGCCCTCAGGAAAGG
CCCTGGTCTCCGGAGTTTCTCGCATTAAGGAGAGAGAGAGAGAGAGTACTCTTTTGACTGGT
```

(B)

```
AUGACCACGCUGGCCGGCGCUGUGCCCAGGAUGAUGCGGCCGGGCCCGGGGCAGAACUACCCGCGUAGC
GGGUUCCCGCUGGAAGUGUCCACUCCCCUCGGCCAGGGCCGCGUCAACCAGCUCGGCGGUGUUUUUAUC
AACGGCAGGUACCAGGAGACUGGCUCCAUACGUCCUGGUGCCAUCGGCGGCAGCAAGCCCAAGCAGGUG
ACAACGCCUGACGUGGAGAAGAAAAUUGAGGAAUACAAAAGAGAGAACCCGGGCGUGCCGUCAGUGAGU
UCCAUCAGCCGCAUCCUGAGAAGUAAAUUCGGGAAAGGUGAAGAGGAGGAGGCCGUCCUGAGCGAGCGA
GGAAAGGCCCUUGGUCUCCGGAGUUUCCUCGCAUUAAGGAGAGAGAGAGAGAGAGUACUCUUUUGACUGGU
```

How many exons does this gene have? Make a little table, showing the nucleotide number (in the genomic sequence) of the start and end of each exon, and the length of each exon and intron. Note that in real genes the introns are likely to be much longer than here.

- (5) Look up the following genes using a genome browser such as the ENSEMBL or SANTA CRUZ browser: *BRCA1*, *GJB2*, *DMD*.
- How many transcripts are recorded for each?
 - Choosing one transcript for each gene, note its ID number and report how many exons it has.
 - What is the relationship between the different transcripts?
 - What is the size of the gene, from the 5' end of exon 1 to the 3' end of the last exon? Note that nucleotides are counted from the tip of the short arm of the chromosome. Because the two strands of the double helix are antiparallel, genes on one strand are transcribed from the short arm tip towards the long arm tip, while those on the other strand are transcribed in the opposite direction. Therefore for genes on one strand the numbering of the nucleotides goes up as you move 5'→3' through the sense strand, for those on the other strand the numbering goes down.

[Hints on questions 2, 3 and 4 are provided in the Guidance section at the back of the book.]

04

How can a patient's DNA be studied?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe the main applications of nucleic acid hybridization
- Describe a microarray and how it might be used
- Draw a diagram showing the principles of the polymerase chain reaction
- Describe appropriate applications of Southern blotting, PCR, FISH, array-CGH, SNP chips and MLPA
- Describe what information can and cannot be obtained by analyzing DNA with these techniques, and what additional information can be obtained through studying RNA or proteins

4.1. Case studies

CASE 11 LIPTON FAMILY

- Baby boy, Luke, with developmental delay
- Family history of learning difficulties
- Unusual features of Fragile-X pedigrees

83 105 395

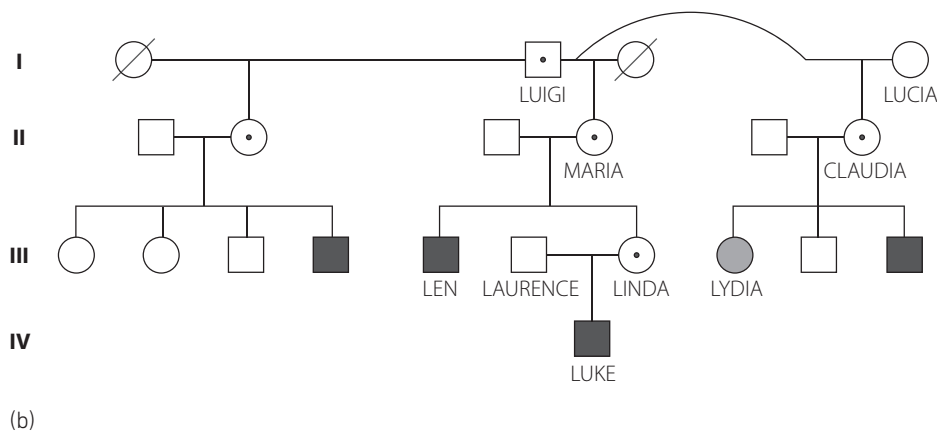
Linda and Laurence Lipton were delighted when their son Luke was born and appeared healthy. He weighed 3.6 kg, fed well and seemed a 'good' baby who didn't cry much and lay still when his diaper was changed. Linda had a friend with a baby of the same age and as time went on Linda began to get concerned about Luke's head control and how slow he was compared to her friend's baby. Linda's anxieties were made worse by the fact that she had an older brother, Len, who had learning difficulties and lived in a group home. She hadn't worried before about risks to her children since she had been told that her brother had been 'starved of oxygen' during birth. Linda told the doctor at the baby clinic about her worries. However, an examination didn't reveal any problems other than Luke was a little hypotonic – in fact the doctor said he had a very good head size for his age. The doctor also asked about family history, particularly in the female line.

Linda had always known that she had two male half-cousins who had learning difficulties. As a biology teacher, she knew about X-linked inheritance and, having drawn up a pedigree (*Figure 4.1b*), she suddenly realized that the pattern in the family might indicate an X-linked condition. But on second thoughts she dismissed the idea because her mother and her two aunts were the products of three separate marriages of her maternal grandfather Luigi. X-linked inheritance would either require Luigi's three unrelated wives all coincidentally to be carriers of the condition, or else would require Luigi himself to be the source of the mutant allele in all three of his daughters. Three independent mutations

would be too much of a coincidence, while Luigi himself was clearly unaffected, having founded a very successful business. Also, there was the question of her cousin Lydia. Lydia was strikingly slow, had left school without any qualifications and worked as an unskilled care assistant. If there were an X-linked condition segregating in the family, Lydia could be a heterozygote – but all the other women in the family who on this hypothesis would also be heterozygotes were graduates following professional careers. Linda asked to be referred to the genetic clinic because she wanted to go through all her reasoning and to see whether an expert agreed the pedigree was incompatible with X-linked inheritance. The doctor in the genetic clinic, however, thought the pedigree and phenotypes were very suggestive of Fragile X syndrome. She explained why she thought this to Linda, gave her some literature about the syndrome and arranged to take blood for DNA extraction.



(a)

**Figure 4.1 – Fragile X syndrome.**

(a) Typical facial appearance of an affected boy. (b) Pedigree of the Lipton family. Affected males are shown in black; III-8 (gray) is a mildly affected female. Obligate carriers are indicated. Note that the normal male I-2 is the source of the fragile X chromosome in his daughters. He is an example of a normal transmitting male (see Section 4.3).

CASE 12 MEINHARDT FAMILY

- Madelena, baby daughter of Margareta and Manfred
- Multiple congenital abnormalities and developmental delay
- Normal 46,XX karyotype under the microscope

84

101

395

Madelena was the second child born to Manfred and Margareta. Their older son was healthy and developing normally. Madelena was born at term weighing 2.7 kg but there were problems from the beginning even though scans in pregnancy had been reassuring. Madelena was hypotonic; she wouldn't suck from a bottle and needed to be tube fed. She was also noted to have a relatively small head, a ridged metopic suture and large ears. In view of these problems the pediatrician arranged for a routine chromosome test which showed a normal 46,XX karyotype. Although her mother managed to get her onto bottle feeds and the family was able to go home, this progress did not last long and Madelena was readmitted to hospital 2 weeks later with constant crying. In spite of tube feeding being re-started she did not settle and so the specialist undertook another detailed examination. He was worried to find that Madelena's corneas appeared cloudy and asked for an urgent ophthalmic check. This confirmed that she had severe congenital glaucoma known as buphthalmos. Surgery was performed immediately to

save her sight. Madelena's further development was slow; she didn't smile until she was 5 months old, she sat at 16 months and the decision was made to feed her directly into her stomach through a gastrostomy tube because of recurrent chest infections due to inhalation of milk. Manfred and Margareta didn't think matters could get much worse but the next problem they faced was the onset of seizures in Madelena at 2 years of age, leading to more hospital admissions and tests. The family was desperate to know what might be the underlying cause of all their daughter's problems, and they were referred to a geneticist who strongly suspected an underlying chromosome problem, despite no abnormality having been apparent under the microscope. However, it was several years before a test became available that could scan the whole genome for submicroscopic copy number changes (duplications or deletions).



Figure 4.2 – A child with multiple congenital abnormalities suggestive of a chromosomal abnormality.

Note the small head, prominent metopic region (above bridge of nose) and large ear. Note also buphthalmos (large eyes) due to congenital glaucoma.

4.2. Science toolkit

A diploid human cell contains 6×10^9 bp of DNA, all chemically identical, all consisting of A, G, C and T nucleotides. Progressing the analysis of most of our cases will require us to examine specific genes or chromosomal regions in DNA samples provided by the patients. How can we achieve this, against a background of so much chemically identical but irrelevant DNA?

With the plummeting costs and soaring capacities of 'next-generation' DNA sequencers, the answer is increasingly 'sequence it all and think afterwards which bits you want to look at'. Next-generation sequencing is transforming much of clinical genetic diagnosis from a hypothesis-driven to a data-driven process, prompting fears among some clinical geneticists that their hard-won diagnostic skills might become redundant. Nevertheless, sequencing is not the answer to every diagnostic question. There are many situations where a more targeted approach makes more sense than simply sequencing a patient's whole exome or genome. Often it is clear which gene needs to be studied – in this chapter, for example, the cases of **Alfred Ashton (Case 1)**; Huntington disease) or **Martin Davies (Case 4)**; Duchenne muscular dystrophy). Larger abnormalities may be found more easily using microarrays (although as the cost of sequencing continues to fall, there has been some gradual movement away from microarray technology). In this chapter we consider some alternatives to sequencing for identifying relatively large DNA changes; then in *Chapter 5* we will look at ways of studying nucleotide-level variation, especially by sequencing.

All the many laboratory methods for examining specified parts of the overall genome boil down to one of two essential approaches:

- (a) hybridize the test DNA (or RNA) to a labeled matching DNA or RNA molecule (a **probe**).
- (b) amplify the sequence of interest. Use the polymerase chain reaction (PCR) to make so many copies of just that sequence that the test sample can be treated as a slightly impure specimen of just the sequence of interest.

Nucleic acid hybridization

The two strands of a double helical DNA molecule are held together by relatively weak chemical bonds (hydrogen bonds) between paired bases. The bonds can be broken by boiling the solution containing the DNA, or by exposing it to a high pH. This is called **denaturing** the DNA. The process is reversible. Complementary single-stranded nucleic acids will **hybridize** or **anneal** (stick together to form a double helix) if they are mixed in solution at a moderate temperature (typically below 50–60°C) and pH (*Figure 4.3*). The ability of matching strands to hybridize is the basis of much of DNA technology.

Hybridization does not necessarily require a perfect match between the strands. Two strands will hybridize if there are enough correctly matching base pairs that can form hydrogen bonds, even if some bases are mismatched. Strands with mismatches will denature at a lower temperature than perfectly matched strands, and equally will require a lower temperature to hybridize successfully. Short strands will denature more easily than long strands because there are fewer hydrogen bonds holding them together. They need a lower temperature for hybridization, and hybridization is more sensitive to mismatches. By manipulating these variables, hybridization can be used in various different ways for DNA testing.

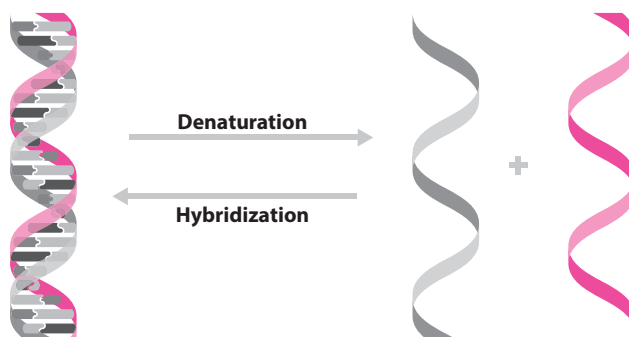


Figure 4.3 – Denaturation of double-stranded DNA; hybridization of two complementary single strands.

Using hybridization as the basis for DNA testing

Applications of hybridization in clinical genetics fall into two general types:

- Sometimes the aim is to allow only perfectly matched sequences to hybridize. For this purpose short **oligonucleotides** (typically 15–30 nucleotides long) are used as the hybridization probes. For such short probes the hybridization

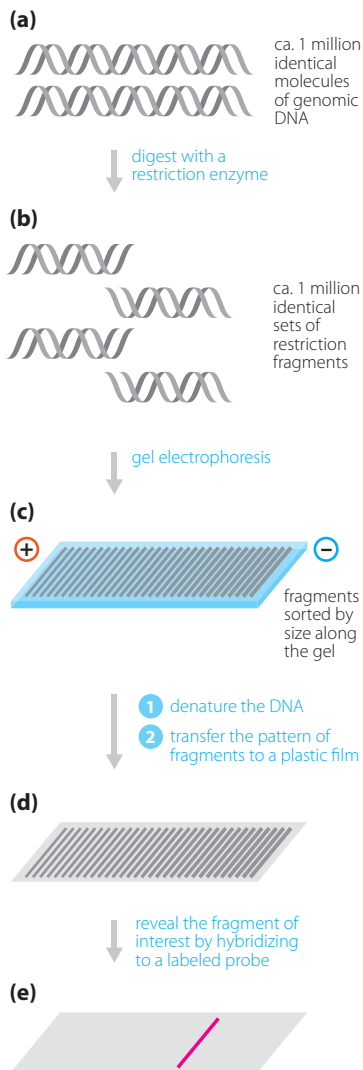


Figure 4.4 – Southern blot analysis of DNA.

The test DNA (a) is digested with a restriction endonuclease (b). The fragments are separated by size by gel electrophoresis (c), denatured by immersing the gel in alkali, and then transferred to a nitrocellulose film (d). The nitrocellulose sheet is then immersed in a solution containing the labeled probe, revealing any fragments that hybridize (e). See *Box 4.1* for more details. This shows whether or not a sequence matching the probe is present and, if so, on what size restriction fragment it is located.

conditions can be made so critical that even a single nucleotide mismatch will be enough to prevent hybridization. Thus they can discriminate between alleles at a locus that differ by just a single nucleotide (**allele-specific oligonucleotides** or **ASOs**). Highly specific hybridization is essential for the microarray-based SNP chip method described below. It is also a prerequisite for PCR (see below). Although PCR is not primarily a test for hybridization, it does depend on specific hybridization for its ability to amplify just one specific sequence. The use of allele-specific oligonucleotides in mutation testing is described in *Chapter 5*.

- At other times we want a hybridization test that will work on anybody's DNA, regardless of any minor differences between the DNA of different people. When we use a hybridization test to check for a chromosomal microdeletion in **George Green (Case 7)** we don't want hybridization to fail just because the corresponding sequence in George may have a few variant nucleotides. Lack of hybridization should mean that the whole sequence is missing. Similarly, comparative genomic hybridization (see below) needs probes that will hybridize to the appropriate sequence in the DNA of any person. For these tests we use a much longer hybridization probe, either chemically synthesized or a cloned piece of natural DNA. A probe of 100 bp or more will hybridize to its cognate sequence even if this varies a bit between individuals.

Whichever procedure we use for a hybridization test, we need to be able to tell whether or not the probe has hybridized to the test DNA. Most commonly this is achieved by labeling one of the two hybridizing partners, then separating out the other partner and seeing whether or not the label has accompanied it. In former times, DNA was usually labeled by incorporation of radioactive ^{32}P ; nowadays it would be tagged with a fluorescent dye or in a way that exploits a strong affinity for some detection system. Examples would include biotin, which has a powerful affinity for streptavidin, or molecules that are bound by specific antibodies. Usually the separation is achieved by fixing one partner to a solid support – a glass slide, a magnetic bead or a piece of nitrocellulose film, for example. This is immersed in a solution containing the labeled partner, left to hybridize, then taken out and washed. The label will then mark any areas of the solid support where hybridization has taken place.

In this section we describe five applications of DNA hybridization in diagnostics:

- Southern blotting
- Fluorescence *in situ* hybridization (FISH)
- Multiplex ligation-dependent probe amplification (MLPA)
- Array-comparative genomic hybridization (aCGH)
- Microarrays to detect single nucleotide polymorphisms (SNP chips).

Southern blotting

In the 1980s Southern blotting (summarized in *Box 4.1* and *Figure 4.4*) was the mainstay of molecular diagnosis. To the relief of most laboratory workers, this demanding technique has been almost entirely superseded by quicker and easier alternatives. It does, however, still have a few applications. Southern blots can characterize sequence changes that are difficult to amplify by PCR (see below). G–C base pairs are held together by three hydrogen bonds, compared to only two for A–T base pairs (see *Figure 11.14*), and so GC-rich DNA does not denature so readily. As a result,

long highly GC-rich sequences are difficult or impossible to amplify by PCR. Here Southern blotting is used in **Case 11 (Lipton family)** to check the exact size of a large expansion of a $(\text{CGG})_n$ trinucleotide repeat in the 5' untranslated region of the *FMR1* gene, as described below.

Principle of Southern blotting

BOX 4.1

The test DNA (isolated from a large number of cells and so consisting of many identical genomes) is first cut into specific fragments using a special enzyme (a restriction endonuclease – see Box 4.2 and Figure 4.4b). These enzymes cut the DNA in a reproducible way, so that DNA from every cell in the test sample gives the same set of fragments. The mixture of fragments is subjected to gel electrophoresis (Box 4.3) so that fragments of different sizes are located at different positions in the gel (Figure 4.4c). The fragments, still in their size-dependent pattern, are denatured in alkali to make them single-stranded. We want to see which fragments will hybridize to our probe, but the hybridization reaction will not work efficiently on DNA immobilized in a gel. The probe molecules cannot move about freely to find their matching partners. The next step is therefore to transfer the fragments onto nitrocellulose paper, while retaining their spread-out pattern. This is the actual Southern blotting procedure. It is usually done by laying the sheet of nitrocellulose paper on top of the gel, putting a stack of dry paper towels on top of that, and very carefully pressing so that the liquid in the gel is sucked through the nitrocellulose paper into the towels. The denatured DNA sticks to the nitrocellulose (Figure 4.4d). This is immersed in a solution containing a labeled probe that will stick to just the fragment of interest and so reveal its position (Figure 4.4e). Depending on the label used, the location of the probe is detected by autoradiography, by fluorescence or by immunological detection of the label. The technique provides information on whether or not a sequence matching the probe is present and, if so, on what size restriction fragment it is located.

Restriction endonucleases

BOX 4.2

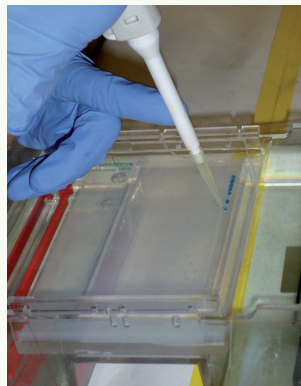
These bacterial enzymes cut double-stranded DNA whenever they encounter some specific short sequence, usually 4, 6 or 8 bp long. Bacteria use them as part of a defense mechanism against foreign DNA. Many different restriction enzymes have been isolated from a wide range of bacterial species. Each has its own recognition sequence – for example, the enzyme *EcoRI* cuts GAATTC.

For molecular biologists, restriction enzymes are very valuable tools that provide a means of cutting a long DNA molecule into reproducible fragments. It is easy to forget that virtually everything one does with DNA involves a large collection of identical molecules. For example, a Southern blot would typically use 5 μg of DNA extracted from blood. One human cell contains 6.4 pg of DNA, so 6.4 μg of DNA would be the DNA content of 1 million cells. Thus the starting material is a very large collection of identical molecules. Unless the restriction enzyme cuts each molecule in exactly the same places, the fragments would be a complete jumble and would not run as sharp bands on a gel. The average size of fragment depends on the choice of restriction enzyme. On average a 4-bp restriction site will occur by chance once every $4^4 = 256$ bp, while a 6-bp site will occur on average every 4096 bp. Restriction sites are randomly distributed in human DNA so there will be a wide spread of fragment sizes around these averages.

Gel electrophoresis

Along with PCR, gel electrophoresis is a central tool of molecular genetics. DNA molecules carry a negative charge because of the phosphate groups (see *Box 3.2*), and so in an electric field a DNA molecule will move towards the positive pole. If the DNA is contained in an agarose or polyacrylamide gel, the DNA molecules have to fight their way through a jungle of long polymer molecules. How fast a double-stranded DNA molecule moves depends almost entirely on its size, and hardly at all on its sequence. Small molecules

move fast, large ones move slowly. Molecules of a given size will form a sharp band. The position of a fragment, revealed by hybridization or staining, is a measure of its size. Electrophoresis can be done manually through a slab gel, as illustrated in *Box figure 4.1a*, but for diagnostic purposes it is usual to use an automated gene analyzer *Box figure 4.1b*. The software displays the result as a trace with peaks (an electropherogram, see *Figure 4.14*). The position of a peak determines the size of a fragment, and the area under the peak measures the quantity.



(a)



(b)

Box figure 4.1 – Gel electrophoresis.

(a) Loading an electrophoretic gel by hand. The DNA samples, mixed with blue dye to make them visible, are pipetted into wells formed in the slab of agarose gel. The gel is submerged in a buffer solution to cool it and make electrical contact. **(b) An automated gene analyzer.** The machine electrophoreses 96 fluorescently labeled DNA samples in parallel through very thin capillaries. For each sample a laser records the time it takes for each fragment to emerge from the capillary, and the intensity of fluorescence. Image of ABI 3730 genetic analyzer courtesy of Thermo Fisher Scientific.

Fluorescence in situ hybridization (FISH)

This is used to check the presence or absence, copy number and chromosomal location of a particular relatively large (100 kb) DNA sequence in chromosomes of a patient. For **George Green (Case 7)**; 22q11 deletion) we want to check for a megabase-sized deletion at a specific chromosomal location. If a deletion is present it will affect only one of his two homologous chromosomes. To retain the information about the chromosomal location the hybridization target is a spread of his chromosomes on a microscope slide, rather than extracted DNA. By very careful treatment the DNA in the spread-out chromosomes can be denatured without destroying the recognizable form of the chromosomes. Although each sequence has its partner nearby, because they are stuck down on the glass slide, the partners cannot move to find each other, so the chromosomal DNA remains single stranded. The slide is exposed to a solution containing the labeled probe. A cloned piece of the relevant DNA, kilobases long, is used as the probe so that, as mentioned above, minor sequence variations do not affect the hybridization. The probe is labeled with a fluorescent dye. After hybridization the slide is washed and examined under the microscope. Where the probe has hybridized, a pair of

fluorescent spots is seen over the relevant chromosome (a pair of spots because at this stage in the cell cycle each chromosome consists of two sister chromatids, see *Chapter 2*). *Figure 4.5* shows the principle, *Figure 4.10* an actual example. Unlike methods that use pooled DNA from many cells, FISH gives a view of individual cells, allowing easy analysis of mosaicism. It is also the only one of the techniques described here that can reveal the chromosomal location of its target.

It is also possible to use FISH on interphase (non-dividing) cells. Counting the number of fluorescent spots would reveal numerical chromosome abnormalities, while the close juxtaposition of labels for different chromosomes could indicate a structural abnormality. For example, cancer cells can be checked for specific chromosomal translocations which may have prognostic value by using two different colored FISH probes for the two chromosomes involved. If the specific translocation is present the different colored spots would always lie close together within the cell nucleus (see *Figure 7.10*).



Figure 4.5 – Principle of fluorescence *in situ* hybridization (FISH).

Chromosomes on a microscope slide are very carefully denatured and then hybridized to a fluorescently labeled probe. See *Figure 4.10* for an example.

MLPA (multiplex ligation-dependent probe amplification)

MLPA checks for specific DNA deletions or duplications using a combination of two processes, hybridization and DNA ligation. Compared to PCR, it produces a reliable quantitative result, and so can detect duplications and heterozygous deletions. Each individual probe consists of two separate oligonucleotides that hybridize to adjacent sequences that are hopefully invariant in the DNA to be tested. The enzyme DNA ligase will covalently seal the gap between them, but only if the nucleotides flanking the gap are correctly base-paired. When the ligase joins the oligonucleotides it creates a single long molecule that can then be PCR amplified. MLPA is a multiplex process, testing up to 45 sequences simultaneously. Each oligonucleotide has a standard sequence at its distal end so that all the products of ligation can be PCR amplified using a single pair of primers. One of the primers is fluorescently labeled, and the reaction mix is analyzed on a fluorescence gene analyzer. *Figure 4.6* shows the principle, *Figure 4.11* shows the result with **Case 4 (Martin Davies)** and his family.

MLPA has gained acceptance in genetic diagnostic laboratories due to its simplicity compared to other methods, relatively low cost, capacity for reasonably high throughput

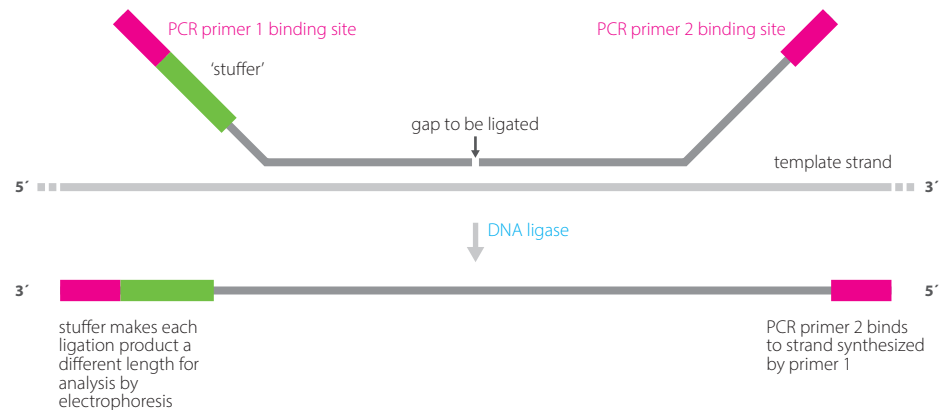


Figure 4.6 –MLPA (multiplex ligation-dependent probe amplification).

This is a method for detecting small deletions or duplications, typically of a single exon. Each probe consists of a pair of oligonucleotides that can be ligated only if their adjacent ends perfectly match the test sample. Ligation creates a PCR-amplifiable molecule. In multiplex MLPA the mixed ligation products are amplified using a single pair of primers that hybridize to sequences present on the ends of every probe pair. A variable length 'stuffer' sequence ensures that each ligation product is a different length, giving a unique peak on a sequencer trace. See *Figure 4.11* for an example.

and perceived robustness. Establishing a new MLPA assay is time-consuming because each individual probe needs to be carefully optimized, so it is mostly used for frequently investigated genes, where probe sets are commercially available.

Array-comparative genomic hybridization (array-CGH)

Both FISH and MLPA are targeted tests: you need to know in advance where in the genome to look. Array-CGH and SNP chips (described below) are global techniques, able to detect any sequence that is present in a greater or smaller number of copies per genome in the test DNA compared to a normal control DNA. They do not supplant MLPA because, as normally implemented in diagnostic labs, they do not have the resolution to pick up the single-exon variants detected by MLPA. Also, compared to FISH, they cannot tell whether any extra copies of a sequence are at their normal chromosomal location or elsewhere in the genome.

In array-CGH it is the probe rather than the test DNA that is anchored to a solid support, and the test DNA rather than the probe that is labeled with a fluorescent dye. The probes are in the form of a **microarray**. A small glass slide is divided into many cells, like the pixels of a digital image. Each cell contains thousands of molecules of one particular probe anchored to the slide. There may be up to a million cells on the slide. To perform the test, the slide is immersed in a solution containing the labeled test DNA.

Array-CGH uses competitive hybridization. DNA from the patient and DNA from a normal control are randomly fragmented and the fragments labeled with two different fluorescent dyes, for example, the patient green and the control red. The two labeled DNAs are mixed in equal amounts and allowed to hybridize to the microarray (*Figure 4.7*). In each cell of the array green DNA fragments from the patient compete with red fragments from the control to hybridize to the molecules of the probe. After hybridization

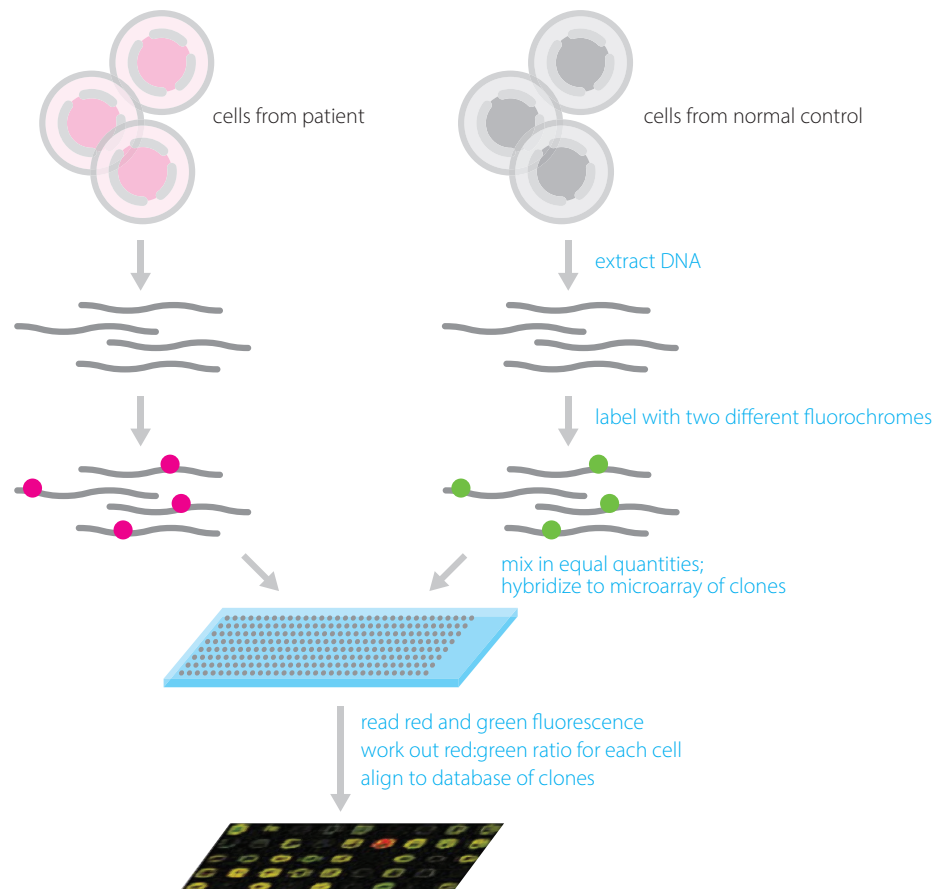


Figure 4.7 – Principle of CGH using a microarray of genomic clones (array-CGH);

adapted with permission from an original by Dr Joris Veltman, Nijmegen. See *Figure 4.12* for an example.

and washing, the slide is examined under a fluorescence microscope to determine the relative amounts of red and green fluorescence bound to each cell. Where the two are present in equal amounts the cell fluoresces yellow. Sequences where one or both copies are deleted in the patient give cells with more red fluorescence, because for that sequence there is relatively more control DNA. Duplications give green cells.

Array-CGH has many powerful features. The availability, through the Human Genome Project, of huge numbers of precisely mapped and characterized clones allows total control over the specificity and resolution of a CGH array. Arrays can be constructed that are specific for one particular chromosome, or that cover the entire genome. The resolution depends on the number of probes used and their distribution across the genome. Because each probe is precisely mapped on the human genome sequence, any deletion or duplication detected can be immediately linked to its chromosomal location and to a list of the genes involved. As mentioned in *Chapter 2*, it was array-CGH that first revealed the existence of hundreds of harmless copy number variants in the DNA of normal healthy individuals (*Figure 2.21*). Interpretation of CGH data requires filtering out these normal variants to concentrate on any that are potentially pathogenic, which can introduce uncertainties. Here we use it to identify and define the chromosome

abnormality in **Case 5 (Elizabeth Elliot)** (Figure 4.12). The main limitation of array-CGH is that it cannot detect balanced chromosomal rearrangements such as translocations or inversions, only copy number changes. Its resolution is also limited: even the highest resolution arrays cannot detect variants less than a few kilobases in size.

SNP chips

An alternative application of microarray technology to detect copy number changes uses so-called SNP chips. The probes on these microarrays target stretches of DNA that contain common single nucleotide polymorphisms or SNPs. Around 1 nucleotide in 300 across the human genome is polymorphic – that is, at a certain position in the genome, two (or occasionally more) alternative nucleotides are each quite frequent in the population. The great majority of SNPs are in non-coding DNA and have no phenotypic effect. For each SNP the chip has two cells containing short allele-specific oligonucleotide probes specific for each of the two variants. SNP chips are versatile tools. They are used in linkage analysis as described in *Chapter 8*, and in the search for genetic susceptibility factors for non-mendelian diseases as described in *Chapter 13*. They offer an alternative to array-CGH for identifying copy number variants, as the case of **Madelena Meinhardt (Case 12)** will illustrate (Figure 4.13).

A typical SNP chip would carry pairs of allele-specific oligonucleotides specific for maybe 500 000 SNPs spaced across the genome. Unlike array-CGH, the test does not use competitive hybridization. Only the DNA of the patient is used. Both the hybridization intensity and the SNP genotypes (homozygous or heterozygous) can provide useful information. If there is a microdeletion, this would show as a series of SNPs from contiguous chromosomal locations where the DNA only hybridized to the probe for one allele, but with a hybridization intensity only half what one would see with DNA from somebody who was homozygous for that allele. A duplication would show as an increased intensity of hybridization, compared to the probes mapping either side of the duplicated region. Comparing genotypes and intensities in the patient and his parents can shed light on the origin and nature of any variants detected.

Amplifying a sequence of interest: the polymerase chain reaction

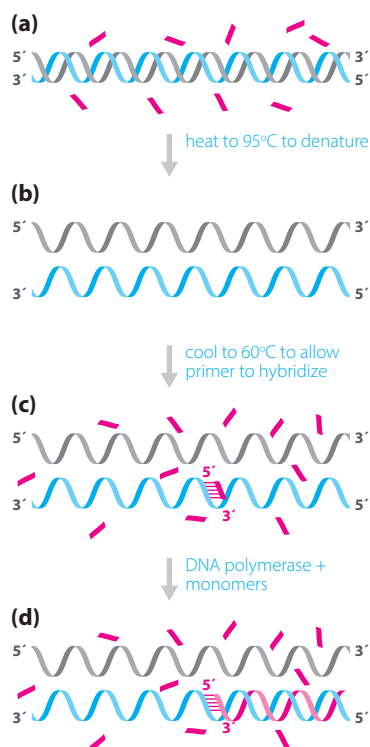
When the sequence of interest forms only a very small fraction of a DNA sample, it may be made visible and followed by means of specific hybridization, as described above. Alternatively, it can be studied after selective amplification. Traditionally this was done by cloning random fragments of the DNA of interest in living cells, usually in *E. coli* bacteria. This enabled the researcher to obtain many copies of the sequence in pure form, uncontaminated by all the other DNA originally present. For diagnostic purposes cell-based cloning has been entirely superseded by PCR. This is a form of *in vitro* cloning. Like cell-based cloning, it works by making many copies of just the sequence of interest, but in a much quicker and easier way. PCR revolutionized molecular genetics; before PCR, only skilled researchers could amplify and characterize DNA, and the whole cell-based cloning procedure was far too complex to be used for diagnostic purposes. PCR has many advantages over cell-based cloning – it is vastly easier and quicker, but crucially, it is selective. With cell-based cloning you cloned a complete mix of random fragments of the DNA of interest, then had to search through the resulting ‘library’ of thousands or millions of random clones to somehow find the one you wanted. With PCR you choose

which bit of the DNA you want to amplify. At the end of the PCR reaction, although all the original irrelevant DNA is still present, the amplified sequence is present in such excess that the product can be treated as a slightly impure preparation of just the target sequence.

PCR can be used to detect the presence or absence of a sequence or to measure its size. Examples of both those applications are shown below. Additionally, PCR products are suitable for sequencing and testing for point mutations, as we shall see in *Chapter 5*. In principle the starting material can be as little as a single DNA molecule, meaning it is possible (though not easy) to analyze the DNA of a single cell, for example, one cell taken from a pre-implantation blastocyst. Single cell analysis requires special expertise, but the sensitivity of PCR means that it is now routine to use non-invasive mouthwashes rather than blood as a source of a patient's DNA.

PCR depends on three features of DNA replication:

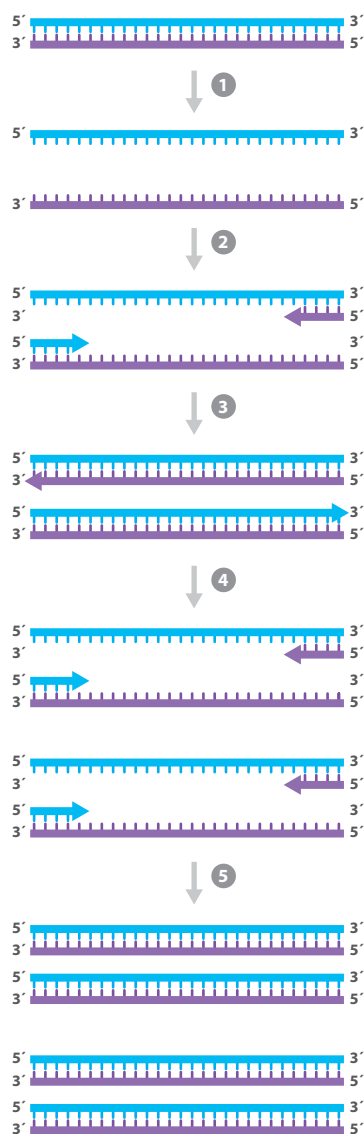
- (1) A new chain requires a **primer**. DNA polymerase (the enzyme that replicates DNA by synthesizing a strand complementary to a single-stranded template) cannot simply start assembling isolated nucleotides; it can only work by extending an existing chain. The primers used in PCR are chemically synthesized single-strand oligonucleotides, normally around 20 nt long.
- (2) Chain extension can only proceed in the 5'→3' direction (see *Box 3.2*).
- (3) The two strands of a DNA double helix are anti-parallel, as shown in *Figure 3.5a*.



The requirement for a primer makes it possible to force DNA polymerase to copy just a selected sequence in a complex DNA sample. A chemical oligonucleotide synthesizer is used to make a large number of molecules of a specific single-stranded oligonucleotide, whose sequence is such that it will hybridize to only one particular sequence in the human genome. A large excess of this primer is added to the DNA. The whole sample is denatured by brief heating to 95°C, then cooled to a temperature at which hybridization can take place (typically 55–60°C). If the starting material consists of DNA from many cells, as is usually the case, some copies of the relevant sequence may re-hybridize to their original partner, but most will end up hybridized to a molecule of the primer because these are present in much greater numbers. If DNA polymerase is then allowed to act on the mixture, most of the action will consist of adding nucleotides on to the 3' end of the primer, building up a strand complementary to that bit of the genomic DNA. *Figure 4.8* summarizes this process.

Figure 4.8 – Using a synthetic primer to force DNA polymerase to synthesize just a strand complementary to a specified part of a large DNA molecule.

(a) The primer (red) is present in a large molecular excess. (b) The target DNA is denatured then cooled to allow hybridization. (c) The primer hybridizes to only one specific sequence. (d) Only sequence downstream of the primer is copied.



The polymerase chain reaction uses this principle, but converts the linear reaction of *Figure 4.8* into an exponential chain reaction by using two primers and multiple cycles of denaturation, hybridization and synthesis. *Figure 4.9* shows how it works. The primers hybridize specifically to opposite strands of the DNA either side of the sequence to be amplified, and are oriented so that 5'→3' chain extension from each primer runs towards the other primer. Thus each strand being synthesized by extending the forward primer comes to include the sequence to which the reverse primer can hybridize, and each strand made by extending the reverse primer comes to include the sequence to which the forward primer can bind. *Box 4.4* explains the process in more detail.

PCR works best for amplifying sequences of 100–400 bp. Sequences longer than 1–2 kb are difficult, and sequences above 20 kb almost impossible to amplify. Uncharacterized DNA fragments, where there is no information for designing primers, can be amplified by using the enzyme DNA ligase to join the same short synthetic oligonucleotides (adapters) to the ends of all the fragments. Primers can then be used that hybridize to the adapters.

PCR reactions can be multiplexed by adding several different primer pairs to the reaction. For multiplexing very large numbers of reactions, rather than using a large number of different primer pairs, the same standard adapters can be ligated to the ends of all the fragments in the mix, so that they can all be amplified using a single pair of primers. Taken to extremes this allows whole genome amplification, which has become an important tool for extracting large amounts of information from small samples.

Figure 4.9 – Principle of PCR.

The PCR reaction contains the target DNA, the DNA polymerase enzyme, large amounts of both primers and a supply of mononucleotides.

1. The starting DNA is denatured by heating to 95°C.
2. At 55–60°C the primers anneal to their target sequences. Note the 5' and 3' directions.
3. At 72°C the polymerase extends the primers in the 5' → 3' direction.
4. Another round of denaturing and annealing primers.
5. Another round of polymerization. After two rounds of PCR we have four strands where originally we had one.

Understanding PCR

In each round of denaturation, hybridization and synthesis, not only the original template but also all the copies made in previous rounds are replicated. Thus the amount of target DNA is doubled in each round – 20 rounds of PCR should suffice to amplify the target sequence one million-fold. Cycles of PCR are controlled by varying the temperature. A few seconds at 95°C denatures everything. Then a few seconds at around 55–60°C allows the primers to anneal to every possible template. Finally, a few seconds at 72°C, the optimum temperature for the special polymerase used, sees each primer extended by adding nucleotides to the 3' end. A typical amplification of 20–25 cycles with each stage lasting 30–60 seconds can be set up and completed in under 2 hours. A programmable heating and cooling block is used that will hold 20 or more individual reaction tubes, so that a number of samples can be PCR-amplified in parallel. Diagnostic labs now commonly perform these reactions in 96-well plastic microtiter plates, using a robotic pipetting machine.

The selectivity of PCR depends on having primers that will anneal only to the desired target and not to any other sequence in the whole genome. The necessary specificity is achieved by using short primers (usually 18–22 nt) and fine-tuning the temperature. If the annealing temperature is too low, the primers may be able to hybridize to mismatched targets, while if it is too high they will not hybridize at all. Usually a temperature in the 55–60°C range allows specific hybridization. Possible primers are checked against the complete genome sequence database to make sure that no other sequence matches the primer. The length and nucleotide composition of the primer can affect the efficiency of the PCR process, and there are computer programs to assist with primer design.

When working out what goes on in PCR it is useful to think of two classes of PCR product:

- Product A is synthesized using the original DNA as the template. Molecules of Product A have a defined 5' end (the 5' end of the primer) but are of indeterminate length.
- Product B is synthesized using Product A as template. Both ends of Product B are defined. The 5' end is the 5' end of the primer used to make it, and the 3' end is defined by the end of its template, which is the 5' end of Product A. All molecules of Product B are exactly the same size.

In later cycles of the PCR reaction, Product B is also used as the template for the next round of synthesis (*Box table 4.1*). The end of a strand made by extending primer 1 is complementary to primer 2 and so, in the next cycle, primer 2 can hybridize to it and prime synthesis of a complementary strand. The end of this strand, in turn, is complementary to



Box figure 4.2 – A typical thermocycler, used to carry out PCR amplification.

When in use an insulating lid covers the tubes.

Box table 4.1 – Progress of the PCR reaction.

	Starting DNA	Product A template is starting DNA; 1 defined end		Product B template can be Product A or Product B; 2 defined ends		
	Single strands	Made in this cycle	Cumulative total	Made in this cycle using Product A as template	Made in this cycle using Product B as template	Cumulative total
After 1 cycle	2	2	2			
After 2 cycles	2	2	4	2	–	2
After 3 cycles	2	2	6	4	2	8 (2 + 4 + 2)
After 4 cycles	2	2	8	6	8	22 (8 + 6 + 8)
After 5 cycles	2	2	10	8	22	52 (22 + 8 + 22)

We imagine starting with one molecule of double-stranded DNA. After the first few rounds, almost all the product consists of just the sequence between the outside ends of the two primers. The numbers are numbers of single strands that are present when all the DNA is denatured.

primer 1. The newly synthesized strand is exactly the same length as its template. After the first few cycles, virtually all the product is made in this way, using Product B as the template.

The best way to understand how PCR works is to take a large sheet of paper and draw out what happens in the first three or four cycles of the reaction. Check your effort against *Figure 4.9*. Take care to mark 3' and 5' ends. Make sure that hybridized strands are always anti-parallel and that strand extension is always 5' → 3'.

BOX 4.4 – continued

4.3. Investigations of patients

We will first consider cases that were investigated by hybridization. These used FISH, MLPA, array-CGH and a SNP chip. Southern blotting might have been used to investigate the **Lipton family (Case 11;** Fragile X), but laboratories try to avoid this difficult and laborious technique by using specialized forms of PCR, as described below. Thus the final three cases used PCR to check for the presence or absence of a sequence and to measure its size. In *Chapter 5* we will see examples of using PCR to check the actual nucleotide sequence.

Cases studied using a hybridization procedure

CASE 7 GREEN FAMILY

- George, aged 3 years
- Developmental delay, mildly dysmorphic
- Normal 46,XY karyotype but suspect microdeletion
- Test for microdeletions
- 22q11 deletion identified by FISH
- Possibilities for therapy

25 39 70 **97** 395

Although conventional cytogenetic analysis showed a normal karyotype (*Figure 2.8*), this did not rule out a chromosomal deletion or duplication smaller than the 5 Mb resolution of standard preparations. George's appearance was suggestive of a deletion of band 22q11. This is caused by recombination between misaligned low-copy repeat sequences on 22q11, as explained in *Disease box 2*. The usual deletion covers about 3 Mb and includes over 20 genes. One of these is the *TUPLE1* gene. A probe containing the *TUPLE1* sequence was used for FISH on a preparation of George's chromosomes. The result showed a deletion on one copy of chromosome 22 (*Figure 4.10*). Because George's features were completely typical of children with the 3 Mb deletion it was not felt necessary to investigate further. Had there been any doubt about it, array-CGH or a SNP chip could have been used to probe the exact extent of the deletion.



Figure 4.10 – 22q11 metaphase FISH.

The green spots are a control probe, used to identify the two copies of chromosome 22 and confirm that hybridization has taken place. The red spots are the *TUPLE1* probe. Only one of the two copies of chromosome 22 contains the sequence that hybridizes to this probe.

CASE 4 DAVIES FAMILY

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy
- X-linked recessive inheritance
- Problems of testing dystrophin gene
- Exon 44–48 deletion identified by MLPA

4

11

68

98

156

285

315

395

As we saw in *Chapter 1*, Martin had features (muscle weakness and calf pseudohypertrophy) suggestive of Duchenne muscular dystrophy (DMD). The pedigree reinforced the suspicion because two deceased maternal uncles had had a progressive neuromuscular disease. The geneticist had obtained their clinical notes. Their clinical course and muscle histology results were all consistent with DMD. The neurologist took a muscle biopsy from Martin, and also a blood sample to measure the level of creatine kinase (CK). CK is an enzyme present within muscle cells. If there is damage to the external cell membrane CK leaks out. The level is often raised in normal people after vigorous exercise, and tends to be permanently raised in female carriers of DMD (it can be used to give a probabilistic indication of a woman's carrier status, see *Figure 14.1*). In affected boys it is strongly and unambiguously elevated.

These tests confirmed the bad news: Martin had DMD. Until very recently the prognosis for boys with DMD was grave – they would suffer increasing disability, be wheelchair-bound by about age 12, and would be unlikely to live beyond their 20s because of problems with their hearts and breathing. As described in *Section 14.4*, there is now hope that new therapies may at least ameliorate the picture, if not provide a full cure.

Initially Martin's parents Judith and Robert were entirely taken up with the emotional shock, and with planning how best to cope with a child with a severe progressive disability. But of course the diagnosis also had implications for any future pregnancies and for the wider family. When they were ready to consider such matters, the geneticist arranged a discussion. It seemed highly likely that Judith was a carrier, in which case any subsequent son would be at 1 in 2 risk. Other female relatives, such as Martin's two sisters, might very well also be carriers. To resolve all these questions, it was necessary to identify the dystrophin gene mutation in Martin.

As we saw in *Chapter 3*, the dystrophin gene is huge, but about two-thirds of all pathogenic changes are deletions of one or more complete exons (see *Figure 3.8*). This is an X-linked condition, so a boy has only a single copy of the dystrophin gene, making deletions easy to detect. If no deletion is found, this would not rule out the diagnosis because one-third of DMD cases have point mutations or duplications rather than missing exons.

The simplest way to check for missing exons would be a direct PCR assay. The 79 exons of the dystrophin gene are all under 300 bp long (except for exon 79, the last exon, which is 2703 bp long – the last exon in a gene is often large, but consists mainly of the 3' untranslated region). Thus individual exons can be readily amplified by PCR. The products can be run out on an electrophoretic gel and any deletion would show up as absence of the relevant band. Primers can be designed to incorporate varying amounts of the flanking intron sequence so as to give a different size product for each exon, allowing multiple products to be run together on the same gel. However, nowadays laboratories would be more likely to use MLPA. Using a commercial kit, 40 exons can be checked in a single reaction, but more importantly, exon duplications can be detected and females can be checked to see if they carried deletions or duplications. These would be in heterozygous form in a carrier female, and the simple PCR test would not reliably pick them up.

The MLPA test showed that Martin had a deletion of exons 44–48 of the dystrophin gene (see *Figure 4.11a*). This confirmed the diagnosis, which also confirmed that Martin's mother Judith was an obligate carrier and clarified the risk of recurrence in any future children of hers. However, it left her two daughters (Martin's sisters Lisa and Jessica, see pedigree, *Figure 1.9*) at a 50% risk of being carriers. The geneticist stayed in contact with the family, and when each daughter was in her late teens she was counseled about the risk and offered a test. This was done by MLPA. The result (*Figure 4.11b*) showed that Lisa was a carrier, while Jessica was not. Lisa then had the option of prenatal diagnosis (probably by fetal sexing using PCR for Y-chromosome sequences followed, if the fetus was male, by PCR tests for the presence or absence of exons 44–48 of the dystrophin gene) in any pregnancy, if that was what she and any partner wanted. Before this test was available many women who were definite or probable carriers opted for fetal sexing and took the traumatic decision to abort all male fetuses, even though there was only at most a 1 in 2 chance of the fetus being affected. With prenatal diagnosis they can be confident, if the tests are normal, of giving birth to boys free of the disease. The implications of this deletion for function of the dystrophin gene are considered in *Chapter 6*, while the impact of these developments on genetic services are described in *Chapter 14* – see *Figure 14.1*.

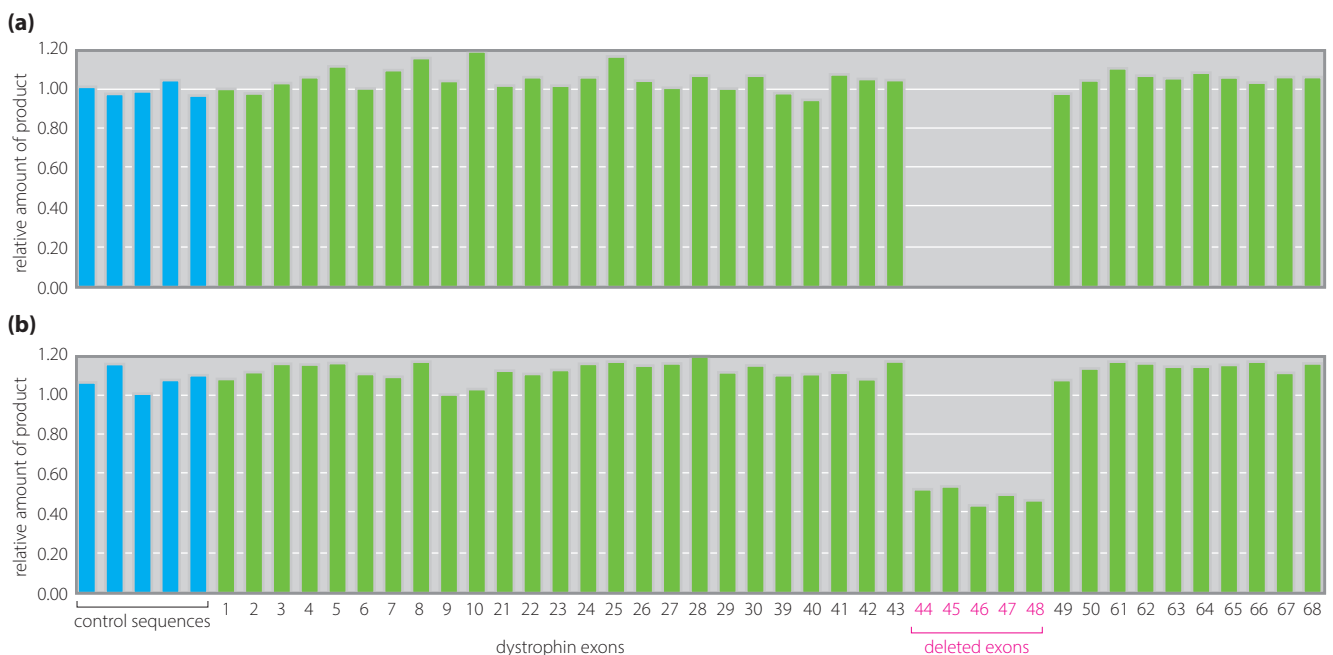


Figure 4.11 – Using MLPA to detect exon deletions in the dystrophin gene in the Davies family.

A multiplex MLPA assay tested selected exons of the dystrophin gene. A separate multiplex MLPA test (not shown) could have been used to check the remaining exons. The bar charts show peak intensities for each product, as measured on a capillary sequencer. Bars in blue are from unrelated control sequences, bars in green are from individual exons of the dystrophin gene, as numbered. (a) This shows that Martin has a deletion of exons 44–48 on his only copy of the X-linked dystrophin gene. (b) His sister Lisa is heterozygous for the deletion. Courtesy of Dr Simon Ramsden, St Mary's Hospital, Manchester.

CASE 5 ELLIOT FAMILY

- Baby girl Elizabeth, parents Elmer and Ellen
- Multiple congenital abnormalities
- Family history of reproductive problems
- ? Chromosome abnormality
- Ellen – balanced 1:22 translocation
- Elizabeth – unbalanced segregation product
- Reciprocal translocation
- Translocation identified by array-CGH
- Possibilities for therapy

4

12

43

68

100

395

In *Chapter 2*, we showed karyotypes of Ellen Elliot, the phenotypically normal carrier of a balanced 1:22 translocation, and her baby Elizabeth who inherited an unbalanced karyotype resulting in multiple congenital abnormalities (*Figures 2.14 and 2.15*). In reality, the first investigation of baby Elizabeth would have used array-CGH. This, or the use of SNP chips, is currently the method of choice for investigating cases with a suspected chromosome abnormality where there is no clear hypothesis as to its location. The karyotype in *Chapter 2* was shown for educational purposes, because it helped make clear the mechanism involved.

The result (*Figure 4.12*) showed that Elizabeth was trisomic for the distal part of chromosome 1 and monosomic for the distal part of chromosome 22. These results strongly suggested she was the result of unbalanced segregation of a balanced translocation in one parent, but proving that required a conventional karyotype to identify the balanced translocation in Ellen (*Figure 2.14*), since array-CGH cannot detect balanced abnormalities.

Eighteen months after Elizabeth's birth, Ellen found herself pregnant again. Knowing there was a substantial risk of another abnormal baby, Elmer discussed whether she should terminate the pregnancy. Ellen reminded him of the geneticist's reassurance that a prenatal test would be possible. They recontacted the geneticist, who arranged a rapid referral for chorion villus biopsy (see *Box 14.5*). The fetal cells were karyotyped,

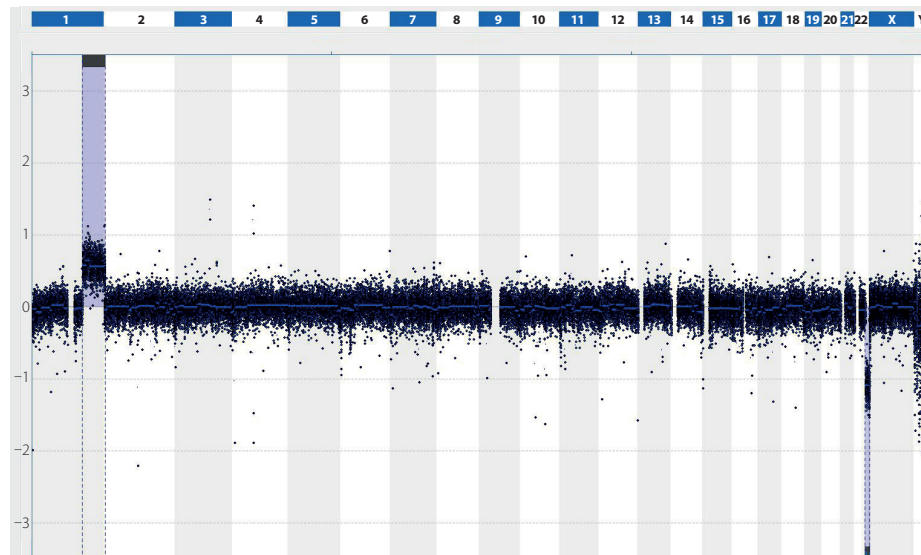


Figure 4.12 – CGH reveals copy number variants in Elizabeth Elliot's DNA.

For each cell of the array the relative intensity of hybridization of Elizabeth's DNA and the control DNA is shown as a dot plotted on the vertical axis. Results are arranged along the horizontal axis by the chromosomal location of each probe. The array contains a total of 60 000 probes, so most of the individual dots fuse together as a continuous band. The result shows that she is trisomic for part of chromosome 1 and monosomic for part of chromosome 22. Data produced using Oxford Gene Technology 8*60 array. Courtesy of Lorraine Gaunt and Ronnie Wilson, St Mary's Hospital, Manchester.

and a normal 46,XY result was obtained. Reassured, Ellen went on to produce a healthy baby son.

This case carries a lesson for health service planners. The genetic intervention achieved the best of all possible results: a worried couple were reassured, an abortion was avoided and a healthy child was born. It would be all too easy to record the outcome of the genetic intervention as 'no action'. Clinical geneticists are very concerned not to have their achievements measured by the number of abnormal fetuses detected and aborted.

CASE 12 MEINHARDT FAMILY

84 101 395

- Madelena, baby daughter of Margareta and Manfred
- Multiple congenital abnormalities and developmental delay
- Normal 46,XX karyotype under the microscope
- 16p microdeletion identified by SNP chip
- Is this microdeletion pathogenic?
- ? Recurrence risk
- Possibilities for therapy

As with **George Green (Case 7)**, the Meinhardt's daughter Madelena had features that suggested a chromosomal abnormality, but no abnormality was detectable on conventional karyotyping. Madelena's combination of intellectual disability and dysmorphic features did not suggest any specific syndrome, so it was not possible to decide on one FISH probe that would give a useful test. Instead, her whole genome was scanned for copy number variations that would indicate a deletion or duplication. This could have been done by array-CGH, as with Elizabeth Elliot, but in this case a high-resolution SNP chip was used. In about 12% of similar cases array-CGH or SNP chips identify submicroscopic chromosomal deletions or duplications. The precise abnormality seen varies widely between different cases.

DNA was extracted from a blood sample and sent for analysis. Madelena's DNA showed a decreased dosage for a series of contiguous clones from position 16p13.11–12.3 on the SNP chip. The data showed that the deletion involved 2.8 Mb of material, including a number of genes, as shown in *Figure 4.13*. The short arm of chromosome 16 is particularly rich in low-copy repeats that predispose to deletions and duplications by non-allelic homologous recombination, as explained in *Disease box 2*. As a result, several different recurrent microdeletions of sequences on 16p have been described. Madelena's deletion overlapped one of these but was larger: 2.8 Mb rather than the 1.5 Mb of the previously described deletions.

Before communicating this result to Madelena's parents, it was important to try to decide whether or not the deletion was the cause of her problems. This seemed highly likely since Madelena's deletion included a 1.5 Mb region, recurrent deletions of which were known to be pathogenic. A necessary check was to see whether either parent carried the deletion, or whether it arose *de novo* in Madelena. If the deletion was *de novo* this strengthens the case for it being pathogenic. The converse is less decisive: finding the variant in one of the clinically normal parents does not completely exclude a role in the pathogenesis, although clearly any role could be only contributory, not completely causative. Several variants (for example, microdeletions at 1q21 or 15q13.3) have been described that are unquestionably more frequent in people with intellectual disability, autism or schizophrenia, but that are also sometimes found in a clinically unaffected parent. It appears that these variants act as susceptibility factors, increasing the chances of a range of psychiatric conditions, but not inevitably causing any one. For a good discussion of the complexities see Girirajan *et al.* (2010).

It turned out that this was a *de novo* event. That did not prove that it is pathogenic, but made it more likely. Combined with the known pathogenicity of smaller deletions of the

same region of 16p, the strong balance of probabilities was that this was the cause of Madelena's problems. This was an important conclusion because it suggested that the recurrence risk was very low. The deletion was a *de novo* event, and there was nothing to suggest any factor in either parent that might predispose to a recurrence. Counseling was that the recurrence risk was not zero but was very low. Because of all the unknown factors, it was not possible to give a precise figure – however, it was small compared to the general risk that attaches to any pregnancy. It would be possible to test for this specific abnormality in future pregnancies if the parents so wished.

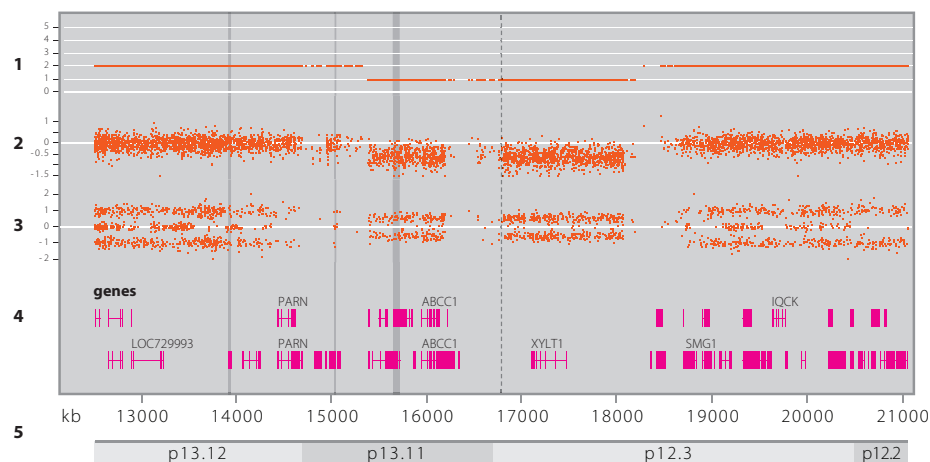


Figure 4.13 – A SNP array reveals a microdeletion in Madelena Meinhardt's DNA.

Across the bottom is an ideogram of chromosome 16, showing the bands and physical distance from 16pter. As with the CGH data (Figure 4.12) dots in Track 2 represent the data from each cell. Hybridization intensity, summed across both alleles of each SNP, is plotted vertically, chromosomal location horizontally. Track 1 shows the interpretation: there is only a single copy of the central part of the sequence, between positions 15 400 and 18 200. Track 3 shows the genotype at each SNP. In the non-deleted regions there are three possible genotypes, 1-1, 2-1 or 2-2 (arbitrary numbering of alleles), while in the deleted region there are only two, 1 or 2. In summary, there is a 2.8 Mb deletion encompassing the genes shown in track 4. Data generated using an Affymetrix SNP 6[®] microarray, courtesy of Lorraine Gaunt, St Mary's Hospital, Manchester.

CASE 3 KOWALSKI FAMILY

- Karol, first son of Kamil and Klaudia
- Developmental delay, hypotonic, severe intellectual disability
- Difficulties of genetic testing in such cases
- Likely need for exome sequencing
- Negative SNP chip test for microdeletions

3 10 67 **102** 134 155 395

Karol's DNA was analyzed on a microarray, as with Elizabeth Elliot, but no pathogenic copy number variant was seen. It seemed likely that his problem was due to a point mutation in some or other gene. This would require DNA sequencing (Chapter 5).

Cases studied using PCR

CASE 9 INGRAM FAMILY

- Isabel, 10 years old with small stature and possibly delayed puberty
- ? Turner syndrome
- 45,X karyotype
- Risk of Y-chromosome DNA
- PCR test for Y sequences negative

26

42

70

103

285

395

The diagnosis in Isabel has been clearly established from her karyotype (*Figure 2.13*). However, as explained in *Chapter 2*, it is important to check whether she has any 46,XY cells, because such cells in the gonads have the potential to become malignant. XY cells might be present if the cause of the syndrome in Isabel's case were loss of the Y chromosome during an early mitotic division of a 46,XY embryo. To check for the presence of XY cells a PCR reaction using Y-specific primers is performed to see if there is any product. This is a much more sensitive method than looking for XY cells in a conventional cytogenetic preparation. Ideally the test would be done on tissue from her streak gonads, because that is the tissue we are concerned about developing a malignancy, but usually the test is done on blood samples. In Isabel's case, no Y chromosome sequence could be amplified, so reassuring us that it is unlikely she has XY cells in her gonads.

CASE 1 ASHTON FAMILY

- John, healthy 28-year-old son of Alfred Ashton
- Family history of ? Huntington disease
- Autosomal dominant inheritance
- Need for diagnostic PCR test
- PCR test confirms diagnosis in John's father
- Pros and cons of predictive test

1

8

67

103

153

395

The next step in investigating this family is to perform a diagnostic test on John's father Alfred. Alfred is already showing signs of disease. Although it seems highly probable that the family disease (see the pedigree in *Figure 1.7*) is Huntington disease, this needs to be proved before John can be given accurate counseling or offered a predictive test. His doctor takes a 3 ml blood sample from Alfred and sends it to the laboratory, where DNA is extracted.

As mentioned in the previous chapter, the laboratory will want to check the size of a run of glutamine codons (CAG) in exon 1 of the *HTT* gene on chromosome 4. The normal range is 5–35 repeats (15–105 bp); anything over 35 repeats is normally pathogenic, although some people with 36–39 repeats remain unaffected. Expansions are normally in the 40–60 repeat range, and are virtually never over 100 repeats or 300 bp. This sort of test is easy to perform by PCR. Primers are designed that flank the region of interest, and the size of the product is measured. The products might be sized on a manual gel; more commonly an automated gene analyzer (the sort of machine shown in *Box figure 4.1b*) would be used. The result shows two peaks because a genomic DNA sample includes DNA from both copies of the *HTT* gene. We see that Alfred's two copies of the gene have 18 and 41 CAG repeats, respectively (*Figure 4.14*, lower trace). As a repeat size over 35 is pathogenic, this result confirms the diagnosis.

Now that Alfred's diagnosis was definite, it followed that John was at 50% risk of developing Huntington disease himself. If he wished, a PCR test on his DNA would tell him for sure whether or not he had inherited the *HTT* gene mutation (see *Box 14.4* for the predictive test protocol). Over several sessions with the genetic counselor, John and his wife Joan pondered long and hard over whether or not to take the test. Initially John thought it would be a bad idea because having a genetic test could affect his ability to get insurance, and maybe his job.

On discussion, he became convinced that these anxieties were misplaced. If he applied for life insurance in the UK he would have to reveal his family history of Huntington

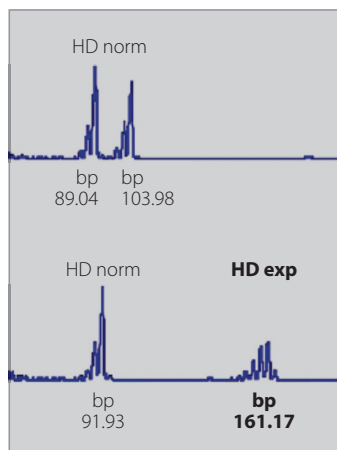


Figure 4.14 – Testing for the Huntington disease mutation.

This is a typical PCR application, amplifying a sequence a few hundred bp long using specific primers, and sizing the product by capillary electrophoresis. Smaller molecules move faster through the capillary, so the position of a peak is a measure of the size of the PCR product. One of the PCR primers carried a fluorescent label to allow the machine to detect the product. The PCR primers used total 38 nt in length, so the number of CAG repeats is calculated by subtracting 38 from the fragment size and dividing by 3. So, for example, the 161 bp product in the lower trace contains $(161 - 38) / 3 = 41$ CAG repeats. Each sample shows two peaks, corresponding to the different sized PCR products produced by the two copies of chromosome 4. Upper trace: a normal result (17 and 22 repeats); lower trace: an abnormal result (18 and 41 repeats). The upper limit of normal is 35 repeats. Courtesy of Dr Simon Ramsden, St Mary's Hospital, Manchester.

disease, and many companies would see this as a reason for heavily loading the premium. In fact a genetic test might help John: if it was negative, the family history would no longer count, while in the UK, if it was positive he wouldn't have to declare the result for life policies up to a certain ceiling under a moratorium agreed between the government and the insurance industry; in any case most of the bad news was already present in the family history. In the USA, the Genetic Information Nondiscrimination Act (GINA) provides wide-ranging protection against genetic discrimination in health insurance, including prohibiting the use of family history information (but does not explicitly disallow genetic discrimination in the provision of life insurance, disability insurance, or long-term care insurance). Unless John developed symptoms that prevented him doing his job properly, his genotype would be of no legitimate interest to any employer. John's wife was then worried about his existing insurance policy. He had a life insurance policy to pay off the mortgage on his house should he die. But when John took out that insurance he was unaware of the nature of the family problem. He had answered all the questions on the form honestly, to the best of his then knowledge, and so that policy was valid and was not affected by later developments.

Having disposed of those concerns, the discussion returned to the basic question of whether or not John and Joan wanted to know. This is a very personal and difficult decision, but in the UK about 80% of people in John's position make the choice not to know. Of course a negative test would be a great relief, and a positive test would at least allow John and Joan to start planning for the future. But most of us like to see our lives as open-ended. Few of us would wish to know, at the age of 28, when and how we will die. After long reflection, John chose not to know. The counselor would have supported him whichever way he chose. Now she arranged to see them over the coming months to continue her support, and assured them that she was always available at any time for further discussion. She also pointed out that there is hope that a recently developed drug could slow the very early development of the disease. If confirmed, it might be in John's interest to have the test, and she promised to keep them informed of developments.

John's wife Joan did have one serious reservation about John's desire not to know. If John did carry the pathogenic variant, any child of theirs would have a 50% risk of inheriting it and thus developing Huntington disease. Would John consent to them having prenatal diagnosis? If the result was negative there would be no problem, but if it was positive that

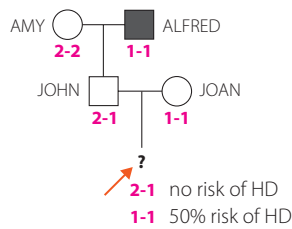


Figure 4.15 – Using a linked marker to identify the risk to the fetus without revealing John's HD status. See text for discussion.

would prove that John must carry the disease allele. John was far from happy about that possibility, and it looked like this might cause major problems between John and Joan. Fortunately, the counselor was able to propose a mutually acceptable solution.

The whole immediate family, including the fetus, could be all typed for a harmless non-pathogenic DNA variant that was located close to the *HTT* gene on chromosome 4p16 (a DNA marker – see Box 8.1). Figure 4.15 shows the logic. In themselves the marker genotypes do not in any way show a person's HD status – for example, in Figure 4.15 Joan and Alfred both have the same marker genotype, although Alfred is affected and Joan is not at any risk. The marker alleles are just labels, with no meaning in themselves, but they can be used to track the transmission of chromosome 4p16 through the family. We know that John's marker allele 2 comes from his unaffected mother Amy, so any fetus that inherits that allele from John is at no risk of HD. John's marker allele 1 comes from Alfred, but we do not know whether it is on Alfred's HD chromosome or his normal chromosome. John's marker genotype tells us nothing new about his own risk – we already knew he was at 50% risk. But any fetus that inherits it also carries a 50% risk. The test requires a marker for which Alfred and Joan were homozygous and John heterozygous – but there are many independent possible markers close to the *HTT* gene, so if one does not have the desired combination of genotypes, another can be tried.

A positive prenatal test would not mean the fetus would definitely have inherited Alfred's HD allele; it would only be at 50% risk. But Joan and John agreed that in that case she would terminate the pregnancy. Thus a difficult potential dispute was resolved.

The final discussion concerned what to do about John's other at-risk relatives. His sister Helen, her two young sons, and his aunt Alice in Australia were all at risk as a result of the positive test on Alfred. It was agreed that John should contact Helen and Alice, explain the situation and give them contact details of their local genetics service. Whether or not they chose to take up the contact was then up to them. Had John chosen to take the test, the geneticist would not have revealed that fact or the result to any other family member. Since John was close to Helen, he might choose to discuss with her his own thoughts and decisions, but there was absolutely no obligation on him to do this.

CASE 11 LIPTON FAMILY

83 105 395

- Baby boy, Luke, with developmental delay
- Family history of learning difficulties
- Unusual features of Fragile-X pedigrees
- Caused by unstable repeat expansion
- Premutations, full mutations and normal transmitting males
- Measuring repeat expansions
- Possibilities for therapy

Fragile X syndrome (OMIM 300624) was first described as a form of X-linked intellectual disability, associated in affected males with a prominent high forehead, a long face, large jaw, large low-set ears and strikingly large testes (macro-orchidism). When lymphocytes from patients were cultured in media deficient in folate (folate is essential for DNA replication), the X chromosome in a proportion of the cells was seen to show a decondensed region (a 'fragile site') at Xq27.3, near the end of the long arm. Pedigrees showed several unusual features. The condition was not fully recessive because carrier women often had some degree of mental slowness and occasional cells showing the 'fragile' X chromosome. The risk of being affected seemed to increase going down through the generations, while at the top of the pedigree there was often a mentally normal male who might have several carrier daughters, implying that he must have carried the pathogenic mutation but not been affected by it – a 'normal transmitting male'. The Lipton family pedigree (Figure 4.1) illustrates these features.

When the disease gene, *FMR1*, was cloned in 1991 it brought yet more surprises. The pathogenic change was an increased number of repeats in a run of tandemly repeated CGG trinucleotides in exon 1 of the gene. The repeats are in the 5' untranslated region, so they are present in the mRNA but not the protein. Repeat numbers up to 45 are stable and non-pathogenic, but larger repeats become progressively more likely to expand on transmission. The classic Fragile X phenotype is seen when the repeat number exceeds approximately 200. This behavior was unprecedented at the time, but subsequently a number of other diseases were found to depend on similar dynamic mutations, including Huntington disease as described above. Over 40 are now known; *Disease box 4* describes a selection.

Alleles with 46–58 repeats are labeled 'intermediate'; they are usually stable but carry some risk of expanding. Those with 59–200 repeats are called premutation alleles: they do not cause the classic Fragile X syndrome, but they are unstable in female meiosis and liable to expand to cause the full syndrome in offspring. People with premutation alleles are at risk of other, apparently unrelated conditions: 15–20% of female premutation carriers have primary ovarian insufficiency (menopause before age 40), while male premutation carriers have a 1 in 3 risk of developing a neurodegenerative syndrome, FXTAS (fragile X tremor/ataxia syndrome, OMIM 300623) after age 50. Female premutation carriers also occasionally develop FXTAS. The problems in premutation carriers are thought to reflect a toxic effect of mRNA carrying the expanded CGG run. The classic Fragile X syndrome, on the other hand, is caused by lack of the FMR1 protein. *FMR1* genes carrying over 200 CGG repeats are not transcribed. The large repeat triggers methylation of promoter sequences which affects the chromatin configuration so as to prevent transcription (see *Chapter 11*).

Fragile X syndrome is diagnosed by checking the size of the CGG repeat. However, there are complications. Normal and most premutation alleles can be amplified by PCR, much as in Huntington disease. However, PCR does not work well for repeat sizes above around 120 units. Full mutations can involve 1000 or more repeats, and standard PCR will fail to provide clear evidence of such a full expansion. Full mutation alleles can be reliably detected using Southern blotting, but this technique is laborious, time-consuming and requires a high degree of laboratory skill. A modified PCR method, triplet-primed PCR, is widely used as a simpler alternative (Tassone *et al.*, 2008). One PCR primer hybridizes to a sequence flanking the (CGG)_n repeat, but the other hybridizes to the repeat itself. On different individual molecules in the sample the triplet-specific primer might hybridize to different positions within the repeat sequence, so the result of the PCR is a ladder of bands stemming from different positions of the triplet-specific primer (*Figure 4.16*). If the tail of the ladder extends into the pathological range this confirms the presence of an expansion, although it would not distinguish premutation (59–200) from full mutation (>200) repeats or identify the actual number of repeats. Southern blotting would be needed to provide that information.

Analysis in the Lipton family (*Figure 4.17*) confirmed the diagnosis:

- Baby Luke Lipton had a full expansion. Southern blot analysis showed approximately 800 CGG repeats.
- Linda herself was heterozygous with a premutation allele with 120 repeats and a normal allele with 38 repeats.
- Linda's mother Maria was heterozygous with a premutation allele with 78 repeats and a normal allele with 43 repeats.

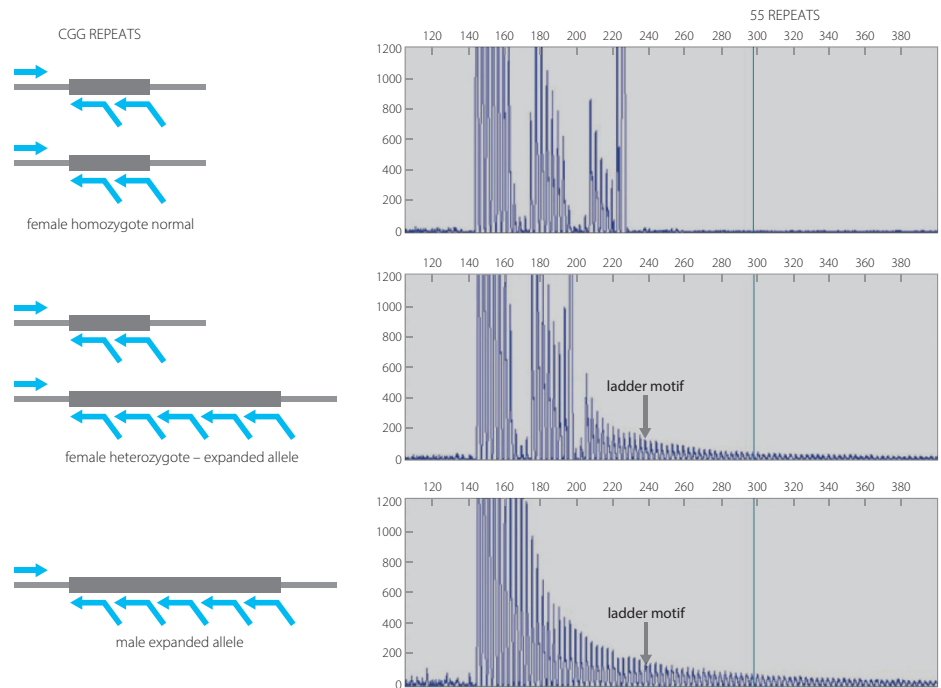


Figure 4.16 – Triplet-primed PCR analysis of Fragile X expansions.

If the tail of the ladder extends beyond 55 repeats, this indicates that either a premutation or a full mutation allele is present. Southern blotting would be required to determine the exact size of the expansion. Results obtained using the FMR1 TP-PCR kit from Abbott Molecular.

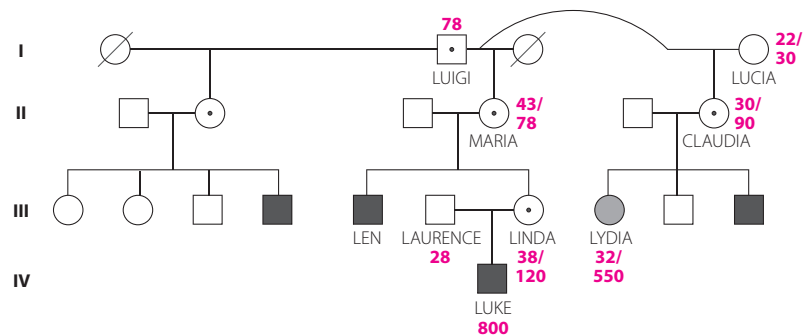


Figure 4.17 – CGG repeat sizes in Linda Lipton's family (Case 11).

See text for details.

- Linda's cousin Lydia was heterozygous with a full expansion (shown by Southern blotting to have approximately 550 repeats) and a normal allele with 32 repeats.
- Lydia's mother Claudia was heterozygous with a premutation allele of 90 repeats and a normal allele of 30 repeats.
- Luigi, the patriarch of the family, was a premutation carrier (a 'normal transmitting male') with 78 repeats.
- Lucia (Claudia's mother) and Laurence (Lydia's husband) show the patterns of normal females and males, respectively.

Linda and Laurence asked about the risks to other children they might have. There was a 1 in 2 chance any child would inherit Linda's premutation allele – the question was, how likely was it to expand, and by how many repeats? Empirical data suggests that the risk of expansion to a full mutation is quite high since Linda's premutation is at the higher end of the repeat range. The geneticist therefore told them that there was 'up to a 50% chance that a child would inherit a full mutation'. They then asked what the implications were if the child was a girl; they had already worked out that a boy might be similarly affected to Luke. This was a simple question but difficult to answer. The geneticist gave them the following information. 'Only about one-third to one-half of affected females have learning problems, and they are usually less severe overall than affected males. However, even some affected girls with normal intelligence have areas of difficulty in learning due to poor attention span, and they can have poor social skills.' The geneticist suggested that if the family were going to have more children and wanted to consider prenatal diagnosis, because of the uncertainties in the clinical interpretation of laboratory results, they might wish to have a longer session with a genetic counselor to help them decide what course of action they would choose in the various scenarios that might occur.

4.4. Going deeper...

Table 4.1 summarizes the techniques covered in this chapter and the type of problem for which each is appropriate. In some cases (Cases 1, 4, 7 and 11) the location and/or nature of the suspected genetic lesion was known in advance. Often this is not the case, and then global techniques (array-CGH or SNP chips) are needed for variants involving kilobases of DNA. DNA sequencing, discussed in *Chapter 5*, is a general method for detecting nucleotide-scale changes, and some additional techniques are considered below.

Table 4.1 – Summary of the methods described so far, and their main applications

Principle	Method	Application
Hybridization	Southern blotting	Checking for large-scale changes (inversions, deletions, etc.) that alter the pattern of restriction fragments, and for trinucleotide repeat expansions that are too large to amplify by PCR.
	FISH on a chromosome spread	Checking for presence / absence and chromosomal location of a known sequence at least several kb long.
	FISH on interphase cells	Checking for copy number of a specific chromosome(s); two-color FISH can detect rearrangements.
	MLPA	Deletion or duplication of single exons of a specific gene.
	Array-CGH	Scanning the entire genome for any copy number change involving sequences of a few kb or longer.
	SNP arrays	As with array-CGH, potentially with higher resolution but lower sensitivity. The genotypes of the patient can be compared with those of the parents to reveal some unusual types of variants.
Amplification	PCR	Checking the presence / absence and size of a specific 50 bp – 5 kb sequence. Refinements of standard PCR (see below) are used to quantify the amount of a sequence or to study RNA.



Figure 4.18 – Digital droplet PCR.

The PCR shows the proportion of droplets that contained a molecule of the test DNA.

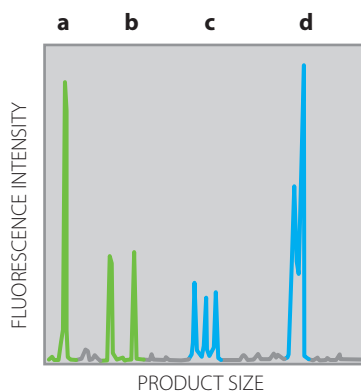


Figure 4.19 – Quantitative fluorescence PCR: a method for rapid detection of common chromosomal trisomies.

Products (a) and (b) are from a normal disomic chromosome, products (c) and (d) are from a chromosome that is present in three copies.

Quantitative PCR

When sequences are amplified using standard PCR protocols the results are not strictly quantitative – that is, the amount of product does not necessarily closely reflect the amount of template that was present in the original sample. This makes standard PCR unreliable for detecting duplications or heterozygous deletions. Quantitative PCR (qPCR) or digital droplet PCR (ddPCR) overcome this limitation.

qPCR follows the progress of the reaction in real time. Early cycles are followed, where accumulation of the product is still exponential, before the supply of primers or monomers starts to limit progress. By various methods (SYBR Green, TaqMan) the accumulation of product can be followed by a growth of fluorescence, so that a special machine can follow it in real time through each cycle of the reaction. The analysis compares the number of cycles needed for the fluorescence in the test sample to reach a certain threshold, compared to the number for a reference sample. One application of qPCR would be as an alternative to MLPA to check females in the **Davies family (Case 4)** to see if they are heterozygous carriers of a deletion of specific exons of the dystrophin gene that had already been characterized in the affected boy.

qPCR depends on comparing the test DNA to a reference sample. **Digital droplet PCR** (ddPCR) (Figure 4.18) directly counts the number of DNA molecules in the test sample, without the need for a reference sample. A special machine processes the test sample into thousands of minute droplets. The trick is to dilute the sample such that some of the droplets do not contain any of the test DNA, while the remainder mostly have only a single molecule. All the droplets are put through a PCR reaction, and the proportion of droplets that contain product is observed. After correction for the fact that some droplets might have contained two molecules of the test DNA, the proportion gives a direct count of the number of molecules in the original sample. Differently labeled primers can be used to compare the proportions of two different sequences, for example a normal and a mutant sequence. By counting enough droplets, even very rare variants can be detected and quantified.

QF-PCR (quantitative fluorescence PCR, Figure 4.19) is widely used as a rapid prenatal test for the common chromosomal trisomies (Allingham-Hawkins *et al.*, 2011). This is not a real-time method; it depends on using a gene analyzer to compare the relative amounts of product from a multiplexed series of microsatellite markers (see Box 8.1) from chromosomes 13, 18 and 21, and sometimes also the X and Y. Typically five loci from each autosome would be amplified. Primers are designed and labeled so that the PCR products from each locus are clearly identifiable by their size and color. Because microsatellite alleles vary slightly in length, different alleles of the same microsatellite often give peaks at slightly different positions. Thus a locus may amplify as one larger peak or two smaller peaks. If there is a trisomy the three alleles of each locus on that chromosome may give three small peaks, two peaks in a 2:1 size ratio, or sometimes a single large peak (an uninformative result for that locus, but hopefully others of the five loci from that chromosome will be informative).

Targeted versus overall testing for chromosome abnormalities. Conventional karyotyping and array-CGH can pick up appropriate abnormalities affecting any part of any chromosome. By contrast, QF-PCR looks only for copy number variants of three or five specific chromosomes. Any other abnormalities would not be picked up. The same is

true of current implementations of testing free fetal DNA in maternal blood (see *Disease box 12*). There has been considerable debate whether this is a good or a bad thing. The majority of those other abnormalities would have uncertain implications and prognosis, and would certainly not fit the criteria for screening set out in *Box 12.4*. But some would result in liveborn abnormal babies. The question is further discussed in *Box 12.2*.

Table 4.2 summarizes the various methods that could be used to detect chromosomal abnormalities. A further option for prenatal diagnosis of fetal trisomies, testing free fetal DNA in the maternal blood, is described in *Disease box 12*.

Table 4.2 – Methods for detecting chromosome abnormalities

	Traditional karyotyping	Array-CGH	QF-PCR	Metaphase FISH	Interphase FISH
All chromosomes studied	+	+	—	—	—
Trisomies 13, 18, 21	+	+	+	—	+
45X, 47XXX, 47XYY	+	+	+	—	+
Triploidy	+	—	+	—	+
Deletions, duplications	+	+	—	+	—
Microdeletions, duplications	—	+	—	+	—
Balanced translocations	+	—	—	+	+ ¹
Unbalanced translocations	+	+	—	+	—

¹2-color interphase FISH is sometimes used to check for a specific translocation in cancer cases, as illustrated in *Figure 7.10*.

Chromosome painting

This technique is a variant of FISH in which a whole cocktail of sequences from one particular chromosome is fluorescently labeled and used as a probe. In a metaphase spread the whole length of that chromosome appears brightly colored. Its use is in identifying the origin of abnormal chromosomes seen on standard karyotyping. For example, if we used a chromosome 1 paint on cells from **Case 5 (Elizabeth Elliot)**, *Figure 2.15*, the paint would highlight both copies of her normal chromosome 1 plus that part of her translocated chromosome that was derived from chromosome 1. In this particular case we don't need chromosome painting to understand the karyotype, but patients sometimes show small extra 'marker' chromosomes, or extra material inserted into a chromosome, whose origin can be impossible to identify under the microscope without chromosome painting. Extending the concept, M-FISH or SKY use a mix of paints, one for each chromosome, each labeled with a different mix of fluorescent dyes, so that each chromosome is painted a different color. We will meet this technique when considering the very complex chromosomal changes typical of leukemia and cancer cells (*Box figure 7.1*). Genome sequencing has now largely replaced these techniques as a means of identifying the origins of abnormal segments, but it cannot reveal their current chromosomal locations.

Testing RNA

Sometimes it is desirable to study RNA rather than DNA. For example, a basic question about any gene is when and where (in which tissue or organ) is it expressed? Also, hunting for mutations, especially those that affect splicing, can be easier if mRNA rather than genomic DNA can be studied. The example of neurofibromatosis 1 was described in *Disease box 1*: the pathogenic change in a patient might be in any of the 59 exons of the *NF1* gene, or indeed deep within an intron if it disrupts the exon–intron splicing. It could be much easier to see if we tested the mRNA rather than genomic DNA. However, mutant genes often produce no detectable mRNA (see *Chapter 6*), and RNA is harder to obtain and handle than DNA. The appropriate tissue has to be sampled, for example, dystrophin mRNA would be obtained by a muscle biopsy. Also, RNA is unstable, requiring stringent precautions in the laboratory. Various companies offer RNA sampling tubes that contain special reagents to stabilize the RNA.

RNA cannot be PCR-amplified or sequenced like DNA (although some novel sequencing technologies may eventually render the latter possible). Most RNA studies therefore first make a DNA copy of the RNA, and study the copy. This is a reversal of the normal DNA → RNA flow of information, as described in the Central Dogma (*Figure 3.2*). Some viruses encode an enzyme, reverse transcriptase, that makes DNA copies of a template RNA. The final product is a double-stranded **cDNA** (complementary DNA). In **RT-PCR** (reverse transcriptase–polymerase chain reaction; not to be confused with real-time PCR) reverse transcription is performed on total extracted mRNA, and gene-specific primers are then used in a PCR reaction on the total cDNA to amplify the chosen gene. RT-PCR can be performed as a single operation and is a powerful tool, provided a tissue can be sampled where the gene of interest is expressed. The complete mRNA repertoire of a cell type or tissue (the **transcriptome**) can be studied by mass sequencing of cDNAs. This has become an important tool of investigation in cancer genetics, and more widely in studies of cell biology.

Testing protein

If a disease is caused by absence of a particular protein, which in turn may be the result of any one of a large number of mutations in the relevant gene, might it not be simpler to test directly for the protein rather than hunting through the whole gene for any possible mutation? In principle the answer should be yes. However, while DNA from any sample can be used to test any gene, the relevant protein may be present only in an inaccessible tissue. Also, DNA tests are generic while protein tests are specific. That is, a laboratory can apply the same well-tried techniques to test the DNA of any gene, whereas each protein needs a specific assay that must be developed, set up and optimized. Protein testing is best done using commercial kits, where the company will optimize the reagents and protocol. However, companies will only develop these for relatively common conditions. We will see examples of protein testing used for population screening in *Chapter 12*.

The total protein content of a tissue (the **proteome**) can be analyzed by mass spectrometry. This can yield information on protein abundance, on tissue-specific patterns of gene expression, and on the range of protein isoforms produced by alternative splicing or post-translational modification. All this information may be highly relevant biologically, but it is not accessible through DNA analysis (Wilhelm *et al.*, 2014). Such proteomics analysis is an important research tool, but currently its use in clinical diagnosis is limited to screening newborn infants for inborn errors of metabolism and to a few commercial tests that use proteomic signatures as markers of disease (see *Chapter 12*).

Diseases caused by expanding nucleotide repeats

The mutation in Huntington disease is unusual but it is not unique. Starting with Fragile X syndrome in 1991, a growing list of human diseases have been identified that are caused by expansions of a nucleotide repeat. Mostly the repeats are trinucleotides, as in Huntington disease, but examples are known that involve 4, 5, 6 and 12 nucleotide repeats. In every case the repeat is present in normal people, where it is stable and non-pathogenic. The number of repeat units varies among normal people as a result of occasional glitches in replicating such repetitive DNA sequences, but it is always below some threshold number. If one of these rare mistakes creates a repeat that is above the threshold, it becomes unstable and has a high probability of expanding yet further on transmission from parent to child. The higher the repeat number, the more unstable it is. The mechanism driving the instability is still not well understood. Separately from this, repeats above a certain size cause disease. Repeats that are big enough to be unstable but not big enough to cause disease are called **premutations**. People carrying premutations are healthy but are at high risk of having affected children. The main examples are shown below.

Box table 4.2 – Diseases associated with pathogenic expanded nucleotide repeats

Disease	OMIM No.	Mode of inheritance	Location of gene	Location of repeat	Repeat sequence	Normal range (repeats)	Pathological range (repeats)
Huntington disease	143100	AD	4p16	Exon 1	(CAG) _n	5–35	37–120
DRPLA	125370	AD	12p13	Exon 5	(CAG) _n	7–34	58–88
SCA1	164400	AD	6p23	Exon 8	(CAG) _n	19–38	40–81
SCA2	183090	AD	12q24	Exon 1	(CAG) _n	15–29	35–59
SCA3 (Machado–Joseph disease)	109150	AD	14q24–q31	Exon 10	(CAG) _n	14–40	68–82
SCA6	183086	AD	19p13	Exon 49	(CAG) _n	6–17	21–30
SCA7	164500	AD	3p21–p12	Exon 3	(CAG) _n	7–17	38–130
SCA17	607136	AD	6q27	Exon 3	(CAG) _n	25–44	50–55
SBMA	313200	XLR	Xq11	Exon 1	(CAG) _n	11–33	38–62
Fragile X site A (FRAXA)	309550	XL	Xq27.3	5'UTR	(CGG) _n	6–45	200–>1000
Fragile X site E (FRAXE)	309548	XL	Xq28	Promoter	(CCG) _n	6–25	>200
Friedreich ataxia (FRDA)	229300	AR	9q13–q21.1	Intron 1	(GAA) _n	7–22	200–1700
Myotonic dystrophy 1 (DM1)	160900	AD	19q13	3'UTR	(CTG) _n	5–35	50–4000
Myotonic dystrophy 2 (DM2)	602668	AD	3q21	Intron 1	(CCTG) _n	12	75–11 000
SCA8	603680	AD	13q21	Untrans. RNA	(CTG) _n	16–37	110 – >500
SCA10	603516	AD	22q13	Intron 9	(ATTCT) _n	10–22	Up to 22 kb
SCA12	604326	AD	5q31	Promoter	(CAG) _n	9–18	66–78
FTD/ALS	105550	AD	9p21	Intron 1	(GGGGCC) _n	2–23	250–4000
Progressive myoclonic epilepsy (PME)	254800	AR	21q22.3	Promoter	(CCCCGCCGCCG) _n	2–3	40–80

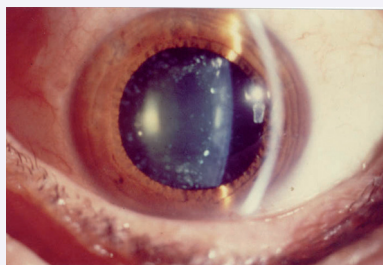
This is not an exhaustive list; around 40 such diseases are known, with new ones regularly added to the list.

DRPLA, dentatorubral pallidoluysian atrophy; FTD/ALS, frontotemporal dementia and/or amyotrophic lateral sclerosis; SBMA, spinobulbar muscular atrophy; SCA, spino-cerebellar ataxia.

A feature of many of these diseases is **anticipation**. This describes the way a disease may get more severe going down the generations. It happens because the expanded repeats are very unstable and tend to expand still further on transmission. The size of expansion is often correlated with the severity of symptoms and/or with onset at younger ages. Thus an expanding repeat is suspected to lie at the root of any genetic disease that shows anticipation. However, reports of anticipation must be viewed with considerable skepticism. If a dominant disease is naturally very variable, as many are, then clinicians will often see severely affected children born to a mildly affected parent. Mildly affected children born to a severely affected parent will be less frequently seen, because severely affected people may not have children, and if they do, they may not feel there is anything wrong with a mildly affected child. Thus a common bias of ascertainment often mimics anticipation.

The first nine diseases in *Box table 4.2* all involve expanding (CAG)_n runs within the coding sequence of a gene. CAG is the codon for glutamine (see *Table 6.1*) so the effect is to encode a protein with an expanded polyglutamine tract. These proteins are in some way toxic to neurons. The cumulative death of neurons leads to a late onset neurodegenerative disease. For most of the other diseases in the table, the expanded repeat prevents expression of a gene, and the disease is the result of lack of the gene product. The two forms of myotonic dystrophy, however, are the result of the toxicity of mRNA containing the expanded repeat. Toxic RNA is also involved in the Fragile X tremor and ataxia syndrome, FXTAS (see *Section 4.3*) and may also be part of the molecular pathology of SCA8 and perhaps other diseases. Thus although all these diseases are caused by expanded repeats, the mechanisms by which they cause disease are quite diverse and for the most part, poorly understood. La Spada and Taylor (2010) give some details; see LaCroix *et al.* (2019) for an intriguing recent example.

Clinically, the polyglutamine diseases are all late onset neurodegenerative conditions. The precise symptoms probably depend on the pattern of expression of the mutant gene and of genes whose products interact with the mutant protein. Friedreich ataxia is also a result of progressive death of neurons, with die-back affecting cerebellar function. FRAXA was described in **Case 11 (Lipton family)**; FRAXE is very similar. The gene products are RNA-binding proteins that assist with the transport and translation of selected mRNAs. Finally, the two forms of myotonic dystrophy are multisystem diseases with muscle myotonia, cataracts, testicular atrophy and frontal balding. Myotonic dystrophy shows especially striking anticipation.



(a)



(b)



(c)

Box figure 4.3 – Anticipation in myotonic dystrophy.

(a) A 'blue-dot' cataract may be the only sign of the disease in the first affected generation. (b) A three generation family showing the grandmother who has bilateral cataracts but no muscle symptoms or facial weakness; her daughter has moderate facial weakness with ptosis and cataracts; the child has the congenital form. Reproduced from *Myotonic Dystrophy* by Peter Harper (Saunders, 3rd edn., 2001) with permission. (c) A baby with the congenital form showing hypotonia. The congenital form is seen only when the child inherits the disease from its mother. It is caused by very large expansions of the CTG repeat, which are never found in sperm.

4.5. References

General descriptions of the techniques covered here can be found in many textbooks, for example, Strachan T and Read AP (2019) *Human Molecular Genetics* 5th edition, CRC Press.

Allingham–Hawkins D, Chitayat D, Cirigliano V, et al. (2011) Prospective validation of quantitative fluorescent polymerase chain reaction for rapid detection of common aneuploidies. *Genetics in Medicine*, **13**: 140–147.

Girirajan S, Rosenfeld JA, Cooper GM, et al. (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**: 203–209.

LaCroix AJ, Stabley D, Sahraoui R, et al. (2019) GGC repeat expansion and exon 1 methylation of *XYLT1* is a common pathogenic variant in Baratela–Scott syndrome. *Am. J. Hum. Genet.* **104**: 35–44.

La Spada AR and Taylor JP (2010) Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* **11**: 247–258.

Mefford HC, Sharp AJ, Baker C, et al. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *New Engl. J. Med.* **359**: 1685–1699.

Tassone F, Pan R, Amiri K, et al. (2008) A rapid polymerase chain reaction-based screening method for identification of all expanded alleles of the Fragile X (*FMR1*) gene in newborn and high-risk populations. *J. Mol. Diag.* **10**: 43–49.

Wilhelm M, Schlegl J, Hahne H, et al. (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**: 582–587.

Useful websites

The Eurogems site (www.eurogems.org) provides links to websites containing material relevant to this chapter at a number of levels.

The Cold Spring Harbor DNA Learning Center has useful resources on PCR and other laboratory techniques: <https://dnlc.cshl.edu/resources/animations>

The Access Excellence Resource Center graphics gallery has graphics of Southern blotting and PCR: www.accessexcellence.org/RC/VL/GG/index.html

For much detailed information on FISH see: www.ihcworld.com/in-situ-hybridization.htm

4.6. Self-assessment questions

- (1) Design 10-nucleotide-long primers to amplify each of the 50 bp sequences underlined so as to produce a 50 bp product.

(a)

CCACTCCCCTCGGCCAGGGCCGCGTCAACCAGCTCGGCGGTGTTTTTATCAACGGCAGGTACCAGG
AGACTGGCTCCATACGTCCTGGTGCCATCGGCGGCAGCAAGCCCAAGGTGAGCGGGCGGGCCTTGC

(b)

AAGAGAGAACCCGGGCGTGCCGTCAGGTACTAGGCCCATTAACCTCTCCCCGCTTCCTTCCTCCTC
CCGCCCCCAGTGAGTTCCATCAGCCGCATCCTGAGAAGTAAATTCTGGGAAAGGTGAAGAGGAGGAG

[Note that real primers would be 16–25 nucleotides long and would be designed to give a product of a few hundred bp; this is an exercise in getting the position and orientation of your primers correct. Do it by hand even if you have access to a primer design program.]

- (2) Extend *Box table 4.1* to show the progress of the PCR reaction up to cycle 10. (It would be neat to make a spreadsheet to do this). How many cycles would it take to produce 100 000 copies of Product B?
- (3) There are 4 different nucleotides in DNA, 16 different dinucleotides, 64 different trinucleotides and 4^n different sequences n nucleotides long. If a restriction endonuclease cuts DNA whenever it encounters a particular 5 nucleotide sequence, and assuming these occur at random throughout the human genome, into how many fragments might it cleave the DNA of a human cell?
- (4) Assuming the human genome consisted entirely of unique sequence DNA, how long would an oligonucleotide probe need to be in order to hybridize to just one sequence in the genome?
- (5) For each of the following sequence changes or effects, choose possible testing methods from the list below that could be used to check for the presence of the change or effect (more than one method may be appropriate for some cases).
 - a G>A change in exon 2 of the *PAX3* gene that results in replacement of valine 60 by methionine in the gene product.
 - a heterozygous 3 bp deletion in exon 6 of the *BRCA1* gene.
 - an A>T nucleotide substitution that changes the codon for arginine 214 (AGA) into a stop codon (TGA) in exon 7 of the *MITF* gene.
 - pathogenic expansion of the (GGGGCC)_n chromosome 9 repeat in a patient with frontotemporal dementia.
 - a GT>GA change in the donor splice site at the end of exon 4 of the *PAH* gene, which encodes the liver enzyme phenylalanine hydroxylase.
 - a C>A change in an intron near a splice site in the ubiquitously expressed actin gene: the question is whether or not it affects splicing of the primary transcript.
 - deletion of several contiguous genes on one copy of chromosome 17 in a child with suspected Smith–Magenis syndrome.
 - a duplication of one or more exons in the dystrophin gene in a boy with Duchenne muscular dystrophy.
 - deletion of one or more exons of the *HYP* gene in a boy with hypophosphatemia (an X-linked dominant condition).
 - insertion of three nucleotides in the promoter of a gene – the question is whether this affects expression of the gene.
 - any material extra or missing on a copy of chromosome 7 in a patient – the cytogeneticist reported that the banding pattern on one copy of chromosome 7 was abnormal but could not work out exactly what events had produced the change.

Options:

- (a) PCR amplification, check for the presence / absence of product
- (b) PCR amplification, check the size of the product
- (c) PCR amplification followed by sequencing (see *Chapter 5* for detail)
- (d) PCR amplification followed by hybridization to an allele-specific oligonucleotide (see *Chapter 5* for detail)
- (e) RT-PCR
- (f) Real-time quantitative PCR
- (g) Southern blotting
- (h) FISH
- (i) Chromosome painting
- (j) Array-CGH
- (k) MLPA.

[Hints on questions 1a and 4 are provided in the *Guidance* section at the back of the book.]

05

How can we check a patient's DNA for gene mutations?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe changes to genomic DNA, cDNA and protein using correct nomenclature
- Describe the principles of DNA sequencing (Sanger and 'next-generation') and read a straightforward Sanger sequencer trace
- Describe the applications of next-generation sequencing in a clinical genetics service
- Describe the circumstances in which a DNA test involves checking for a specific change, scanning a gene for variants, sequencing a panel of candidate genes or sequencing a patient's whole exome or genome
- Describe briefly the principles of two methods by which a person's DNA can be checked for a specified change

5.1. Case studies

CASE 13 NICOLAIDES FAMILY

- Spiros and Elena both carriers of β -thalassaemia

117 129 159 316 395

Spiros Nicolaides is an IT graduate who was born in the UK but whose family originally came from Cyprus. He is healthy, and on a recent trip to see his grandparents in Cyprus met and fell in love with Elena, who recently returned there from the USA where she was studying. Both of the families are delighted and a big engagement party is planned. However, Elena's older sister tells her that before she married she was advised to have some blood tests to see if she was a carrier for β -thalassaemia. Luckily, although Elena's sister was shown to be a carrier, her future husband was not and they now have a healthy son. Elena and Spiros request an appointment in the genetic clinic to discuss their risks.

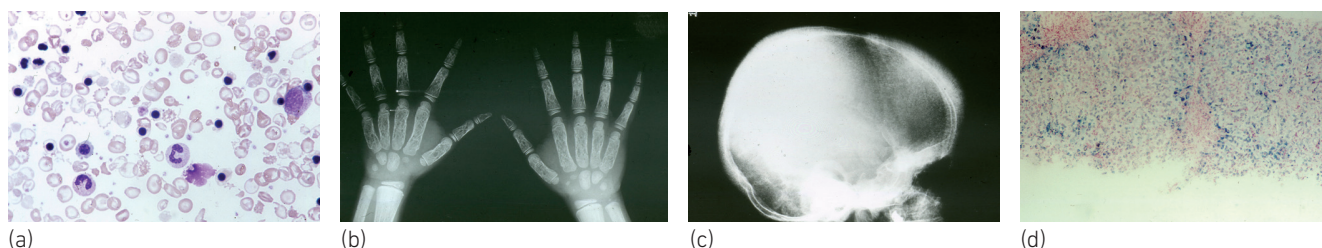


Figure 5.1 – Effects of thalassemia.

(a) Blood film with very marked hypochromia and many nucleated red cells. (b) Osteoporotic appearance of hands due to bone marrow extension. (c) 'Hair on end' skull. (d) Liver biopsy with Perl's stain, showing iron overload. Courtesy of Dr Andrew Will, Royal Manchester Children's Hospital.

5.2. Science toolkit

In the previous chapter we saw how PCR or various hybridization methods can be used to examine any chosen small section of the DNA in a sample. Sometimes no further investigation is needed. That was the case for several of the families:

- in **Case 1 (Ashton family)**, the size of the PCR product identified the pathogenic Huntington disease allele.
- for **Case 4 (Davies family)**, MLPA identified and characterized a partial deletion of the dystrophin gene.
- for **Case 11 (Lipton family)**, triplet-primed PCR identified the family members having premutations and full mutations of the *FMR1* gene. Southern blotting defined the exact size of the full expansions.

But in many diseases the pathogenic variants are substitutions of one nucleotide for another somewhere within the DNA of a gene, which would not generate any visible difference in a PCR product or any of the other assays discussed so far. Equally, insertion or deletion of one or two nucleotides would not make a noticeable difference to the size of a PCR product. We need therefore to consider how to check a gene for such point mutations.

The available methods fall into three classes.

- Methods that check for a specific sequence change. These have many applications in diagnosis and screening for specified mutations.
- Methods that rapidly scan a specific gene for any change, without identifying the nature of the change. In past times these helped reduce the need for sequencing. Now that sequencing is so cheap these once important methods are falling into disuse, so they are only briefly described below.
- DNA sequencing – the falling cost and increasing capability of DNA sequencers means that sequencing is now the default approach to many questions.

Methods for detecting specific sequence changes

It often happens that the laboratory is looking for the presence or absence of one specific small sequence variant in a sample. This happens under various circumstances.

- The disease in question may always be caused by exactly the same sequence change. The reasons why some diseases are always caused by one specific variant are discussed in *Chapter 9*.
- A disease may be caused by various different variants, but one or a few may be so frequent in a particular population that it is worth first checking for these before going on to a more general search. An example is cystic fibrosis. Over 1700 different pathogenic variants have been reported, but 80% of all changes in Northern Europeans are one particular deletion of 3 nucleotides (see *Table 12.1*). Beta-thalassemia is another example: 80% of all mutations in Greek Cypriots are c.93–21G>A (see *Table 5.3*); the nomenclature is explained in *Box 5.1*.
- The test may be to check somebody for a family variant that has already been identified and characterized in other family members.
- The test may be to check samples from healthy controls to make sure a variant found in a patient is not a non-pathogenic variant present in the normal

population. The availability of exome or genome sequences of hundreds of thousands of ostensibly healthy individuals in databases such as GnomAD (see *References*) has made this much less of a problem than in past years, but it can still be necessary if the patient comes from an ethnic group that is not well represented in public databases.

The same techniques are also used for genotyping people for non-pathogenic single nucleotide polymorphisms (SNPs), for example, in linkage analysis (*Chapter 8*) and association studies (*Chapter 13*).

Many methods (in addition to sequencing) are available to detect a specific sequence change, including the following.

- Hybridizing the PCR-amplified sample to an allele-specific oligonucleotide (ASO) probe. As mentioned in *Chapter 4*, under suitable conditions even a single mismatched nucleotide can prevent a short (16–25 nt) probe from hybridizing. Used on a massive scale, this is the basis of SNP chips, but single ASOs also have many uses.
- Performing an allele-specific PCR reaction. PCR primers may work even if there are slight mismatches to their target, provided the primer is long enough to hybridize, but the 3'-end nucleotide of the primer is critical. Unless it correctly base-pairs to the template the reaction will not work. We can therefore check whether a given nucleotide is A or G by setting up a PCR reaction where this nucleotide must pair with the 3'-end nucleotide of one of the primers. A primer ending in T will amplify only the A allele, while one ending in C will amplify only the G allele. Two parallel reactions are run with the two different primers, or alternatively the primers are differently labeled (e.g. with different dyes) so as to give separately identifiable products in a single mixed reaction. *Figure 5.2* shows how allele-specific PCR can be used to detect the sickle cell mutation; *Figures 5.9* and *5.12* show other applications.
- Digestion with a restriction enzyme. If the variant happens to either create or destroy a sequence that is the recognition site for a restriction enzyme, this allows an easy test. The relevant region is PCR amplified, the product is digested



Figure 5.2 – Detecting the sickle cell mutation by allele-specific PCR.

A PCR reaction will fail if the 3'-end nucleotide of a primer is not correctly base-paired. Slight mismatches elsewhere in the primer sequence may be tolerated, but the 3' end must match. One primer (not shown, off to the left) is standard; the second primer is specific for one allele of the c.20A>T (p.Glu6Val) change in the β -globin gene that causes sickle cell disease (green). Two reactions are set up, one using the common plus A-specific primers, the other using the common plus T-specific primers. See *Figures 5.9* and *5.12* for more examples.

A brief guide to nomenclature of variants

A variant can be described in terms of the change in the genomic DNA, the cDNA, or the encoded protein. The description is prefixed with g., c., or p. according to which of these it describes.

For DNA > means 'changes to'. Thus G>A means that G in the Reference Sequence is replaced by an A nucleotide. Deletions and insertions are symbolized by del and ins respectively. The affected nucleotide is numbered from an agreed starting point and database file. For cDNA, nucleotides are counted from the A of the AUG start codon. A nucleotide in an intron is given the number of the last nucleotide of the preceding exon, a plus sign and the position in the intron, for example, c.77+1, c.77+32, etc. If it is near the end of a large intron it may be given the number of the first nucleotide of the following exon, a minus sign and the position (*Tables 5.3 and 11.1* show examples). For genomic DNA, the start nucleotide position must be specified.

For protein changes the one-letter or three-letter amino acid abbreviations are used (see *Box 3.6*). X, Ter or * means a stop codon. Amino acids are counted from the initiator methionine of the protein (even though this is usually removed in post-translational processing).

Examples:

c.76A>C means that at nucleotide 76 of a cDNA the normal A is replaced by C

c.76_78del means a deletion of 3 nucleotides, from nucleotides 76 to 78 inclusive

p.Ala26Val or p.A26V means that amino acid alanine-26 is changed to a valine.

p.Cys318Ter or p.C318* means the codon for cysteine 318 is changed to a stop codon.

p.Arg123LysfsTer34 shows the effect on the protein of a frameshift: the first changed amino acid is arginine 123, changed to lysine, and the new reading frame encounters a stop codon 34 amino acids later.

This level of detail is sufficient for understanding everything in this book. Nomenclature has been defined for describing every possible sequence variation. You can find full details at the Human Genome Variation Society website, <http://varnomen.hgvs.org/>

Alongside this descriptive nomenclature there are some less formal ways of naming variants. Sometimes a name refers to an established list of alleles. This system is particularly used for variants relevant to pharmacogenetics. Thus in *Chapter 10* we will meet *TPMT*2* and *HLA-B*1502*. Sometimes variants are labeled by the place they were first described. This is widely used for hemoglobin variants, for example, Hemoglobin Beirut, Hemoglobin Lepore (see OMIM for full lists). Sometimes the same system is used with other proteins, for example, Factor V Leiden.

with the appropriate restriction enzyme and then run out on a gel. The size of the fragments shows whether or not the enzyme was able to cut the DNA. Heterozygotes show both the cut and uncut fragments. Examples are illustrated below for β -thalassemia (**Case 13 – Nicolaides family**, *Figure 5.8*) and LHON (**Case 6 – Frank Fletcher**, *Figure 5.10*).

Methods for scanning a gene for any sequence change

A diagnostic laboratory often needs to check every exon of a candidate gene in a patient to look for variants. Given the average sizes of exons and introns (145 bp and 3365 bp, respectively, see *Chapter 3*) this usually meant PCR amplifying and sequencing each exon individually. In the past the cost of doing this routinely for a gene with many exons such

as *CFTR* (**Case 2, Brown family**: 27 exons) or *COL2A1* (**Case 10, O'Reilly family**: 54 exons) could be prohibitive. Various methods were developed to save sequencing costs by scanning each exon quickly and cheaply to eliminate those that apparently contained no variants. This approach is seldom used nowadays, except for scanning a gene to check for deletions or duplications of whole exons, as was done for **Case 4 (Davies family)**. Sequencing costs have fallen, next-generation sequencing is routinely used to check every exon of one or more genes, and RNA analysis (see *Section 4.4*) offers an alternative way of checking multi-exon genes. Because they are no longer widely used, the methods for scanning for point mutations are only briefly described here; earlier editions of this book give more detail. The commonest methods were based on either of two principles.

- *Properties of heteroduplexes.* If a person is heterozygous for a sequence variant in an exon, PCR amplification of the exon will give a mix of the variant and normal sequences. If the mix is heated to denature the DNA and then slowly cooled, some of the resulting double helices will be **heteroduplexes**, containing one strand from each of the two alleles. Heteroduplexes denature more easily than fully matched duplexes. This can be noted in various ways, for example, by their altered mobility on a denaturing high performance liquid chromatography (dHPLC) column. An alternative technique, melting curve analysis, follows the denaturation by monitoring the fluorescence of a dye such as SYBR Green, which fluoresces strongly in the presence of double-stranded, but not single-stranded DNA.
- *Properties of single-stranded DNA.* In conditions that favor hybridization, single-stranded DNA ties itself into knots, as bases in different parts of the strand pair with each other. The precise shape of the knot depends on the sequence, and this affects the rate at which the knot migrates through an electrophoretic gel. Single strand conformation polymorphism (SSCP) analysis looks for any differences between the migration of the test sequence and a normal sequence. Compared to dHPLC or melting curve analysis, SSCP requires no expensive equipment and is simple to set up, but probably misses more variants (claimed sensitivity is 80–90%).

These methods all have in common that they report (with variable sensitivity) whether or not the sequence of an exon appears to differ from the Reference Sequence, but they do not identify any specific change. That requires the exon to be sequenced.

DNA sequencing – the ultimate test

Sequencing DNA – that is, determining the specific sequence of A, C, G and T nucleotides in a DNA molecule – first became practicable with the publication by Fred Sanger of his dideoxy technique in 1977. For almost 30 years Sanger's method remained unchallenged as the universal way to sequence DNA. The Human Genome Project was achieved by massive, industrial-scale application of Sanger sequencing. Over the years incremental technical progress had made the process more user-friendly – radiolabeling was replaced by fluorescence labeling, and machines like the one illustrated in *Figure 5.4* allowed semi-automated sequencing of 96 samples in parallel – but the essential principle remained the same.

Starting in 2005, this all changed. In quick succession several independent commercial companies announced revolutionary new sequencing technologies. Different competing

companies pioneered different technologies, but what they all have in common is that they are *massively parallel*. Thousands or millions of sequencing reactions are run in parallel, allowing a dramatically increased throughput. Collectively called next generation sequencing (NGS), these new technologies have changed the face of molecular genetics. They generate previously unimaginable amounts of sequence data for quite modest costs per nucleotide. This makes all sorts of sequencing-intensive applications possible that would previously have been unthinkable. Sequencing whole genomes is now routine. The Human Genome Project took 15 years to produce the Human Reference Sequence, and it cost \$3 billion. The latest NGS machines can sequence an entire human genome in hours for \$1000.

Despite the revolution, Sanger sequencing is not dead. As discussed below, NGS faces challenges in terms of its accuracy and the difficulty of selecting the targets for sequencing. Thus for targeted applications where high throughput is not important – typically sequencing a single PCR product, a single exon or a single small gene – Sanger sequencing is still the technique of choice. But NGS uniquely makes it possible to sequence large panels of genes, whole exomes or whole genomes. Thus each technology has its place in clinical laboratories. In this section we will first describe Sanger sequencing, and then discuss NGS, focusing on capabilities, applications and potential rather than on the details of the competing technologies. Readers who would like to know more should consult a review. Chapter 6 of Strachan and Read's *Human Molecular Genetics* (2019) describes the various technologies in considerable detail, as does the review by Goodwin *et al.* (2016). Because of the rapid pace of developments, readers should consult the most recent review available for details.

Sanger (dideoxy) sequencing

Rather like PCR, dideoxy sequencing uses a DNA polymerase to make many copies of the fragment of interest (an approach called *sequencing by synthesis*). The starting material is a collection of identical DNA molecules – usually a PCR product, or sometimes a cloned copy of the fragment of interest. Unlike in PCR, for sequencing we want to make copies of just one of the DNA strands. Therefore a single primer is used, as in *Figure 4.8*. Starting with the single-stranded target DNA, we add the primer, the four nucleotide monomers A, G, C and T, and a DNA polymerase enzyme to synthesize the complementary strand. However, the pool of monomers is spiked with chain-terminating molecules. This is the key to the sequencing technique. The idea was developed by Fred Sanger, and earned him a share of the 1980 Nobel prize for chemistry (his second! – he had already won the 1958 Nobel prize for determining amino acid sequences of proteins).

The chain terminating molecules are modified versions of the standard A, G, C and T nucleotides (*Figure 5.3*). They are incorporated into a growing polynucleotide chain just like their regular counterparts, but they then prevent the chain growing any further. The chemical trick that does this is to remove the hydroxyl group on position 3' of the deoxyribose sugar: chain terminators are *dideoxy* nucleotides. Since this hydroxyl group provides the link to the next nucleotide in the chain, its absence means that no new nucleotide can be added to that particular chain.

To make the whole thing work, we need to add just the right amount of chain terminators. When it is incorporating an A into the growing chain the polymerase will randomly pick a normal or a dideoxy molecule of the A nucleotide. If there is 1% as much dideoxy-A as

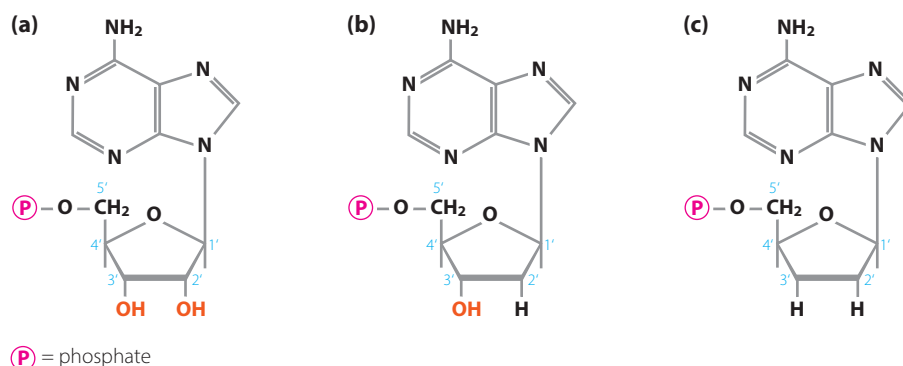


Figure 5.3 – Formulae of (a) a ribonucleotide (b) a deoxyribonucleotide (c) a dideoxynucleotide.

The dideoxynucleotide can be added to a growing DNA chain through its 5' phosphate, but because it lacks a 3' hydroxyl group, no further growth of the chain is then possible. In sequencing reactions the four dideoxynucleotides are labeled with four different fluorescent dyes.

normal A, then at each A position around 1% of the growing chains will incorporate the chain terminator and grow no further. As a result, a series of nested fragments accumulate, each terminated by a dideoxy-A. Putting this into a concrete example, suppose we use a 20-nucleotide primer and the DNA we are copying has T at positions 27, 30, 35, 41, etc. nucleotides from the 5' end of the primer. Opposite each T in the template strand, the polymerase will incorporate an A in the growing strand. About 1% of growing strands will terminate at each of these positions by incorporation of dideoxy-A, if this is present as 1% of the pool of A monomers. Thus there will be fragments of length 27, 30, 35, 41, etc. nucleotides. When the product is size-separated by electrophoresis, if we can read off the lengths of the fragments, we can read off the position of each A in the newly synthesized strand (or each T in the template strand). If dideoxy-A was the only dideoxy nucleotide used, there would not be fragments 28, 29, 31, 32, etc. nucleotides long, because at those positions the polymerase would not be incorporating an A into the growing chain, and so there would be no risk of a dideoxy-A preventing the chain from growing further.

To generate the full sequence, the reaction is spiked with terminator versions of all four nucleotides (*Figure 5.4a*). The four dideoxynucleotides are tagged with four different colored fluorescent dyes; the standard nucleotides are unlabeled. Each fragment that ends with an A will be (say) green, each fragment ending in C will be blue, and so on. A DNA sequencing machine separates the fragments by length using electrophoresis through a fine capillary, and a laser reads the color of each fragment as it emerges from the capillary. The result is displayed as a series of colored peaks (*Figure 5.4b*). Sophisticated software interprets the sequence and can provide other information, such as quantitating each peak and defining the length of each fragment. These features make automated sequencers useful for other tasks as well as sequencing. Increasingly, PCR-based tests are formatted to use fluorescently labeled primers, so that the PCR product can be sized and quantitated on a sequencing machine, rather than run on a manual electrophoretic gel.

A 3-D animated video of the process can be accessed at www.yourgenome.org/video/dna-sequencing.

Compared to NGS, Sanger sequencing has much lower throughput, but offers three compensating advantages.

- First, it is highly accurate. All NGS technologies have much higher rates of random errors. To counter this they need to sequence the same stretch of DNA many times so that the random errors cancel out. Many clinical laboratories still prefer to rely on Sanger sequencing to confirm critical variants detected by NGS.
- Second, Sanger sequencing will deliver around 600–800 nt of accurate sequence from a single run. Most NGS technologies are *short-read*, delivering no more than 100–200 nt of sequence per fragment (but see below). This has implications for the accuracy with which sequences can be identified, as discussed below.
- Finally, Sanger sequencing is *targeted*. By the choice of starting material and primer, a scientist can specify precisely which segment of DNA is to be sequenced. NGS technologies all bulk-sequence whatever DNA molecules are loaded into the machine.

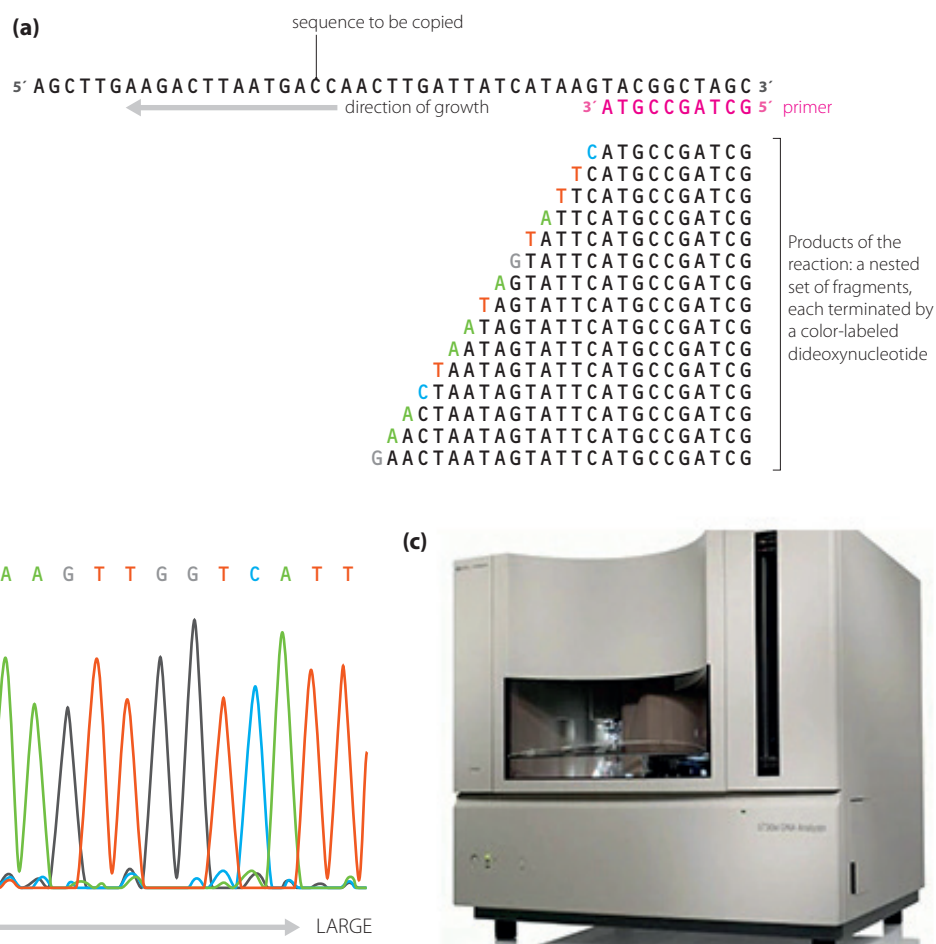


Figure 5.4 – Principle of DNA sequencing.

(a) Each time a nucleotide is incorporated into the growing chain, there is a small chance that it will be a dideoxynucleotide, which will terminate growth of that chain. The result is a series of nested fragments. (b) Electrophoresis separates the fragments by length. Each dideoxynucleotide carries a different colored fluorescent dye. (c) Sequencing machines such as this ABI 3730 genetic analyzer run up to 96 samples simultaneously and automate the separation, detection, sizing and quantitation of the fragments. Image courtesy of Thermo Fisher Scientific.

Next generation sequencing (NGS)

There is not just one NGS technology. A dozen or more companies are developing competing technologies, and some of their products are in routine use in laboratories around the world. The key novelty shared by all the NGS technologies is that they are *massively parallel*. That is, they simultaneously perform millions of independent sequencing reactions. Depending on the technology, these may be on millions of tiny beads, in individual cells of a microarray, in individual droplets of an emulsion, or in individual pores in a membrane. Compared to Sanger sequencing, they generate a staggering amount of data per run at a far lower cost per base sequenced. However, they lack a key feature of the Sanger method: they are not targeted. In Sanger sequencing the choice of primer determines the sequence to be read. The massively parallel machines simply sequence everything in the test sample. Unless the aim is to sequence a whole genome, it is necessary to first select a subset of a patient's DNA for sequencing. This might be just a single gene, exons of a panel of genes of interest, or all exons of all genes (the whole exome). To do this, genomic DNA from the patient is fragmented, usually by sonication, denatured and exposed to a cocktail of probes that, between them, represent all the sequences it is desired to check. Hybridizing fragments are isolated (typically using biotin-labeled probes and streptavidin-coated magnetic beads) and used for sequencing.

The input material for an NGS run is a vast collection of millions of DNA fragments produced by random fragmentation of the original sample. Unlike with a restriction enzyme digest (Box 4.2) the fragmentation is random, so that identical DNA molecules from different cells will give different fragments. Thus a given nucleotide in the DNA will be present on overlapping fragments of different sizes. Assuming the initial sample consisted of DNA from thousands of cells, any given sequence will be represented by thousands of independent fragments. The raw output from the sequencing run is a collection of millions of short 'reads' (stretches of sequence), representing all the different fragments in the input material. A video showing what is involved in preparing a DNA sample for sequencing, and then sequencing it on one of the most widely used NGS machines, can be accessed at www.yourgenome.org/sites/default/files/downloads/video/life-in-the-lab-a-dna-sequencing-pipeline/life-in-the-lab-dna-pipelines.mp4.

The first step in converting the millions of individual reads into useful information is to align them to a reference sequence. For clinical and other human genetic purposes this is currently the GRCh38 Human Reference Sequence. Powerful computer programs attempt to match each fragmentary sequence to a unique position in the Reference Sequence. Here we see the importance of both **read length** and **read depth**. Figure 5.5 shows a simplified example and illustrates some of the problems.

All NGS technologies have a much higher frequency of sequencing errors compared to Sanger sequencing. When a read matches the Reference Sequence except for a single nucleotide (positions marked 'b' or 'c' in the figure), this might be a true variant or it might be a sequencing error. Read depth is important in deciding which alternative is true. If there are only two reads, one of which has G and the other A at a certain position, it is impossible to know whether this is a true variation or a sequencing error. If there are 50 reads, 49 Gs and 1 A, the A is likely an error. If the 50 reads have 29 Gs and 21 As the person is probably heterozygous for the two variants. Read depth is ultimately determined by the amount of input material loaded into the machine (which in turn depends on how much one is willing to spend on a run), but it is not uniform across the genome – some

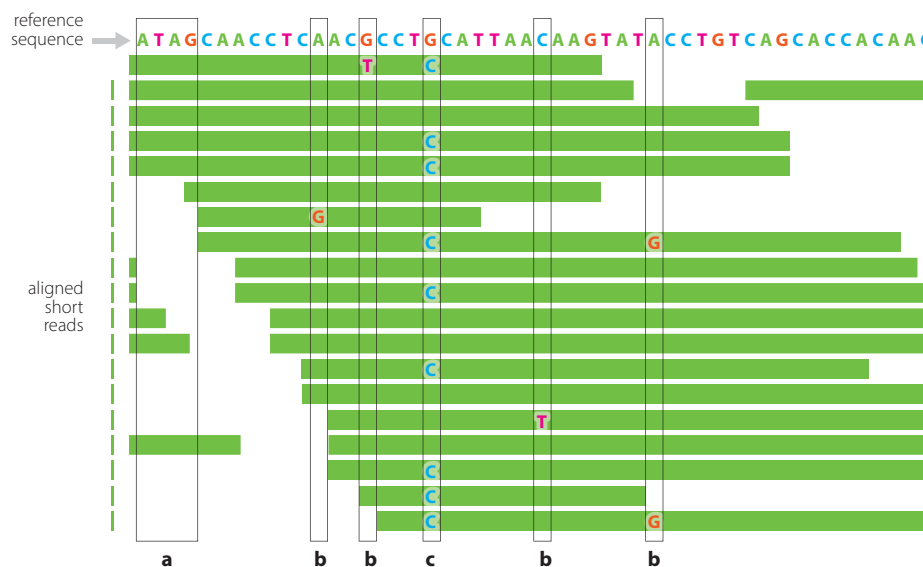


Figure 5.5 – Aligning NGS reads to the reference genome.

Each horizontal green bar represents an individual sequencing read. Nucleotides that do not match the reference are shown; otherwise all positions match. (a) A region with poor coverage. (b) The variants in these tracks are most likely sequencing errors. (c) At this position the subject is most likely heterozygous G/C. A real example would have much greater read depth.

sequences are more efficiently captured, amplified or sequenced than others, so a high average read depth is needed to ensure acceptable minimum cover of all sequences. Figure 5.6 shows an example. For clinical purposes a 80X average read depth is commonly recommended.

The length of a read affects how easily it can be aligned to the reference sequence. Short reads are less likely to find a single unambiguous match. The various NGS technologies can be categorized into *short read* and *long read*. All the original versions, and most current implementations in routine clinical use, are short read. Intense competition between companies is driving rapid technical development, in particular aimed at increasing the average read length, but for the most part it remains in the 100–200 nt range, far below the 600–800 nt sequenced in a typical Sanger run. The many regions of the genome

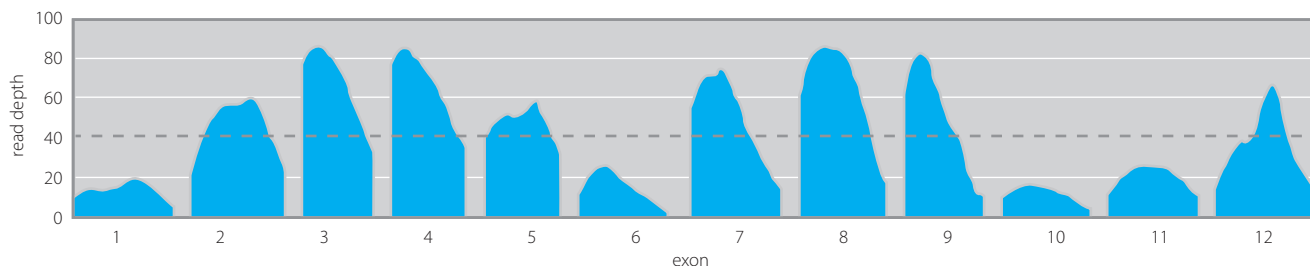


Figure 5.6 – Read depth is not constant.

The *PCSK9* gene has been sequenced to an average read depth of 40X (dotted line), but the actual depth varies across the 12 exons, and within each exon. Data from exac.broadinstitute.org.

with complex repetitive structures cannot be sequenced by short read NGS because the reads cannot be unambiguously aligned. Alternative long read technologies are offered or being developed by companies such as PacBio and Oxford Nanopore. These can produce single reads of tens of kilobases, making alignments much more certain and allowing complex repetitive regions to be resolved. Many people predict that, once costs come down and accuracy improves, long read technologies will come to dominate the field, but at present cost and/or accuracy limit their widespread adoption.

Paired-end sequencing offers an extra level of information. If, say, 50 nucleotides from each end of a longer fragment of known size are sequenced, this can help with the alignment. For example, one end may map to a repetitive region, but could be placed correctly by reference to the other end. Alignments can also reveal insertions, deletions or inversions in the test DNA if the ends, which are a known distance apart in the test DNA, match sequences that are a different distance apart, or in a different orientation, in the reference (*Figure 5.7*).

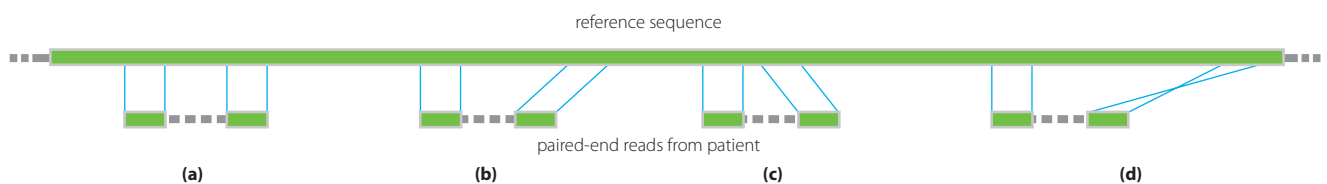


Figure 5.7 – Paired-end sequencing can be used to identify structural variants.

(a) Sequence from each end of a 2 kb fragment aligns as expected to the Reference Sequence. (b) Sequences that are 2 kb apart in the patient's DNA are 3 kb apart in the reference genome. The patient must have a 1 kb deletion somewhere in the region between the paired ends. (c) Sequences that are 2 kb apart in the patient's DNA are only 1 kb apart in the reference genome. The patient must have a 1 kb insertion somewhere in the region between the paired ends. (d) Sequences that are in the same orientation in the patient's DNA are in opposite orientations in the reference genome. The patient must have an inversion with one breakpoint in the region between the paired ends.

After reads have been aligned, a variant calling program produces a list of variants – positions at which the test DNA differs from the Reference Sequence. Mostly these will be single nucleotide variants. Quality control programs monitor the alignments and remove reads with insufficient coverage or excessive numbers of mismatches (which are probably incorrectly aligned). Fragments that truly contain an insertion or deletion of one or a few nucleotides in the test DNA compared to the reference also risk failing quality control because they align poorly. Thus NGS analyses tend to underestimate such variants in a patient. Different variant calling programs do not always produce the same list of variants, and there are problems with variants that look like artefacts but are real. Thus transforming the raw data from any NGS machine into useable clinical sequence requires substantial computing capacity and extensive bioinformatic skills – the programs cannot simply be left to perform automatically.

After all the processing and quality control, a typical NGS whole genome would reveal:

- 3–3.5 million single nucleotide variants compared to the Reference Sequence (including approximately 20 000 variants in coding sequences)
- 500 000 small insertions or deletions.

Filtering such a list to identify the single pathogenic variant in a patient with a mendelian condition can be a daunting challenge, requiring both substantial bioinformatic skills and clinical insight. We will consider several examples in this and later chapters.

5.3. Investigations of patients

The stories so far...

Table 5.1 summarizes the achievements and requirements for testing in the cases described so far. This section describes the further tests performed on cases 2, 3, 6, 10 and 13, which illustrate the various techniques described above. The six cases are discussed in the order of the tests used: first Case 13, analyzed entirely using tests for specific sequence changes; next Case 6, where the mitochondrial genome was checked for specific mutations. Then Cases 2 and 10, where a candidate gene was searched for mutations. Finally, Case 3 used massively parallel sequencing of the whole exome, because there was no candidate gene.

Table 5.1 – Summary of testing performed to date in the cases featured, and further tests required

Case	Problem	Tests so far	Tests needed
1.	Huntington disease	Size of expansion checked by PCR	None needed
2.	Cystic fibrosis		Check <i>CFTR</i> gene for mutations
3.	Intellectual disability	SNP chip scan for pathogenic copy number variants – none found	Sequence whole exome to look for pathogenic change
4.	Duchenne muscular dystrophy	Deletion of exons 44–48 defined by MLPA, family members tested	None needed
5.	Chromosomal translocation	Identified by karyotyping and array-CGH	None needed
6.	Leber hereditary optic neuropathy		Check for mitochondrial mutation, homoplasmic or heteroplasmic
7.	? Deletion of 22q11	Deletion confirmed by FISH	None needed
8.	Down syndrome	Confirmed by karyotyping	Prenatal test (see Chapter 12)
9.	Turner syndrome	Confirmed by karyotyping Check for Y sequences by PCR	None needed
10.	Stickler syndrome		Check <i>COL2A1</i> gene for mutations
11.	Fragile X	Check size of trinucleotide repeat by PCR and Southern blotting	None needed
12.	? Chromosomal abnormality	Check for chromosomal imbalances with SNP chip; check if it is <i>de novo</i>	No further routine investigation
13.	β-thalassemia		Identify mutations

CASE 13 NICOLAIDES FAMILY

- Spiros and Elena both carriers of β -thalassemia
- Need to define mutations for prenatal diagnosis
- Allele-specific PCR shows Spiros carries the p.Gln39X variant
- Restriction digest shows Elena carries the c.316–106C>G variant

117

129

159

316

395

Hemoglobinopathies are the most intensively studied of all genetic diseases, and rightly so, because they affect millions of people in many countries. Children homozygous for β -thalassemia have a difficult life. They require endless blood transfusions, but then have major problems with iron overload. Hopefully gene therapy will soon come to the rescue – results from trials are promising. Among Greek Cypriots, it has been estimated that one person in seven is a carrier of β -thalassemia. Thus we would expect one in 49 marriages to be between carriers. This has triggered a national population screening program. The reasons why hemoglobinopathies are so remarkably common in some populations are discussed in *Chapter 9*. Carrier testing is done using routine hematological methods. This indicated that both Spiros and Elena were carriers, though entirely healthy in themselves. They were offered molecular tests to identify their precise mutations. Because they thought they would opt for prenatal diagnosis in any pregnancy, they took up the offer. It was better to do it now, when there was no urgency, than to wait until Elena was pregnant and would require urgent testing.

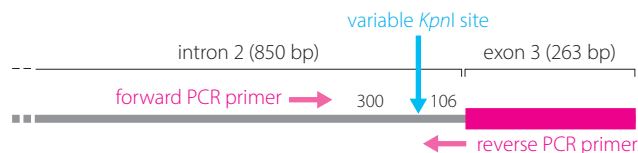
Spiros and Elena each provided a mouthwash sample from which DNA was extracted. Five specific variants in the β -globin gene account for 98.4% of all β -thalassemia alleles among Greek Cypriots (*Table 5.2*). Their DNA was therefore first tested for each of these variants. Had the results been negative, their β -globin genes would have been sequenced. The β -globin gene is small (3 exons and 2 introns, only about 1500 bp in total), so it is straightforward to sequence the exons, introns and promoter to check for variants.

Neither of them carried the most common Cypriot variant, c.93–21G>A, so the DNA was then checked for the other four common causes. Spiros turned out to be carrying the p.Gln39X variant while Elena carried c.316–106C>G. All these could have been identified by sequencing, allele-specific PCR or by hybridization to allele-specific oligonucleotides; the c.316–106C>G variant also creates a restriction site for the enzymes *RsaI* (GTAC) and *KpnI* (GGTACC). *Figures 5.8* and *5.9* illustrate the use of different methods for detecting these variants.

Table 5.2 – Common β -thalassemia alleles among Greek Cypriots

Variant	Location	Percentage of all β -thalassemia alleles in Greek Cypriots	Sequence change: (normal, mutant)
c.93–21G>A	Intron 1	79.8	ctatt g gtctatattttccc ctatt a gtctatattttccc
c.92+6T>C	Intron 1	5.5	AGgttgg t at AGgttgg c at
c.92+1G>A	Intron 1	5.1	AG g ttggtat AG a ttggtat
c.316–106C>G	Intron 2	5.1	cag c taccat cag g taccat
p.Gln39X	Exon 2	2.9	TGGACCC CAG AGGTTC TGGACCT AG AGGTTC

Exon sequences are in upper case, intron sequences in lower case. See *Box 5.1* for the nomenclature of variants and *Chapter 6* for discussion of why these variants cause disease. Data from HbVar database: <http://globin.cse.psu.edu/globin/hbvar>.



	<i>KpnI</i> site	undigested PCR product	<i>KpnI</i> -digested PCR product
normal sequence	absent	406 nt	406 nt
c. 316–106 C>G	present	406 nt	300 + 106 nt

Figure 5.8 – Identification of the c.316–106C>G allele.

The variant creates a *KpnI* restriction site (GGTACC) in intron 2 of the β -globin gene. A suitable size fragment including the variant site is PCR-amplified and the product incubated with the restriction enzyme. The fragment sizes in the digested DNA are measured by electrophoresis on a manual gel or a gene analyzer.



Figure 5.9 – Identification of the p.Gln39X allele by allele-specific PCR.

See Figure 5.2 for the principle. The C>T change converts a codon for glutamine (CAG) into a stop codon (TAG), as explained in Chapter 6. The C-specific and T-specific primers are used with a common primer that hybridizes to a sequence to the left of the region shown. A deliberate mismatch at position 3 of each primer increases the specificity of the reaction. Note that the primer sequences are written in the 3'–5' direction.

CASE 6 FLETCHER FAMILY

- Frank, aged 22, with increasingly blurred vision
- Family history of visual problems
- Possible mitochondrial inheritance
- ? Leber hereditary optic neuropathy
- Test mitochondrial genome
- m.G3460A mutation identified

5 13 69 130 157 395

LHON (OMIM 535000) is the result of inadequate function of the mitochondria, usually caused by variants in the mitochondrial DNA (mtDNA). Confirming the diagnosis in Frank depends on demonstrating the presence of one of the specific variants associated with LHON. The molecular genetics of LHON is quite complicated. Eighteen different single nucleotide variants in the mtDNA have been associated with this one disease. Presumably there are many ways in which mitochondrial function can be impaired, leading to the disease. Five of the 18 have a sufficiently serious effect to cause LHON by themselves; the others are found in combination with each other and presumably cause disease by an accumulation of smaller effects.

Three single nucleotide variants cause the great majority of cases, at least in people of European origin (we use the usual mtDNA nomenclature here).

- m.G11778A – substitution of A for G at nucleotide 11778 of the 16.5 kb mitochondrial genome. This results in replacement of arginine 340 by histidine in the ND4 protein that is part of the oxidative phosphorylation machinery.
- m.G3460A – the G→A nucleotide change replaces alanine 52 in the ND1 protein by tyrosine.
- m.T14484C – this nucleotide substitution causes methionine 64 in the ND6 protein to be replaced by valine.

The usual diagnostic procedure is first to test for these three specific variants. If none is present, then a wider search is needed, which may include sequencing parts of the mtDNA. Any of the methods described previously could be used for checking for these three variants. *Figure 5.10* shows a restriction enzyme-based method for m.G11778A. The result showed that Frank Fletcher did not carry the G11778A variant.

	SEQUENCE			FRAGMENT SIZES	
				<i>Sfa</i> NI	<i>Mae</i> III
normal	CGAACGCACT	CACAGTC <u>G</u> CA	TCATAATCCT	417 + 91	233 + 218 + 57
m.G11778A	CGAACGCACT	CACAGTC <u>A</u> CA	TCATAATCCT	508	233 + 131 + 87 + 57

Figure 5.10 – Detecting the G11778A allele of mitochondrial DNA by its effect on restriction enzyme recognition sites.

The variant abolishes a site for the enzyme *Sfa*NI (GCATC) but creates one for the enzyme *Mae*III (GTNAC, where N is any nucleotide). The assay shown here involves PCR-amplifying a 508 bp portion of the mtDNA that includes nucleotide 11778. Separate samples of the PCR product are digested with the two enzymes and the fragments sized by gel electrophoresis. Restriction sites are underlined, the changed nucleotide is highlighted in color and the sizes of fragments are shown.

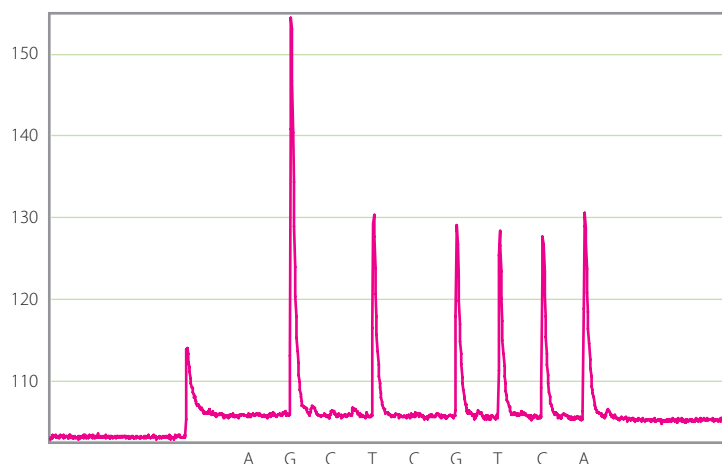


Figure 5.11 – Detection of the G3460A mtDNA variant by pyrosequencing.

The pyrosequencing machine tries adding each nucleotide in turn to the end of a primer. Successful addition triggers a bioluminescent reaction, recorded by the machine as a peak. Upper trace: variant sequence GTGTCA; lower trace: normal control GCGTCA (the sequences are of the reverse strand).

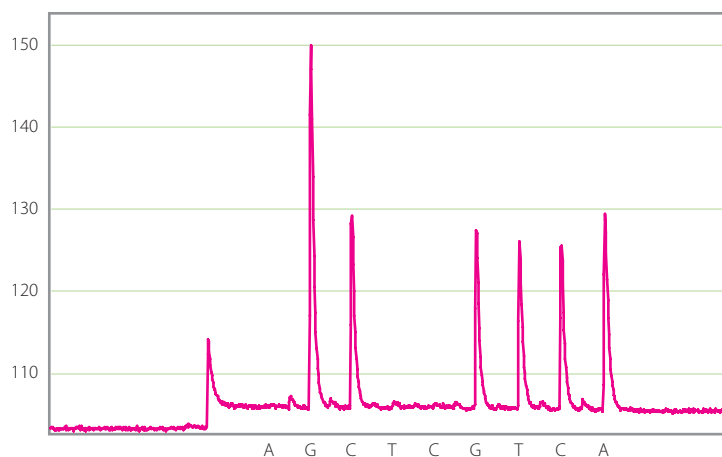


Figure 5.11 shows the use of an alternative technique, pyrosequencing, to check for m.G3460. Pyrosequencing is a special method of sequencing by synthesis, not widely used, but it has the advantage of producing a quantitative result. As explained in *Chapter 1*, people can be homoplasmic or heteroplasmic for mitochondrial mutations. Pyrosequencing of a short stretch of the mitochondrial genome around position 3460 would give a readout of the relative proportion of mitochondria carrying G or A at this position. The process is described by Ronaghi *et al.* (1996). Briefly, a primer is extended using specially modified monomers that produce a flash of light when they are incorporated into the growing chain. The pyrosequencing machine presents each monomer in turn to the polymerase. Unincorporated monomer is then degraded before the next monomer is presented. The result showed that Frank was homoplasmic for the m.G3460A variant, thus confirming the diagnosis of LHON.

CASE 2 BROWN FAMILY

- Baby Joanne, recurrent infections, poor growth
- Sweat test confirms she has cystic fibrosis
- Autosomal recessive inheritance
- Need for molecular test
- *CFTR* variants identified

2 10 67 **132** 154 313 395

There were two reasons for wishing to identify the precise pathogenic variants in Joanne's *CFTR* genes. Some new drugs show promise in ameliorating some of the symptoms of the disease, but they act only against certain specific variants, so it would be necessary to check Joanne's two *CFTR* genes to discover whether or not she might be able to benefit. Secondly, David and Pauline indicated that they might consider having more children in the future but would definitely wish for prenatal tests because they felt that coping with the extra needs of one child with cystic fibrosis was as much as they could manage. Before that could be done, it would be necessary to define both variants in Joanne. Additionally, after discussion at family reunions, other members of the large extended family became concerned about their own carrier risk and several relatives expressed a desire for carrier testing. Again, this has to be done by checking for the variants present in Joanne. Although over 1700 different variants in the *CFTR* gene have been described in cystic fibrosis patients, most of these have been seen in only one or a very few cases. A small number of pathogenic changes are relatively frequent. Testing in CF therefore usually starts by checking for these common variants before proceeding to full sequencing if necessary. Several commercial kits are available for this purpose.

Joanne's DNA was screened using a multiplex allele-specific PCR assay. The result (Figure 5.12) shows that she has one copy of the most frequent cystic fibrosis allele in Northern Europeans, a deletion of three nucleotides, c.1521_1523delCTT that results in the protein lacking phenylalanine 508 (p.F508del). This variant is often called by the non-standard name delta-F508. It accounts for 70–80% of all CF alleles in many Northern European populations. Sanger sequencing was used to look for her second variant. This identified the following changes (in addition to p.F508del):

Exon 3 c.236G>A

Exon 16 c.2620–15C>G (Figure 5.13)

In *Chapter 6* we will see how the laboratory attempted to decide whether either of these changes was likely to be pathogenic.

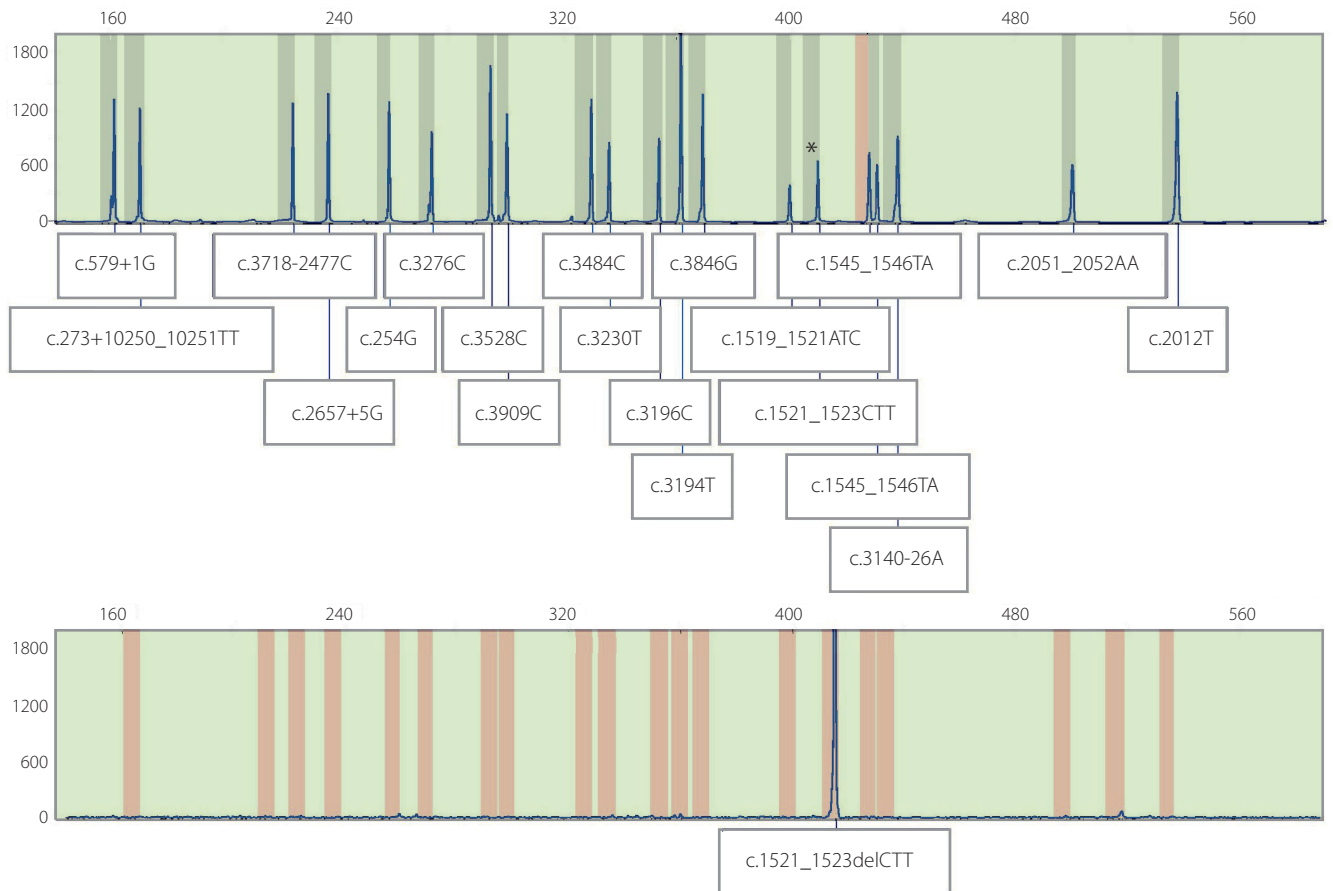


Figure 5.12 – A multiplex allele-specific PCR test for 36 common CFTR alleles.

Joanne's sample was amplified in two reactions, each using a cocktail of primers specific for particular *CFTR* alleles. The products were separated on a gene analyzer. The upper trace shows part of the results with primers that detect the normal counterpart of each variant. The lower trace shows the corresponding variant alleles. The result showed both the normal (upper trace, asterisk) and variant (lower trace) alleles for the p.F508del variant (here given its proper name, c.1521_1523delCTT). Thus Joanne is heterozygous for this variant, but must have a second variant that is not one of the 36 targeted by this kit. Data using Devyser Core kit, courtesy of Simon Ramsden and Jenny Henchcliffe, St Mary's Hospital, Manchester.

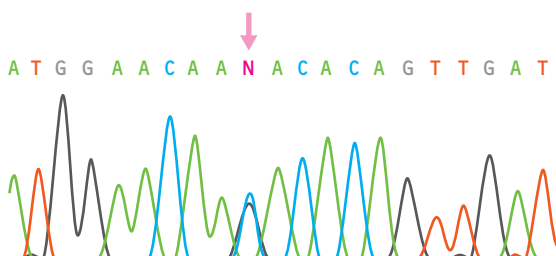


Figure 5.13 – DNA sequencer trace of part of the exon 16 PCR product from Joanne Brown's CFTR gene.

At the arrowed position G and C nucleotides are both present, showing that Joanne is heterozygous for a nucleotide substitution (remember that the products of PCR and sequencing are normally a mix of the products from the two alleles). Control samples show only the G. It is usual to sequence both strands of the DNA separately to confirm any change. In this case the sequence shown is of the reverse strand.

CASE 10 O'REILLY FAMILY

- Orla has severe myopia, short stature and hip problems
- Family history of similar problems
- ? Stickler syndrome
- Test collagen II genes
- Sequencing identifies *COL2A1* variant

57

70

134

158

395

Orla's combination of high myopia and joint problems, inherited in an autosomal dominant manner, is characteristic of people who have a mutation in Type II (or occasionally Type XI) collagen. As described in *Box 3.4*, there are at least 27 different human collagens, encoded by at least 30 genes. To recapitulate, each collagen gene encodes a polypeptide, procollagen, which is subject to extensive post-translational modification. The final processed collagen molecule contains tightly wound triple helices of polypeptide chains. Some are homotrimers, some are heterotrimers. Collagen II is a homotrimer of polypeptides encoded by the *COL2A1* gene on chromosome 12q13.

The *COL2A1* gene has 54 exons. Orla's mutation might be anywhere within exons 8–49. These exons encode the triple helical domain of Type II collagen; changes outside this region affect parts of the procollagen that are cleaved off during post-translational processing, as described in *Box 3.4* and are not associated with Orla's phenotype. In Orla's case, sequencing revealed a deletion of 2 nucleotides, c.2488_2489del, in exon 40. In *Chapter 6* we will consider the effect such a deletion would have on the collagen II molecule.

CASE 3 KOWALSKI FAMILY

- Karol, first son of Kamil and Klaudia
- Developmental delay, hypotonic, severe intellectual disability
- Difficulties of genetic testing in such cases
- Likely need for exome sequencing
- Negative SNP chip test for microdeletions
- Exome sequencing

3

10

67

102

134

155

395

As described in *Chapter 4*, DNA from the boy Karol, who has severe intellectual disability, was tested for copy number variants on a microarray, but no clearly pathogenic variant was found. It is likely that he has a point mutation in some or other essential gene – but mutations in any of hundreds, maybe thousands of genes could cause intellectual disability. Before the availability of NGS there was no way of investigating further. But with NGS his whole exome could be sequenced, offering a fair chance of identifying the cause of his problem. We describe here a real case from Manchester, but with the names changed.

Karol's DNA was fragmented by sonication, then standard adapter oligonucleotides were ligated to each end of the fragments to allow PCR amplification of the whole library using a single pair of primers. The library of prepared fragments was hybridized to an exon capture kit of oligonucleotides (see above). Non-hybridizing fragments were washed away, and the library was sequenced on a next generation machine to an average read depth of 80x across the exome, resulting in coverage of 94% of the reference exome at 20x depth or greater.

After alignment to the Reference Human Genome, removing reads that failed quality control, and running a variant caller program, the result was a list of 16 400 variants. The first step in reducing the list to a manageable number of candidates for more detailed consideration was to eliminate common variants. Karol's problem is rare, so the cause must be a rare variant. Before deciding how rare a variant must be to allow it to remain in the analysis, it was necessary to form a judgement about the likely cause of Karol's problem: was it a recessive or a dominant condition? If it was dominant it must be the result of a *de novo* mutation, since neither parent was affected, nor was there any family history of any similar condition. The relevant mutant allele must be correspondingly rare – certainly occurring in well under 1% of the population. If Karol's condition is recessive a rather less stringent upper limit on the frequency would be appropriate, since carriers of rare recessive conditions are not so very infrequent in the population – see *Chapter 9*

for calculations. Either possibility could be correct, and the analysis could be conducted using the two alternative hypotheses, but experience suggests that in countries like the UK where consanguineous marriages are infrequent, the great majority of such cases are caused by new dominant mutations.

In the early days of NGS there was only limited information about the frequencies of variants in any population, making it uncertain how many of an extensive list of variants could be disregarded as being too common. As more and more genomes have been sequenced, the quality of information has rapidly increased. Two important public databases, ExAC and GnomAD are invaluable resources. Published exome or genome sequences from a variety of sources were collected and subjected to a standardized reanalysis (important because, as mentioned above, different quality control procedures and variant calling programs can produce somewhat different lists of variants). The data was mostly from apparently healthy control individuals from various clinical trials or the genomewide association studies described in *Chapter 13*. Some caution would be necessary if using this data to eliminate variants that might contribute to late-onset diseases, but we can be confident that none of the subjects had severe pediatric conditions like Karol Kowalski's. The ExAC database (exac.broadinstitute.org) comprised 60 706 exomes; in its successor, the GnomAD database (<https://gnomad.broadinstitute.org>), this was expanded to 125 748 exomes and 15 708 complete genomes. The only limitation on this extremely powerful resource is that the majority of subjects are white and of European ancestry. Efforts are underway to correct this bias.

Filtering Karol's list of variants against these databases reduced the list to 410 variants. Further filtering depended on considering the likely effect of each variant on the gene involved, and is discussed in *Chapter 6*.

5.4. Going deeper...

The three questions

The methodology of mutation testing depends crucially on the precision of the question being asked. Consider three possible questions:

- (1) Does **Joanne Brown (Case 2)** have the p.F508del variant in the *CFTR* gene?
- (2) Does Joanne Brown have *any* variant in the *CFTR* gene?
- (3) Does Joanne Brown have *any* variant in *any* gene that would account for her condition?

Question (1) is quick and cheap to answer, using any of the methods described in the section on detecting specific sequence changes. That section also describes the circumstances under which it is possible to ask such a specific question. Question (2) could be answered by sequencing each exon of the candidate gene. This would most likely be done by Sanger sequencing, although if a gene has a large number of exons it might be easier to use a next generation method. Question (3) was unanswerable before the advent of massively parallel sequencing. Our new-found ability to answer this question has led to the rapid identification of the genes and variants responsible for each of a large number of rare conditions previously characterized only by clinical description.

In a service context there are three possible ways of harnessing the power of the new technology.

- **Sequencing a panel of candidate genes.** Conditions such as deafness, blindness or congenital heart defects are often caused by pathogenic variants in a single gene, but in each case the gene responsible could be any one of 100 or more candidates. A common diagnostic approach is to draw up a list of candidate genes and then create a panel of custom oligonucleotides matching each exon of each of those genes. These are used as hybridization probes to capture all the desired exons from a patient's genomic DNA. Non-hybridizing sequences are washed away and those remaining are sequenced. So, for example, in Manchester a panel of 180 genes is used in investigation of retinal degeneration, and a different panel of 115 genes is used for congenital cataract. Compared to whole exome or whole genome sequencing, this reduces the amount, and hence the cost, of sequencing. It also focuses attention on plausible candidates, thus reducing the amount of irrelevant and unwanted information generated. The disadvantage is that much effort is needed to develop and validate a suitable panel, and the panel needs periodic updating to accommodate new gene discoveries. Additionally, it means that laboratory protocols are split across several different panels, making workflow and quality control more complex.
- **Whole exome sequencing.** Here every exon of every protein-coding gene is sequenced (a total of around 180 000). As before, the desired sequences are selected from the patient's genomic DNA by hybridization. Various companies sell exome capture kits that also include varying amounts of flanking intron and some additional functional non-coding sequences. Typically these would capture 30–60 Mb of sequence, roughly 1–2% of the total genome. Sequencing 20 000 genes rather than the 100 or so on a targeted panel clearly costs more, but there are advantages to the process. Using a commercial kit saves the laboratory the effort of developing and updating its own disease-specific gene panels, and makes for a more unified workflow. However, a patient's exome would typically include around 20 000 variants compared to the reference human genome, and filtering these to home in on those that are relevant to the patient's condition is a substantial task. Additionally, when every gene is sequenced there is a chance of uncovering information that is possibly important but unrelated to the condition for which the patient was referred. A child referred for hearing loss might be coincidentally discovered to have a pathogenic variant in a cancer-causing gene. This may or may not be a welcome finding. It raises considerable questions about consent and ethics. How to handle these so-called 'incidental findings' has generated much discussion and soul-searching (see *Section 12.4*).

One widely favored strategy is to use a *virtual gene panel*. That is, the whole exome is sequenced, but only a limited set of genes is analyzed, at least in the first place. In one initiative, expert committees convened by Genomics England have defined a large number of virtual gene panels that are used in the analysis of cases from the UK 100 000 Genomes Project. *Table 5.3* shows some examples.

Table 5.3 – Examples of virtual gene panels used in the UK 100 000 Genomes Project.

Condition	No. of genes considered
Kabuki syndrome	4
Familial hematuria	8
Congenital hypothyroidism	27
Anophthalmia or microphthalmia	56
Primary ciliary disorders	138
Epileptic encephalopathy	182
Congenital hearing impairment	356
Primary immunodeficiency syndromes	388
Intellectual disability	1997

Data from <https://panelapp.genomicsengland.co.uk/panels/>.

- Whole genome sequencing.** As sequencing costs continue to fall, it becomes feasible to think of sequencing a patient's whole genome as a diagnostic investigation. This would allow changes outside coding exons to be identified. At present that is a questionable benefit because in the current state of knowledge the vast majority of small changes in non-coding sequence are uninterpretable. But a major advantage is that structural variants (whose breakpoints are usually in non-coding DNA) can be identified. This is of particular importance in oncology when tumor genomes are being analyzed (*Chapter 7*). A second advantage of whole genome sequencing is that it avoids exon capture. A study at Nijmegen University in the Netherlands (Gillisen *et al.*, 2014) illustrated the value of that. Whole genome sequencing was performed on 50 trios (patients with unexplained severe intellectual disability and their unaffected parents). All had previously been investigated for copy number variations using microarrays, and had had their whole exomes sequenced, without any causative variant being identified. Whole genome sequencing allowed conclusive diagnoses to be reached in 20 of the 50 cases. In each case the causative variant identified was not in the non-coding DNA, but a *de novo* coding sequence variant or small *de novo* copy number variant that had been missed before (despite the extensive experience of this group) because of inefficient exome capture or the inadequate resolution of microarray analyses, or had been wrongly excluded by the quality control procedures. Hopefully with better technology and more experience, these problems are reducing.

Where's it all going?

Undoubtedly the time is soon coming when many healthy people will carry around with them an electronic record of their own genome sequence. How far this will simply serve personal curiosity or vanity, and how far it will serve any useful clinical purpose, is a wide open question – and one to which clinical geneticists would dearly love an answer.

Sudden arrhythmic death syndrome

Genetic conditions that predispose to fatal heart arrhythmias can be the cause of devastating family tragedies. It has long been recognized that sudden death can occur in apparently healthy young people without there being any clue on post-mortem examination as to why the tragedy happened. Occasionally more than one person in a family is affected. Here we describe how a family might come to attention and be investigated. The story starts with a referral letter.

Letter from family doctor to Genetics Clinic:

"Please see this young woman aged 22 years whose brother sadly died suddenly and unexpectedly last year. The family have been encouraging her to be checked out before her marriage later this year and she is concerned whether there might be any risks for any children she might have."

Information needed at/before the clinic appointment:

- Family pedigree, particularly noting any other sudden deaths in the family, even if they were thought to be caused by epilepsy, drowning, etc.
- Circumstances of death
- Post-mortem findings and whether any tissue is available for further testing

Consider the main differential diagnoses:

- Suicide
- Substance abuse
- Epilepsy
- Cardiac event

Establishing a diagnosis:

If the first three causes have been excluded or are thought unlikely, then the main diagnosis to consider is **sudden cardiac death**. In older people this is most likely due to coronary artery disease but in young people where no structural heart problem has been identified then **sudden arrhythmic death syndrome (SADS)** is most likely. Any medical records should be scrutinised and if tissue is available genetic tests can be performed.

The commonest cardiac **channelopathies**, known as arrhythmia syndromes, are:

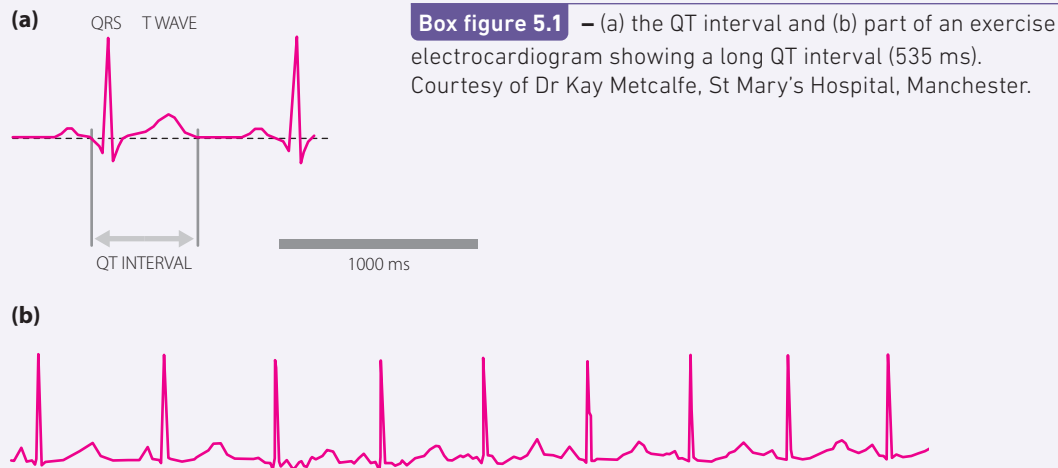
- long QT syndromes (most affect potassium channels that control the flow of potassium ions from the inside to the outside of cells)
- Brugada syndrome, which affects sodium channels regulating flow of sodium ions into the cells
- catecholaminergic polymorphic ventricular tachycardia (CPVT), which affects calcium regulation in the cell.

Cardiomyopathies are a group of inherited diseases of the heart muscle that usually give symptoms such as breathlessness but can cause SADS. There may be preceding symptoms and, in some individuals, obvious hypertrophy at post-mortem. However, some individuals may have no symptoms, and any histological signs are detectable only by experts in the field. The main conditions are:

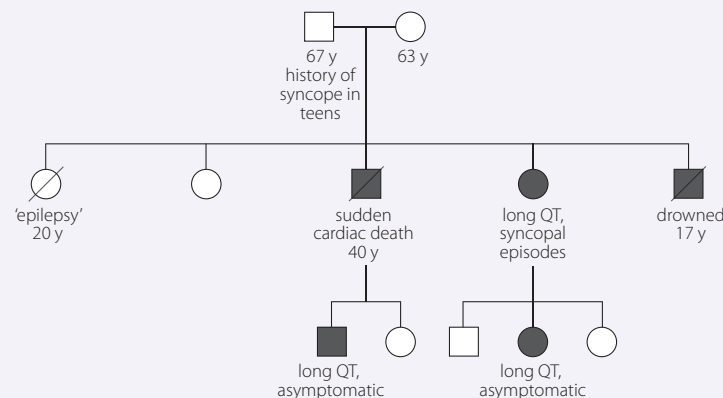
- Hypertrophic cardiomyopathy (HCM)
- Dilated cardiomyopathy (DCM)
- Arrhythmogenic right ventricular cardiomyopathy (ARVC)

Long QT syndromes are characterized by ECG abnormalities (*Box figure 5.1*). These consist of QT prolongation (indicating prolonged or disordered ventricular repolarization) and T-wave abnormalities associated with a tendency to tachycardia (rapid beating of the heart) which may cause syncope (fainting). Twitching occurring during the syncope may lead to the misdiagnosis of epilepsy. Often these episodes self-terminate, but they can progress to ventricular fibrillation

where the heart ceases to pump blood effectively and sudden death occurs. In 10–20% of all cases of drowning, long QT or other genetic pro-arrhythmic conditions may be the cause (see Choi *et al.*, 2004). Long QT syndromes may also explain some cases of sudden infant death syndrome, but at most account for only 10%.



Families have been described where the condition is compatible with dominant inheritance, but others are reported where affected individuals have sensorineural deafness and autosomal recessive inheritance. The dominant types (*Box figure 5.2*) are collectively called Romano–Ward syndrome (OMIM 192500) and the recessive types Jervell and Lange–Nielsen syndrome (JLN; OMIM 220400). JLN-affected individuals have profound sensorineural hearing loss in addition to a long QT interval. Half of those untreated die before the age of 15 years.



Box figure 5.2 – A typical pedigree of a family in which dominant long QT syndrome is segregating.

Linkage studies revealed that both Romano–Ward and JLN syndromes are heterogeneous, with different loci implicated in different families. Mutations have been found that implicate the following ion channel genes:

- *KCNQ1* and *KCNE1* – these encode components of the IKs potassium channel. Romano–Ward patients can be heterozygous and JLN patients homozygous for mutations in either gene. Syncope and sudden death are particularly triggered by physical exertion, especially swimming.

- *KCNH2* and *KCNE2* – these encode components of the IKr potassium channel. Heterozygous mutations in either gene have been described in some Romano–Ward patients. Emotional excitement or loud noise (especially noises that wake somebody up) are the commonest triggers of sudden death.
- *SCN5A* – encodes a cardiac sodium channel. Heterozygous mutations have been described in some Romano–Ward patients. Death most commonly occurs during quiet rest or sleep. Mutations in this gene have also been described in some patients with Brugada syndrome, where the ECG abnormalities are different.

Up to 30% of families do not have a detectable mutation in any of these genes. It is highly desirable to identify a causative mutation, so that at-risk relatives can be clearly identified and those not at risk reassured. Treatment for all the long QT syndrome disorders is directed at avoiding known precipitating factors, reducing the tendency to tachycardia by the use of beta-blocker medication and, for some, the use of implantable cardioverter defibrillators. Knowledge of the specific mutation may enable anticipatory guidance about avoiding specific precipitating factors.

5.5. References

- Choi G, Kopplin LJ, Tester DJ, et al.** (2004) Spectrum and frequency of cardiac channel defects in swimming-triggered arrhythmia syndromes. *Circulation*, **110**: 2119–2124.
- Goodwin S, McPherson JD and McCombie WR** (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**: 333–351.
- Gilissen C, Hehir-Kwa JY, Thung DT, et al.** (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature*, **511**: 344–347.
- Katsanis SH and Katsanis N** (2013) Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.*, **14**: 415–426.
- Ronaghi M, Karamohamed S, Pettersson B, et al.** (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**: 84–89.
- Soothill PW and Lo YMD** (2014) Non-invasive prenatal testing for chromosomal abnormality using maternal plasma DNA. *Scientific Impact Paper 15*. Royal College of Obstetricians & Gynaecologists, London.
- Strachan T and Read AP** (2019) *Human Molecular Genetics*, 5th edition, CRC Press – see Chapter 6 for details of sequencing technologies.

Useful websites

Gene panels: <https://panelapp.genomicsengland.co.uk/panels/>

GnomAD: <https://gnomad.broadinstitute.org>

www.yourgenome.org – this site, from the Wellcome Trust public education team, has animations and videos explaining various techniques and interviews with laboratory workers using them.

5.6. Self-assessment questions

- (1) For each of the following sequence changes or effects, choose possible testing methods from the list below that could be used to check for the presence of the change or its effect (more than one method will be appropriate in many cases).
- Seeking a point mutation in any of the 27 exons of the *CFTR* gene.
 - Checking for duplication of exons 50–54 of the dystrophin gene in a boy with Duchenne muscular dystrophy.
 - Checking for expansion of the (CAG)*n* repeat in the *SCA3* gene in a woman suspected of having Machado–Joseph disease.
 - Identifying a small additional ‘marker’ chromosome seen in a dysmorphic baby.
 - Checking for insertion of 4 nucleotides in exon 7 of the dystrophin gene in the sister of a boy known to have this mutation.
 - Seeking the cause of profound hearing loss in a newborn baby.
 - Checking for any balanced chromosome structural variant in the normal parents of a baby who died with multiple congenital abnormalities.
 - Identifying a deletion of any of the exons of the dystrophin gene in a boy with Duchenne muscular dystrophy.
 - Checking for a G>C change (TGGAATTGCAGCAG > TGGAATTCCAGCAG) in exon 4 of the *CFTR* gene in the cousin of a child who had cystic fibrosis with this variant.
 - Looking for a second mutation in a child with cystic fibrosis where conventional screening has identified only a single pathogenic variant.
 - Checking for a heterozygous 1.5 Mb deletion at 7q11.23 in a child suspected of having Williams–Beuren syndrome.
 - Checking whether a variant in the promoter of a gene has any functional effect.
 - Characterizing a full expansion in the *FMR1* gene in a young woman from a Fragile-X family.
 - Identifying acquired changes in the genome of cells from a patient’s tumor.

Possible methods:

- (a) G-banded karyotyping
- (b) FISH
- (c) Array-CGH
- (d) Multiplex ligation-dependent probe amplification (MLPA)
- (e) PCR, check for presence / absence of product
- (f) PCR, note size of product
- (g) PCR, check for restriction digestion of product
- (h) PCR, Sanger sequence product
- (i) Allele-specific PCR
- (j) Reverse transcriptase–PCR (RT–PCR)
- (k) Hybridization to allele-specific oligonucleotide (ASO)
- (l) SSCP (single-strand conformation polymorphism)
- (m) Southern blotting
- (n) Whole exome sequencing
- (o) Whole genome sequencing

(2) For each of the following tests, note whether it depends on a property of single-stranded or double-stranded DNA:

- Denaturing high-performance liquid chromatography (dHPLC)
- Array-CGH
- Checking for heteroduplexes by gel mobility
- PCR
- MLPA
- Checking for creation / abolition of a restriction site.

(3) The restriction enzyme *EcoRI* cuts the sequence GAATTC. Part of the coding sequence of a certain gene reads as follows:

CAA	AAC	CTC	AAG	TCA	ACG	AGT	TCG	GTA	ACG	TAC
Gln	Asn	Leu	Lys	Ser	Thr	Ser	Ser	Val	Thr	Tyr

This part of the gene is PCR amplified from DNA of a patient with a disease that is often caused by mutations in this gene. It is noted that the PCR product, which in normal people is not cut by *EcoRI*, is now cut into two fragments. Assuming the mutation changes a single nucleotide in the segment shown, identify it.

[Hints on questions 2 and 3 are provided in the *Guidance* section at the back of the book.]

06

What do mutations do?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe and identify silent, mis-sense, nonsense and frameshift changes in coding sequences, and splice site mutations
- Use a table of the genetic code to define the effect on the gene product of a change within a coding sequence
- Explain and give examples of loss of function, gain of function, haploinsufficiency, dominant negative effects, nonsense-mediated decay, dosage sensitivity
- Discuss how far genotype-phenotype correlations can be established, and reasons why they are often imprecise
- List ways in which a change in non-coding DNA may affect expression of a gene

6.1. Case studies

CASE 14 JENKINS FAMILY

- James Jenkins, achondroplasia diagnosed in infancy
- No family history
- Father was 58 years old when James conceived
- James's wife Joanne also has achondroplasia
- Obstetric problems and risks to children

143

160

395

28 year old James Jenkins was a young man well known to the genetics department. His parents, Jessica and John had attended soon after the birth of James. They had two normal healthy children but in this pregnancy they had been worried about the risk of Down syndrome because Jessica was then 39 years old (John was 58, but they understood that paternal age was not a major risk factor for Down syndrome). They were reassured when an amniocentesis had revealed no abnormality. James was noticed to have a large head at birth, but it was only when he had his regular check at 6 weeks that his head control was noted to be poor, and the doctor commented that his limbs seemed short. He was referred to a pediatrician who arranged for a skeletal survey.

The survey revealed the typical radiological features of achondroplasia. This was a shock to his parents, both of whom were healthy and of normal stature. They wanted to find out more about the condition, and also to meet other families with children with the same condition, so even though they didn't plan to have any more children themselves they asked to be referred to the genetic clinic.

In the genetic clinic the doctor and counselor told Jessica and John about the guidelines that were being developed to help families care for a child with achondroplasia. A sleep study was arranged to check for sleep apnea, the parents were given growth charts that had been developed for children with achondroplasia so their progress could be

monitored, and they were given advice about what medical problems to look out for. They were advised about suitable seating for home, in the car and in a pushchair since a child with achondroplasia needs good support of their back and head. The counselor gave them contact details of a patient organization and some of the helpful literature which had been produced.

Over the years the family contacted the genetics department and the patient organization many times for support about medical problems and to ensure all appropriate facilities were in place for James to access school and social activities. James was generally a healthy boy, although his frequent upper respiratory infections and constant noisy breathing meant he needed to have his tonsils and adenoids removed. He was keen on sport but disheartened when he competed against his class-mates at school. However, the family found out about the Dwarf Sports Association UK (www.dsauk.org) and James became a keen competitor. Soon he was meeting many other young people with short stature and winning medals. In time he traveled to the international World Dwarf Games representing his country (see *Figure 6.1*), and that is how he met Joanne, who was a swimmer who also had achondroplasia.

In time James and Joanne's relationship blossomed and they were married. James worked as a sports coach and Joanne was an adviser to a large company on accessibility issues, having studied psychology at university. They asked to attend the genetic clinic



Figure 6.1 – Three young boys with achondroplasia at the World Dwarf Games.

Note each has a relatively large head, short limbs and a normal length trunk.

because they wanted to find out about risks for Joanne if she became pregnant, and also what the risks were for a child. Joanne's mother had achondroplasia too, so they knew there was a chance that a child could be like themselves, but they had heard that there was a risk of a severe and lethal condition if both parents were affected; also they were not sure how they would feel if any child they had was of normal stature.

In the clinic they had a detailed discussion about pregnancy risks and Joanne understood she would need to give birth in hospital by caesarean section at around 38 weeks of pregnancy, since women with achondroplasia cannot deliver a baby vaginally. Ideally, she should be referred to a consultant obstetrician early in pregnancy so a plan could be drawn up, and later she should meet her anaesthetist since there are sometimes issues with intubation. The risks for each pregnancy were explained: there is a 50% chance of the child inheriting the altered gene from either mother or father and having achondroplasia, a 25% chance of a child inheriting the normal gene from both and being of normal stature, and a 25% chance of inheriting the altered gene from both parents – a so-called double dose – which results in a child with severely short limbs and a very narrow chest who would only survive for hours or days. They were informed that tests were available in early pregnancy to see which of the possibilities had occurred and to give them the option, if the child was homozygous, of a termination of pregnancy.

6.2. Science toolkit

In the last two chapters we have seen some of the methods used to detect DNA sequence variants, and in some cases considered their effects. In this chapter we will look more systematically at molecular pathology: that is, about how a variant DNA sequence can affect the biochemical functioning of a cell or the characteristics of a person. In this current section we will look at the way different types of variant can affect transcription or translation of a protein-coding gene. Then in *Section 6.4* we will consider the wider questions of whether a variant causes a loss of function or a gain of function of a gene, and how far genotypes are predictive of phenotypes. But first we draw attention to a question of the appropriate use of words (*Box 6.1*). Partly this is about using words that are scientifically accurate but also, for people involved in talking with patients, there is a need to be sensitive to the unintended effects of words – in particular the word 'mutation'. Although we used this word in the title of this chapter, to give a snappy effect, it may be better to use an alternative except for very specific purposes.

Words to describe DNA sequence variants

The word '**mutation**' can be used to describe either an event that produces a DNA sequence variant, or the resulting variant, even if that had been inherited maybe through many generations. In other words, it can describe the process or its product. It may be better to describe the product as a 'variant'. Workers in diagnostic laboratories often use 'mutation' to mean a pathogenic variant, while non-pathogenic variants are termed 'polymorphisms'. The latter term has a specific meaning in population genetics (see *Chapter 9*), and this loose usage is not recommended. Meanwhile, in discussion with patients it is better to avoid the word 'mutation' altogether, except when describing the actual process of genetic change. For some people the word may have pejorative connotations (*'John Smith is a mutant...'*); it is wiser to use a word such as 'variant' or 'fault'.

BOX 6.1

An overview of types of variants

If a protein-coding sequence is to function as proposed in the Central Dogma (*Figure 3.2*), a series of steps is necessary. Any of these steps might be affected by a variant. The main types of variant to consider are:

- deletions of a whole gene
- duplications of a whole gene
- disruption of a gene by a chromosomal rearrangement
- deletions or duplications of one or more exons of a gene
- variants in the promoter or other *cis*-acting regulatory sequence
- variants that affect splicing by altering an existing splice site
- variants that affect splicing by activating a cryptic splice site
- variants in coding sequence that alter the triplet reading frame (frameshifts)
- variants in coding sequence that introduce a premature stop codon (nonsense changes)
- variants in coding sequence that replace one amino acid in the protein with another (mis-sense changes)
- variants in coding sequence that alter one codon for an amino acid into another codon for the same amino acid (synonymous substitutions)

In *Section 6.4* we will consider an alternative classification based on gene function. Variants can be classified as null or amorphic (no product is made, or there is no function), hypomorphic (too little product, or too little function), hypermorphic (too much product, or excessive function) or neomorphic (a new function).

Deletion or duplication of a whole gene

These would be expected to decrease or increase the amount of gene product proportionally to the change in gene number, though this may be modified by feedback controls regulating the level of expression according to the need for the gene product. Not all gene deletions or duplications are pathogenic. It has recently become apparent that there is considerable variation among normal people in the copy number of some genes. Comparative genomic hybridization (*Figure 4.7*) has revealed unexpected large-scale copy number variants that are common and not evidently pathogenic (see *Figure 2.21*). Mostly these do not involve coding sequences, but some do. Examples include the following:

- people vary in the number of tandemly repeated green color vision pigment genes on the X chromosome
- people vary in the number of genes for salivary amylase at chromosome 1p21; those from populations with traditional high-starch diets tend to have higher numbers than people from populations with traditional low-starch diets (Perry *et al.*, 2007)
- some major histocompatibility complex haplotypes at chromosome 6p21 (see *Chapter 10*) contain different numbers of HLA genes.

For most genes, however, copy number changes are abnormal and often pathogenic. A gene is called **dosage-sensitive** if a 50% decrease or increase in copy number (having 1 or 3 copies of a gene that is normally present in 2 copies) causes a phenotypic change. Duplications are less likely than deletions to be pathogenic. Dosage-sensitive genes on the relevant chromosome must explain the pathogenic effects of chromosomal trisomies, and the still more drastic effects of monosomies.

Disruption of a gene

If a gene is disrupted by a chromosomal rearrangement the 5' fragment of the gene retains the promoter and may still be transcribed, but there can be no production of the normal full-length transcript. For example, half of all cases of severe hemophilia A (OMIM 306700) are caused by an inversion in the *F8* (blood clotting Factor VIII) gene (Figure 6.2). Stability of an mRNA is largely mediated by the 3' untranslated region, and a partial mRNA is unlikely to be stable or to encode any product. Thus such disruptions abolish expression of the gene. Occasionally a chromosomal rearrangement may create a novel chimeric gene by bringing together exons of two different genes. Changes of this sort are important in cancer (see *Chapter 7*) and form a partial exception to the rule that disruptions prevent expression of a gene.

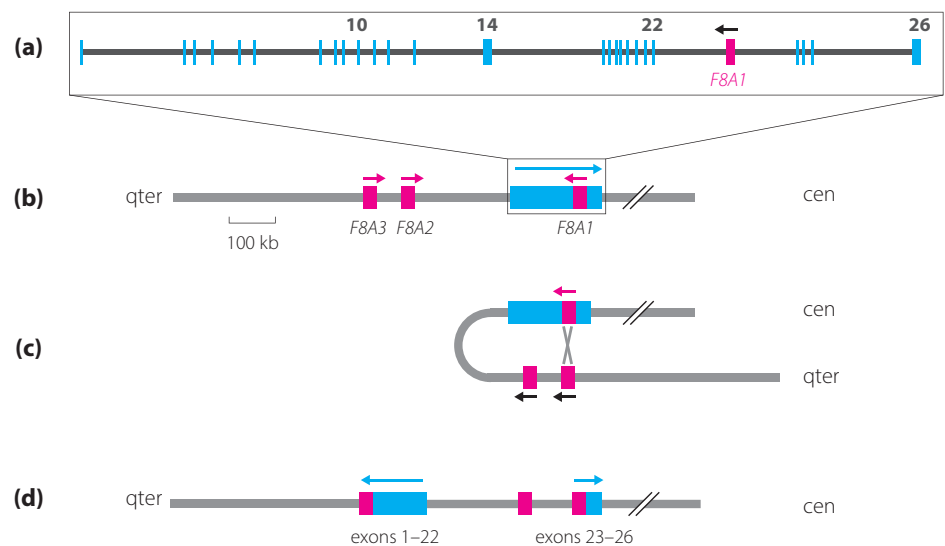


Figure 6.2 – An inversion that disrupts the *F8* gene.

(a) The *F8* gene has 26 exons spanning 187 kb of genomic DNA on the X chromosome at location Xq28. (b) The red boxes show a non-coding sequence present in intron 22 of the *F8* gene, of which two additional copies are located 360 kb and 435 kb away from the start of the *F8* gene. The red arrows show the orientation of the repeats. Note that these are inverted repeats (compare with the repeats in *Disease box 2*, which are direct repeats). (c) During male meiosis this part of the X chromosome has no matching partner. The DNA may loop round to allow two copies of the repeated sequence to pair in the same orientation. Occasionally there is recombination between the paired sequences. (d) The result is a chromosomal inversion of about 500 kb of DNA that disrupts the *F8* gene. Blue arrows show the 5'→3' orientation of the *F8* gene.

Variants that affect the transcription of an intact coding sequence

Correct regulation of gene expression is critical for correct functioning of a cell. Genes must be turned on or off as appropriate for the type of cell and the job it is currently doing. For a coding sequence to be expressed it must have a functioning promoter, appropriate enhancers and a permissive ('open') chromatin environment – see *Chapter 11* for more details of these controls. Interfering with any of these elements could prevent transcription. Thus sequence variants in promoters or enhancers can silence a gene, even though the coding sequence itself may be intact.

Although many individual examples of such effects are known (Gordon and Lyonnet, 2014), they are not as frequent as might be expected. Partly this is because diagnostic laboratories seldom search for such changes. If they found a change in a promoter or enhancer, checking its effect would require experimental investigations that would not fit into the workflow of a diagnostic laboratory. In the current state of knowledge, computer analysis of a change in non-coding sequence would not usually suffice to decide whether it might be pathogenic or not. In addition to this bias of ascertainment, it appears that enhancers in particular are remarkably tolerant of small sequence changes. They seem to have evolved something akin to the fault-tolerant engineering of manned spacecraft. Minor sequence changes may cause small changes in the level of expression of the gene they regulate, and an accumulation of such changes may predispose to various common diseases such as diabetes (see *Chapter 13*), but despite cases such as those cited by Gordon and Lyonnet (2014), they seldom have individual dramatic effects on gene expression.

Changes in the chromatin environment, triggered by a DNA sequence change, are perhaps more frequently pathogenic. One category of non-coding change that is important in pathology concerns DNA methylation. As described in *Chapter 11*, special enzymes attach methyl ($-\text{CH}_3$) groups to cytosine bases in DNA as part of systems regulating chromatin conformation and gene expression. Genes with intact coding sequences can be shut down by inappropriate methylation. One trigger of such methylation is the unstable repeat expansions described in *Disease box 4*. The Fragile X full mutation (**Case 11, Lipton family**) is an example (*Figure 6.3*). An expansion beyond 200 repeats of the $(\text{CGG})_n$ repeat in the 5' untranslated region of the *FMR1* gene triggers methylation of the DNA that prevents transcription of the gene. Similar methylation effects are important in cancer (see *Chapter 7*).

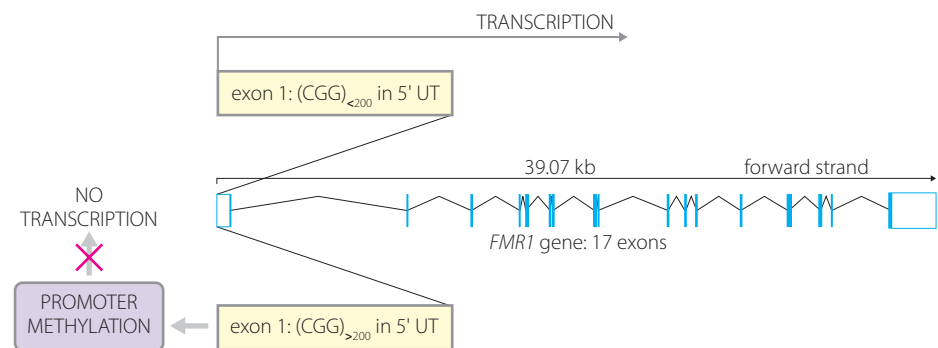


Figure 6.3 – The Fragile X $(\text{CGG})_n$ expansion shuts off transcription of the intact *FMR1* coding sequence.

Fortunately, for most mendelian diseases careful studies of coding sequences and splice sites can usually identify the causative variant in the great majority of cases. Thus regulatory changes may have their main relevance in complex multifactorial disease (*Chapter 13*), where susceptibility may depend on combinations of subtle changes in gene expression, rather than mendelian diseases, which usually have single mutations of large effect.

Variants that affect splicing of the primary transcript

As mentioned in *Chapter 3* (see *Figure 3.10*), a large multimolecular machine, the spliceosome, physically cuts the primary transcript at exon–intron boundaries and splices the exons together. Almost all splice sites consist of a GU dinucleotide in the RNA (GT in the DNA) at the start of an intron and AG at the end, each embedded in a loosely defined consensus sequence and surrounded by short motifs that act to enhance or suppress the activity of the site. Variants can affect splicing either by stopping a normal splice site from being used, or by causing a sequence that is not a normal splice site to be incorrectly treated as one.

Splice sites are not all-or-nothing; potential sites vary in their strength (that is, their affinity for the splicing machinery). Changes to a splice site may completely abolish its function or may alter its strength. Any change to the (almost) invariant GT...AG dinucleotides in the DNA that mark the beginning and end of introns will prevent the affected site from being used. What the cell will do in response to this is hard to predict. The exon involved may be skipped, or sequence from the intron may be retained within the mature mRNA. Often some other nearby site is used as an alternative, with consequent changes to the sequence of the mature mRNA. Other changes close to an exon–intron junction also often affect splicing, but less predictably. They may affect the consensus sequence surrounding the GT...AG or they may affect splicing enhancer or silencer sequences nearby that bind components of the splicing machinery. None of these sequences is rigidly defined. They may be located in the intron near a splice site, but they may also lie in the exon. If a change within a coding sequence alters an exonic splicing enhancer, it may be pathogenic for reasons unconnected with any predicted amino acid sequence change. Two examples will illustrate these effects.

- Near the 3' end of intron 8 of the *CFTR* gene is a run of T nucleotides. In different people there may be 5, 7 or 9 Ts. The 5T variant reduces the strength of the nearby splice site so that exon 9 is often skipped, leading to loss of function. The loss of function is only partial because a proportion of transcripts are correctly spliced. Thus the 5T variant is associated with mild and atypical forms of cystic fibrosis.
- Spinal muscular atrophy (SMA or Werdnig–Hoffmann disease, OMIM 253300) is caused by loss of function of the *SMN1* gene on 5q13. A duplicate copy of the gene (*SMN2*) lies only 500 kb away on the same chromosome, and at first sight should be able to replace the function of the mutated gene. The two genes have only apparently insignificant sequence differences. One of these is a C>T change in exon 7, six nucleotides downstream of the 5' splice site. The substitution apparently only replaces one phenylalanine codon with another (UUC>UUU). But in fact this change disrupts splicing, probably by creating an exonic splicing silencer, so that exon 7 is skipped in about 90% of transcripts. Thus the *SMN2* gene is largely non-functional, and cannot compensate for any loss of function of the *SMN1* gene. Interestingly, some SMA patients have several copies of this gene. Although each one produces very little protein, together they make enough to render the disease in these patients milder and of later onset. See *Disease box 10* for function of the SMN protein and *Box 14.9* for further discussion of SMA.

Some sequence changes affect splicing by activation of a **cryptic splice site**. That is, there is a sequence that by chance has many of the features of a splice site, but is sufficiently

different for the cell not to use it for splicing. A change may increase the resemblance to the point where the cell does start using it. The cryptic site can be either in an exon or an intron, as the following examples show.

- In hemoglobin E the β -globin gene has a G>A substitution 14 nucleotides upstream of the 3' end of exon 1. This would be predicted to cause an amino acid substitution p.Glu26Lys. However, the pathogenic effect is on splicing. Changing this one nucleotide causes the changed sequence to be used as an alternative splice site. Transcripts that use this site are non-functional, and the effect is β -thalassemia.
- One cause of cystic fibrosis is a single nucleotide change c.3849+12191C>T, 12 kb inside the large intron 22 of the *CFTR* gene. This activates a cryptic splice site, causing aberrant splicing and loss of function of the gene. As with the 5T variant described above, only some transcripts use this splice site, so the effect is relatively mild disease.

Variants that activate cryptic splice sites deep within introns cannot be found by the usual approach of sequencing exons; it would be necessary to study mRNA (as cDNA) rather than genomic DNA (although once defined, such variants can be genotyped by any test for a known sequence change). They are undoubtedly under-reported.

As mentioned in *Section 3.4* (see *Figure 3.10*), many transcripts are subject to **alternative splicing**. That is, some splice sites are used by only a proportion of molecules of the primary transcript, resulting in a range of **splice isoforms**. Browsing any gene in the ENSEMBL database, as described in *Section 3.4*, will likely reveal a plethora of different transcripts, most of which are splice isoforms (though some involve using different promoters). Alternative splicing may be tissue-specific, with some isoforms being found only in a certain tissue. The average number of isoforms per locus found by the ENCODE project (see *Section 3.4*) is 5.4. If a change in a consensus splice sequence or a splicing enhancer changes the strength of a splice site, it may affect the balance of splice isoforms, and this may produce a phenotype. Again, such changes are more likely to act as susceptibility factors for common disease rather than mendelian disease mutations.

In summary, effects on splicing are widespread but, apart from changes to the invariant GT...AG dinucleotides, hard to predict. Computer programs are available that try to predict the effect of a sequence change on splicing. The best ones, used in combination, are probably right 90% of the time, but they are far from infallible. In addition to effects of variation in a target sequence, variants in the splicing machinery itself can be pathogenic – see *Disease box 10*.

Variants that cause errors in translation

When there is a change in a protein-coding sequence the effect needs to be considered by reference to the genetic code (*Table 6.1*). Changes are of three types:

- mis-sense changes replace a codon for one amino acid with a codon for a different one, producing a substitution of a single amino acid
- nonsense changes replace a codon for an amino acid with a STOP codon (UAG, UAA or UGA)
- frameshift changes insert or delete nucleotides so as to alter the reading frame.

We consider each category in turn.

Table 6.1 – The genetic code

1st base in codon	2nd base in codon				3rd base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Some amino acids such as serine and arginine have multiple codons, whereas tryptophan and methionine have only one each. AUG doubles as the initiator codon and as the codon for internal methionines. Note the three stop codons.

Mis-sense changes. A single nucleotide change within a coding sequence changes one codon into another. Assuming the changed codon is not a stop codon, there are two possible results.

- *Synonymous (or silent) substitutions* are single-nucleotide changes that replace a codon with a different one that encodes the same amino acid – for example, a T>C change in the DNA that converts a UUU codon in the mRNA to UUC. Both code for phenylalanine. These would normally have no effect unless, like the change in the *SMN2* gene mentioned above, they affect splicing.
- *Mis-sense changes* replace one amino acid with another. Use *Table 6.1* to identify the effect of a codon change on the amino acid sequence.

Deciding whether a mis-sense change would be pathogenic is difficult. How important is that amino acid to the functioning of that particular protein? Does it form part of the active site of an enzyme, or some other essential functional element? Maybe it is crucial for establishing the 3-dimensional structure of the protein. And how well is the replacement tolerated? Amino acids differ in their chemistry and size; some pairs are more compatible than others. However, for many positions in many proteins, replacing one amino acid with another has no effect on the function. Various computer programs (e.g. POLYPHEN-2, SIFT) try to address these questions for any given change in any given protein. Mostly they work by looking at related proteins in humans and other species, and asking if the changed amino acid is always the same in related proteins – if so, probably the protein would not tolerate a replacement.

It is important to remember, as mentioned above, that seemingly innocuous synonymous or mis-sense changes may have a major pathogenic effect by disrupting splicing.

Nonsense changes. The three codons UAG, UAA and UGA in mRNA are stop codons (Table 6.1). A single nucleotide change that converts any other codon into a stop codon (TAG, TAA or TGA in the DNA) causes the ribosome to detach and protein synthesis to terminate at that point. Such variants are called **nonsense changes**. Contrary to common belief, mRNAs containing premature termination codons do not normally produce truncated proteins. Cells have a very interesting mechanism (**nonsense-mediated decay**) for detecting and degrading mRNAs that contain premature termination codons (Figure 6.4). In some cases a certain amount of a truncated protein may be produced, but the usual effect of a nonsense change is the same as a complete deletion of the gene.

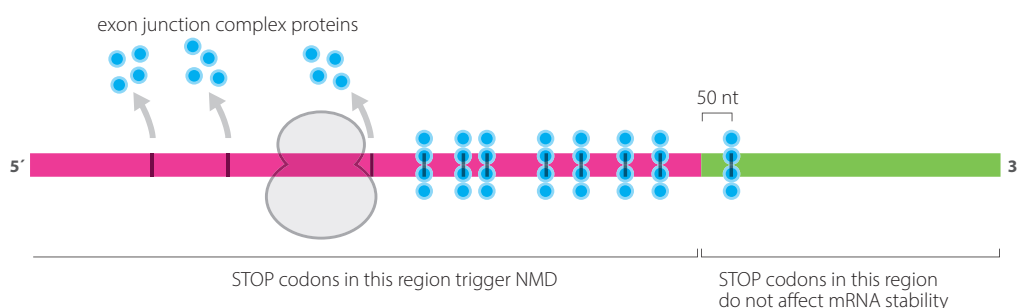


Figure 6.4 – Nonsense-mediated decay (NMD).

When a spliced mRNA is exported from the nucleus to the cytoplasm certain components of the splicing machinery (the exon junction complex, EJC) remain attached to each splice site. As the first ribosome moves along the mRNA it displaces each EJC until it reaches the stop codon and detaches. If this happens in the red segment of the mRNA, one or more EJCs will remain in place. This triggers degradation of the mRNA. Thus premature termination codons in the red segment trigger mRNA decay and no truncated protein is produced, but a nonsense change in the green zone is able to direct production of a truncated protein that may be pathogenic.

Nonsense-mediated decay probably evolved to protect cells against possibly toxic truncated proteins (see below). It presumably also explains why the last exon of many genes is very large – you can't split the 3' untranslated region between several exons or use of the normal stop codon would trigger nonsense-mediated decay. Some rather mystifying differences between the effects of nonsense codons in different parts of the same gene are the result of nonsense-mediated decay operating in the red zone (Figure 6.4) but not in the green zone (see OMIM 602229 for an example). For further details see the review by Holbrook *et al.* (2004).

Frameshifts. Insertion or deletion of any number of nucleotides that is not a multiple of 3 produces a frameshift. The effect is to change completely the reading of all the message downstream of the change (see Box 3.3). In principle, a completely novel polypeptide might be produced by translation of the frameshifted mRNA. But even if such a protein were produced, it would probably be unstable. Only a very small subset of all the myriad possible polypeptide chains can fold so as to produce a stable protein. Cells detect and degrade those that fail to fold correctly. More commonly, however, no novel polypeptide is produced because it is usually not long before the frameshifted message includes

a stop codon. *Figure 6.5* shows an example. As explained above, mRNAs containing premature termination codons are normally degraded and do not produce protein. Thus, one way or another, no protein is produced.

Deletions or duplications of whole exons are a frequent cause of frameshifts. When one or more complete exons of a gene are deleted or duplicated, the effect will partly depend on whether or not this results in a frameshift. Deletion of a 125-nucleotide exon will produce a frameshift, while if the exon had 126 nucleotides it would not. The major role of frameshifts in the molecular pathology of DMD is described in *Section 6.3*.

Table 6.3 in *Section 6.4* gives a general guide to the likely effect of the various types of mutation.

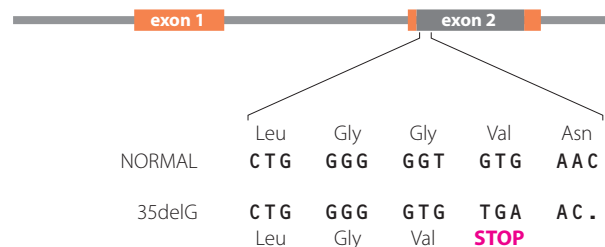


Figure 6.5 – The c.35delG allele of the connexin 26 (*GJB2*) gene.

This mutation is a frequent cause of autosomal recessive congenital deafness; the connexin 26 protein has important functions in the inner ear. The diagram shows the 2 exons of the *GJB2* gene, with coding sequence in gray and the 5' and 3' untranslated regions in color. The gene sequence includes a run of 6 consecutive G nucleotides. When such homopolymer runs are replicated, if the DNA polymerase drops off the template and then re-associates, it may miss or repeat a nucleotide (replication slippage). Thus homopolymer runs are hotspots for single nucleotide insertions or deletions. Reading the frameshifted *GJB2* message the ribosome quickly hits a stop codon.

6.3. Investigations of patients

Here we consider the molecular pathology of the variants that have already been identified, as well as the investigation of **Case 14 (Jenkins family)**.

CASE 1 ASHTON FAMILY

- John, healthy 28-year-old son of Alfred Ashton
- Family history of ? Huntington disease
- Autosomal dominant inheritance
- Need for diagnostic PCR test
- PCR test confirms diagnosis in John's father
- Pros and cons of predictive test
- Molecular pathology
- Possibilities for therapy

1 8 67 103 **153** 395

The trinucleotide repeat expansion was described in *Chapter 4*. It adds CAG triplets into the coding sequence of the *HTT* gene. When the gene is transcribed and translated the expansion does not disrupt the reading frame. The protein product is the normal 3142 amino acid huntingtin protein, but with an expanded run of glutamines near the N-terminal end (CAG is a codon for glutamine). Affected people are normally heterozygotes, but rare homozygotes have been described, and they are usually clinically indistinguishable from heterozygotes. This suggests that it is the presence of the mutant protein, rather than the absence of the normal protein, that is causing the problems. No Huntington disease patients have been described who have deletions, nonsense changes or other pathogenic variants in the *HTT* gene. Additionally, at least one person has been described who does not have Huntington disease despite having a chromosomal rearrangement that disrupts one copy of the *HTT* gene. Together, these observations make clear that Huntington disease must be caused by a gain of function.

The mutant protein is somehow toxic to neurons, especially in the striatum and caudate nucleus of the brain. Gradual cell death leads to the late-onset disease. Exactly why the mutant protein is toxic is controversial. It forms aggregates inside neurons but it is not clear that the aggregates are themselves harmful. More probably an intermediate in the aggregation process is toxic. The normal protein interacts with many other proteins – it probably functions as some sort of scaffold for assembling multiprotein complexes – making it very difficult to identify one key altered function. It is also possible that abnormal translation of mRNA molecules carrying expanded repeats may produce toxic RNA molecules (Cleary and Ranum, 2013).

CASE 2 BROWN FAMILY

- Baby Joanne, recurrent infections, poor growth
- Sweat test confirms she has cystic fibrosis
- Autosomal recessive inheritance
- Need for molecular test
- *CFTR* variants identified
- Molecular pathology

2

10

67

132

154

313

395

Cystic fibrosis is a recessive disease caused by complete loss of function of the chloride channel encoded by the *CFTR* gene. Being affected, Joanne must have no functional copy of the gene. DNA testing (*Chapter 5*) revealed three sequence changes: c.236G>A in exon 3, c.1521_1523delCTT (p.F508del) in exon 10 and c.2620–15C>G in intron 15. Of these, p.F508del is very well understood, being by far the most frequent variant in Europeans with cystic fibrosis (see *Table 12.2*). The change is an in-frame deletion of three consecutive nucleotides in the coding sequence. A virtually full-length protein is produced with 1479/1480 correct amino acids – but the one missing amino acid affects the structure. The changed protein is not correctly processed after synthesis, and fails to locate to the apical cell membrane where it is needed. The result is a complete loss of function. Joanne's other two variants were not common and required further analysis. The intron 15 change was thought to be non-pathogenic for two reasons:

- (1) The C>G change was in intron 15, 15 nucleotides before the start of exon 16. Intronic changes can sometimes be pathogenic by affecting splicing. Splice prediction programs did not suggest any effect on splicing, but without RNA studies an effect could not be ruled out. However, in this case there was a second line of evidence:
- (2) When DNA from Joanne's parents David and Pauline was checked, both the p.F508del and c.2620–15C>G variants came from David. They must both be in the same gene copy. Even if the c.2620–15C>G variant was pathogenic, it would not explain Joanne's disease, because the gene she inherited from Pauline must also carry a pathogenic change.

Joanne's other variant, c.236G>A did come from her mother. The nucleotide change is within exon 3 and converts the codon for tryptophan 79 (TGG) into a stop codon (TAG). This is an unambiguously pathogenic change. Because of nonsense-mediated mRNA decay (see above) it is unlikely that the variant gene would produce any protein, and any that it did produce would certainly be non-functional since the stop codon occurs very early in the sequence.

As previously mentioned, one reason for identifying the pathogenic variants in Joanne was to see whether she would be eligible for one of the new variant-specific drugs. A drug, Ivacaftor, partially corrects the effect of one of Joanne's variants, p.F508del. Drugs like Ivacaftor work by improving the function of a defective CFTR protein. Different types of CFTR variants produce CFTR proteins that are defective in different ways, and each type of defect requires a different specific drug. But a nonsense variant like Joanne's second

mutation, p.W79X, produces no protein and so its effect cannot be ameliorated by one of these drugs. An additional problem is the extremely high cost of these new drugs, which the patient would need to take for their entire life. This has led to questions whether the money could not achieve greater overall health benefit by being used in other ways. The issue has become political, with high-profile campaigns by patients and their families to gain access. Manfredi and colleagues (2019) provide an interesting commentary.

CASE 3 KOWALSKI FAMILY

- Karol, first son of Kamil and Klaudia
- Developmental delay, hypotonic, severe intellectual disability
- Difficulties of genetic testing in such cases
- Likely need for exome sequencing
- Negative SNP chip test for microdeletions
- Exome sequencing
- *De novo* *ARID1B* variant identified
- Possibilities for therapy

3

10

67

102

134

155

395

As described in *Section 5.3*, Karol's entire exome was sequenced, revealing 16 400 differences from the Reference Human Genome (a typical figure for exome sequences). After eliminating variants that were too frequent in the population to be plausible causes of Karol's rare condition, 410 remained. The analysis then proceeded as follows.

- (1) Variants predicted not to change the amino acid sequence of any protein were excluded. These would include non-coding variants and synonymous changes that replace one codon for an amino acid with another for the same amino acid. The remaining candidates are non-synonymous coding changes, insertions or deletions, or changes that affect splice sites or truncate the protein. Note that this process might have wrongly excluded occasional pathogenic synonymous variants (like the TTT>TTC change in the *SMN2* gene), and pathogenic changes in promoters, enhancers or the untranslated regions of an mRNA.
- (2) Mis-sense changes that were judged unlikely to be pathogenic by the POLYPHEN2 or SIFT programs (see above) were excluded. However, these programs are only around 80% accurate on average, so some variants might have been wrongly excluded (or included) at this stage.
- (3) At this point in the analysis it was decided to test the data on the alternative hypotheses that Karol's problem was either autosomal recessive or dominant. On the recessive hypothesis a candidate gene should have plausibly pathogenic changes in both copies. In fact, no such gene was identified, so the list of variants was searched for a plausible dominant change.
- (4) The remaining short list of possible dominant variants was further filtered by looking for *de novo* changes that were not present in either parent. DNA was obtained from Karol's parents and their exomes sequenced. This reduced the list to a single variant, a *de novo* single nucleotide substitution c.3304C>T in exon 12 of the *ARID1B* gene. This is a classic nonsense change, changing codon 1102 from CAG (glutamine) to TAG (STOP). The change is in exon 12 of the 20-exon gene, so the mutant mRNA should trigger nonsense-mediated decay, thus effectively acting as a null allele. The other allele is intact. Thus Karol has one functional and one non-functional *ARID1B* gene.
- (5) Finally, Sanger sequencing was used to double-check that the sequence change was correctly identified and was truly *de novo* (an important check because apparent *de novo* changes are often sequencing errors), and the status of *ARID1B* as a candidate gene was reviewed. The result was highly satisfactory. In *Section 6.4* we present evidence showing that loss of function of a single copy of *ARID1B* would be pathogenic; it was not necessary that Karol should have a homozygous change. The ARID1B protein forms part of a chromatin remodeling complex, a multiprotein

machine that controls the local chromatin conformation, and hence the expression of many other genes. Variants in several proteins of these complexes were associated with intellectual disability, and in fact *de novo* mutations in this gene had been described previously in children with severe intellectual disability. Thus the cause of Karol's problem had been identified.

CASE 4 DAVIES FAMILY

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy
- X-linked recessive inheritance
- Problems of testing dystrophin gene
- Exon 44–48 deletion identified by MLPA
- Molecular pathology

4

11

68

98

156

285

315

395

Martin has a deletion of exons 44–48 of the dystrophin gene (*Figure 4.11*). Considering the deletions in this part of the dystrophin gene that have been reported in different patients, a seemingly paradoxical situation is observed (*Table 6.2*).

Multi-exon deletions often have less severe consequences than deletions of single exons. For example, deletion of any one of exons 43–46 causes the severe Duchenne muscular dystrophy, yet larger deletions that take out pairs of the same exons, or even all four, cause a much milder condition, Becker muscular dystrophy (OMIM 300376). In other words, having two or even four lethal mutations simultaneously results in a milder disease.

The key to the paradox is to ask whether or not a deletion produces a frameshift. For example, *Table 6.2* shows that if either exon 43 or 44 is deleted there is a frameshift, but if both are deleted the two frameshifts cancel each other out. The long dystrophin molecule functions within the muscle cell a bit like a rope with hooks at each end (*Figure 6.6*). The body of the rope consists of repeated units, the precise number of which is not critical to dystrophin function. If Martin simply made slightly smaller than normal dystrophin molecules, but with the hooks at each end intact, he would have the milder Becker muscular dystrophy (BMD) with a relatively good prognosis. But adding up the exon lengths in *Table 6.2*, we can see that his deletion produces a +1 frameshift. Any protein produced would have a completely wrong amino acid sequence downstream of the frameshift – but in fact a ribosome reading the message in this novel reading frame would soon encounter a stop codon. As before, nonsense-mediated decay means that Martin's gene would produce no protein product, rather than a truncated protein. Studying a muscle biopsy with labeled dystrophin antibody could confirm the absence of dystrophin protein (see *Figure 1.4*). In Martin's case this adds no useful new

Table 6.2 – Effect of whole exon deletions in the dystrophin gene

Exon	Size (bp)	Frame	Effect of single exon deletion	Multi-exon deletions causing BMD		
42	195	0	BMD			
43	173	−1	DMD	Deletion of exons 43 + 44		Deletions of exons 43–46
44	148	+1	DMD		Deletion of exons 44 + 45	
45	176	−1	DMD			
46	148	+1	DMD			
47	150	0	BMD			
48	186	0	BMD			

DMD Duchenne muscular dystrophy, a lethal condition; BMD Becker muscular dystrophy, a milder condition.

Reproduced from Strachan & Read (2019) *Human Molecular Genetics 5e*, with permission from Garland Science/Taylor & Francis LLC.

information, but in boys where the MLPA deletion screen identifies no mutation it is a valuable diagnostic test. Searching all through the huge dystrophin gene (79 exons, 11 kb of coding sequence, 2.4 Mb of genomic DNA) for a point mutation can be seriously challenging, and a muscle biopsy provides a much easier confirmation of the diagnosis.

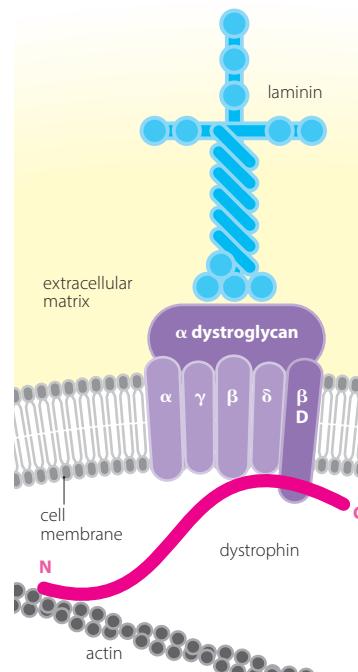


Figure 6.6 – The dystrophin molecule anchors the cytoskeleton of muscle cells to the extracellular matrix, via the dystrophin glycoprotein complex.

This includes the α , β , γ and δ sarcoglycans (mutations in which cause limb-girdle muscular dystrophies) and the α and β (labeled βD) dystroglycans. Muscle cells that lack dystrophin are mechanically fragile, and fail after a few years, hence the progressive muscle weakness.

CASE 6 FLETCHER FAMILY

- Frank, aged 22, with increasingly blurred vision
- Family history of visual problems
- Possible mitochondrial inheritance
- ? Leber hereditary optic neuropathy
- Test mitochondrial genome
- m.G3460A mutation identified
- Molecular pathology
- Possibilities for therapy

5

13

69

130

157

395

The pathogenic variant was identified as m.G3460A in Frank's mitochondrial DNA (*Chapter 5*). Most mitochondrial proteins are encoded by nuclear genes, synthesized on cytoplasmic ribosomes and then imported into the mitochondria, but the 13 protein-coding genes in mitochondrial DNA are transcribed and translated within the mitochondrion in a manner closely similar to the way nuclear genes are expressed. They encode components of the electron transport chain of oxidative phosphorylation, the process by which mitochondria generate energy in the form of ATP molecules (*Figure 3.9b*).

Comparing mitochondrial with nuclear DNA variants, large deletions and duplications are relatively more common, and splicing mutations are absent because mitochondrial genes do not have introns. Heteroplasmy introduces an extra layer of variability between genotype and phenotype. The three common LHON alleles are all mis-sense variants affecting three different proteins in the electron transport system: p.Ala52Tyr in the ND1 protein, p.Arg340His in ND4 and p.Met64Val in ND6. Frank's variant is the first of these. Each variant reduces the efficiency of oxidative phosphorylation. Defects in oxidative phosphorylation typically affect tissues that have high energy demands, such as the retina. It is not clear why these particular defects should predispose to a sudden irreversible loss of vision.

CASE 10 O'REILLY FAMILY

- Orla has severe myopia, short stature and hip problems
- Family history of similar problems
- ? Stickler syndrome
- Test collagen II genes
- Sequencing identifies *COL2A1* variant
- Molecular pathology
- Possibilities for therapy

57

70

134

158

395

Orla's combination of high myopia and joint problems suggested a problem with Type II collagen, which has important roles in cartilage and in the vitreous humor of the eye. Sequencing of her *COL2A1* gene revealed she was heterozygous for a deletion of 2 nucleotides in exon 40 (c.2488_2489del, *Section 5.3*). Reading the frameshifted message, the ribosome would encounter a stop codon after a further 42 codons. Nonsense-mediated decay would probably ensure that the affected allele produced no product. Thus Orla makes only half the normal quantity of the Type II collagen.

Defects in synthesis of collagens cause chondrodysplasias – a group of some 150 clinically defined phenotypes with defects of cartilage that often lead to defects of modeling of long bones. The *COL2A1* chondrodysplasias form a spectrum from achondrogenesis type II, an intrauterine or perinatal lethal condition, through hypochondrogenesis, spondyloepiphyseal dysplasia (SED) and Kniest dysplasia, to Stickler syndrome, a relatively mild condition that is often diagnosed late. Interestingly, the variants with the most drastic effect on protein synthesis are not the ones that produce the most severe clinical phenotypes.

- The most severe phenotypes are caused by mis-sense changes that replace glycines in the Gly–X–Y units of the triple helical region (see *Box 3.4*). Only glycine, the smallest amino acid, can fit into the interior of the tightly packed triple helix, so replacements severely disrupt assembly of the fibril.
- Whole exon deletions and exon skipping due to aberrant splicing are also associated with relatively severe phenotypes. The *COL2A1* gene has 54 exons, of which exons 8–49 encode the triple helical region. These exons are all frame-neutral (45, 54, 99 or 108 bp). Thus deleting or skipping exons in this region does not cause a frameshift or nonsense-mediated decay; the variant protein is just a little shorter. However, when heterozygous people try to combine chains of different lengths into the triple helix, this disrupts the structure.
- Nonsense changes and frameshifts cause the mildest phenotype, Stickler syndrome, as in Orla. Because of nonsense-mediated decay there are no abnormal chains to interfere with assembly of the normal chains, but there is a quantitative deficiency of collagen II. A heterozygote for a null mutation would have a 50% level of normal triple helices, whereas if molecules are selected at random from a 50:50 mix of normal and abnormal polypeptide chains, only one triple helix in eight would consist of three normal chains.

This series shows that it can be better (in a heterozygote) to make no protein than one that can interfere with the function of its normal counterpart. Variants are called **dominant negative** if the changed product prevents the normal product from functioning (*Figure 6.7*).

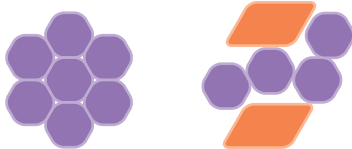


Figure 6.7 – A dominant negative effect.

The variant allele produces an abnormal protein that disrupts the assembly of a multiprotein complex. In heterozygotes dominant negative alleles produce more severe phenotypes than null alleles.

CASE 13 NICOLAIDES FAMILY

- Spiros and Elena both carriers of β -thalassemia
- Need to define mutations for prenatal diagnosis
- Allele-specific PCR shows Spiros carries the p.Gln39X variant
- Restriction digest shows Elena carries the c.316–106C>G variant
- Molecular pathology

117

129

159

316

395

Spiros has the p.Gln39X variant. A single nucleotide substitution converts a codon for glutamine (CAG) into a stop codon (TAG) (*Figure 6.8*). The changed gene produces no product, causing a β^0 thalassemia (complete absence of β -chains) in homozygotes.

Elena's mutation (*Figure 6.9*) is more subtle. The change, c.316–106 C>G, is a single nucleotide substitution deep within intron 2 of the β -globin gene, 106 nucleotides upstream of the start of exon 3. It is an example of the sort of mutation that could only be first noticed by studying mRNA, although in Elena's case the laboratory was able to check using a simple restriction digestion, see *Figure 5.8*, because it was looking for a known variant. This apparently innocuous change activates a cryptic splice site that is then preferentially used compared to the normal donor splice site. The result is a β^+ thalassemia – that is, the mutant gene produces some correct β -chains, using the normal splice site, but not a sufficient quantity.

INTRON 1	Leu	Leu	Val	Val	Tyr	Pro	Trp	Thr	Gln	Arg	Phe	Phe	Glu
ccacccttagGCTG	CTG	GTG	GTC	TAC	CCT	TGG	ACC	CAG	AGG	TTC	TTT	GAG	
ccacccttagGCTG	CTG	GTG	GTC	TAC	CCT	TGG	ACC	TAG	AGG	TTC	TTT	GAG	
INTRON 1	Leu	Leu	Val	Val	Tyr	Pro	Trp	Thr	STOP				

Figure 6.8 – The p.Gln39X allele of the β -globin gene.

This single nucleotide change near the start of exon 2 converts a glutamine codon into a stop codon. Lower case letters represent intron, upper case exon.

NORMAL INTRON 2	c t a a t a g c a g c t a c a a t c c a g t a c c a t t c t g c t
MUTANT- part handled as exon, with new donor splice site	C T A A T A G C A G C T A C A A T C C A G g t a c c a t t c t g c t

Figure 6.9 – The c.316–106C>G variant in the β -globin gene.

This single nucleotide change activates a cryptic splice site deep within intron 2 of the β -globin gene. Lower case letters represent intron, upper case are the abnormal exon.

Table 5.2 listed three other β -globin variants that are common causes of thalassemia in Greek Cypriots. Interestingly, all three affect splicing.

- The commonest variant, c.93–21G>A, again activates a cryptic splice site in an intron, intron 1 this time. Studies of the mRNA showed that 80% of the time the new splice site is used, so this allele produces only 20% of the normal quantity of β -globin. For many genes a 20% level of function would be sufficient, but β -globin is required in large amounts, so the result is a β^+ thalassemia.
- The second variant, c.92+6T>C, is an example of a change that reduces the efficiency of a normal splice site. The changed nucleotide is in intron 1, 6 nucleotides downstream from the start of the intron (GGCAGgttggtatca..., where exon sequence is in upper case letters and the nucleotide that is changed is underlined). The T nucleotide is not part of the invariant GT found at the 5' end of every intron,

but forms part of the context that is necessary for the splice site to be recognized efficiently. Again the result is a β^+ thalassemia.

- The third common Greek Cypriot variant is c.92+1G>A. This directly changes the obligatory GT at the exon 1 – intron 1 splice site to AT. No transcripts can be correctly spliced, so this allele produces β^0 thalassemia.

CASE 14 JENKINS FAMILY

- James Jenkins, achondroplasia diagnosed in infancy
- No family history
- Father was 58 years old when James conceived
- James's wife Joanne also has achondroplasia
- Obstetric problems and risks to children
- All cases have same *FGFR3* mutation
- Reasons for apparent high mutation rate
- Possibilities for therapy

143

160

395

All achondroplastic individuals are heterozygous for exactly the same mis-sense change, replacement of glycine 380 by arginine in fibroblast growth factor receptor 3. This is a very surprising finding, given that 80% of cases are caused by *de novo* mutations in unrelated individuals. That was the case for James, both of whose parents were of normal stature, although Joanne's mother was achondroplastic. Achondroplastic individuals like James and Joanne are fit and fertile, but face many social disadvantages, and thus have on average far fewer children than people of normal stature. Thus, as with neurofibromatosis (NF1, *Disease box 1*), the condition is maintained in the population by recurrent mutation – see *Chapter 9* for some calculations. Given the relatively high incidence of achondroplasia (around 1 in 20 000 livebirths), this implies that the *FGFR3* gene has a high mutation rate. But the figure is doubly surprising because virtually all the mutations causing achondroplasia affect exactly the same nucleotide, c.G1138 (a tiny minority involve the next-door nucleotide, but have the same effect, p.G380R, at the protein level). This nucleotide appears to have by far the highest mutation rate of any nucleotide in the human genome. What is so special about it?

For many years this remained a mystery. Eventually, pioneering work by researchers at the University of Oxford showed that the answer was not a uniquely high mutation rate, but selection within the male testes for spermatogonia carrying the mutation (Goriely *et al.*, 2009). In fact, *FGFR3* mutations occur at much the expected frequency, but mutant cells in the male germline have a proliferative advantage, resulting in disproportionate numbers of sperm carrying this specific variant. Almost all the mutations occur in males, and the frequency rises markedly with age because of the continuing reproductive advantage of mutant spermatogonia. James's father John was 58 when James was conceived.

As discussed in *Section 6.4*, variants in the fibroblast growth factor receptor genes provide some of the clearest examples of genotype–phenotype correlations.

6.4. Going deeper...

A central ambition of clinical molecular genetics is to be able to predict the phenotypic effect of any DNA sequence change. This goal, of establishing genotype–phenotype correlations, can be divided into two tasks:

- deciding how a variant affects the function of a gene – does it cause loss of function (partial or total), gain of function, or does it have no effect?
- deciding how loss or gain of function of a gene will affect the phenotype. Loss of function is not necessarily pathological – it depends on the gene. Loss of function variants in X-linked genes have been found in healthy boys (Tarpey *et al.*, 2009), while McArthur *et al.* (2012) identified healthy individuals among

the 1000 Genomes cohort who had homozygous loss of function of genes. Evidently we can get by perfectly well without the function of certain genes.

Loss of function and gain of function changes

A first question to ask concerning the effect of a variant is, does it cause a loss of function, a gain of function, or have no effect? To put it differently, does the altered gene product simply fail to do its normal job (partially or totally), does it do something actually harmful, or does it simply work just as before? *Table 6.3* summarizes the likely effects of different types of variant.

Table 6.3 – Common types of variant and their likely effects

Type of variant	Likely effect on a gene
Large deletion or inversion	Most likely to completely abolish function.
Duplication of whole gene	Will increase the amount of product by 50% (from 2 to 3 gene copies). This will generally have no phenotypic effect, unless the precise level of the gene product is critical.
Change in promoter or regulatory sequence	May reduce or increase the level of transcription, or alter the response to control signals. Changes to enhancers may affect the tissue specificity of expression – see Gordon and Lyonnet (2014). Any protein produced has the normal structure and function.
Change in intron	Most likely to have no effect – but can sometimes affect splicing, usually resulting in a loss of function.
Change in 5' or 3' untranslated region of mRNA	Most likely to have no effect – but can sometimes affect the stability or translation efficiency of the mRNA, generally leading to a loss of function.
Splicing variant	Mutation of a canonical GT...AG splice site is likely to abolish function of that allele unless there is an alternative splice form that skips that exon. Other variants may have more subtle effects, causing a proportion of transcripts to be incorrectly spliced or changing the pattern of alternative splicing. This can produce a partial loss of function. A single nucleotide substitution deep within an intron may activate a cryptic splice site.
Frameshift	Likely to abolish function of that allele. The polypeptide downstream of the frameshift bears no resemblance to the correct sequence. Usually a stop codon will be encountered quite soon as the ribosomes read the frameshifted codons. The effect will then be the same as a nonsense variant.
Nonsense variant	Likely to abolish function of that allele. Most mRNAs containing premature termination codons are not translated to produce a truncated protein. Instead, they are degraded and not used at all (nonsense-mediated mRNA decay).
Mis-sense variant	Effect very variable, depending on the nature and function of the amino acids concerned. Could be loss or gain of function, or no effect. Replacing an amino acid by a chemically very similar one is likely to have less effect than a more radical change. Some amino acids in a protein are essential to its structure or function, others are not. Apparent mis-sense variants may actually be pathogenic because of an effect on splicing.
Synonymous codon substitution	Most likely to have no effect – but can sometimes affect splicing.

As one would expect, there are many different ways of altering a gene sequence so as to cause a loss of function, but only a few very specific ways to produce a gain of function. Any such gain would most likely not involve gain of a radically new function (although there are cases of such changes – see, for example, the p.M358R variant of the alpha-1-antitrypsin protein, OMIM 613490). More usually, a gain of function involves a loss of regulation such that the product functions when it should not. It may become insensitive to signals that should shut it off, or be expressed at too high a level. A cell surface receptor may become constitutionally active, transmitting a signal to the cell interior even when no ligand is present. Huntington disease is an example of an alternative mechanism, where a mutant protein has a toxic gain of function (see above). Gain of function mechanisms require the mutant allele to produce an abnormal protein (or sometimes a toxic mRNA). Gain of function variants are therefore mis-sense or regulatory variants, not deletions, nonsense or frameshift changes. Considering the spectrum of reported pathogenic variants in any particular gene, a very clear difference can be seen between loss of function and gain of function conditions (*Figure 6.10*).

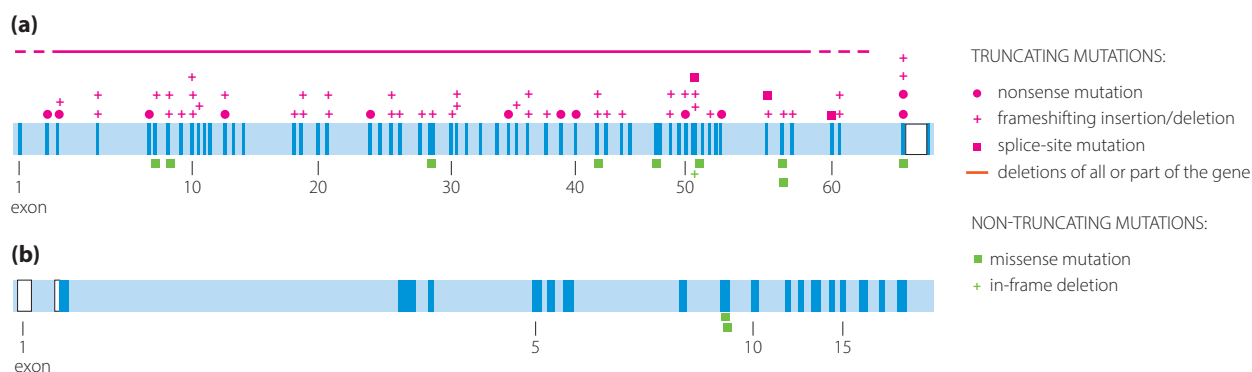


Figure 6.10 – The contrasting spectra of variants causing a loss of function or a gain of function condition.

(a) Variants in the *ATM* gene reported in patients with ataxia-telangiectasia (OMIM 208900). (b) Variants in the *FGFR3* gene reported in patients with achondroplasia (OMIM 100800). Part (a) reproduced from Strachan & Read (2019) *Human Molecular Genetics 5e*, with permission from Garland Science/Taylor & Francis LLC.

As mentioned earlier, chromosomal rearrangements sometimes create a novel gene by juxtaposing exons of genes that are normally far apart in the genome. Such chimeric genes may gain novel functions. Exon shuffling has probably been important in evolution – many proteins can be seen to be made of varying combinations of a limited repertoire of functional domains, each encoded by separate exons (see *Figure 3.11*). Because chromosomal translocations cause major problems in meiosis (*Chapter 2*), such rearrangements are unlikely to cause inherited disease. However, cancers develop entirely by mitosis from mutant somatic cells, and so they face no obstacle to the propagation of cells with chromosomal rearrangements that create an oncogenic gain of function. See *Disease box 7* for an example.

Dominant or recessive?

Dominance and recessiveness are properties of characters or phenotypes, not of genes. We should not really talk of ‘dominant genes’ and so on, though it is sometimes hard

to avoid doing so. A character is dominant if it is seen in a heterozygote, and recessive if not.

- Gain of function changes are expected to produce dominant characters. The gain of function is present in a heterozygote, regardless of the presence of the normal allele.
- Loss of function variants can produce either dominant or recessive characters, depending how sensitive the organism is to a partial loss of that particular function (*Figure 6.11*). For many gene products we can get by perfectly well on 50% of the normal level, so any loss of function phenotype is recessive – cystic fibrosis is one example among many. For some gene products in some cells 50% of the normal function is not sufficient (**haploinsufficiency**), and heterozygous carriers of a loss of function variant have a dominant condition.
- Dominant negative variants are a special class of loss of function alleles, where the abnormal product interferes with the function of the normal product (see *Figure 6.7*). A heterozygous carrier of such a variant will have less than 50% of the normal level of function, and the result is often a dominant condition. The effect depends on the presence of the abnormal product, so the causative variant is always a mis-sense change, not a frameshift or nonsense change. For example, a simple loss of function variant in the *GJB2* gene was shown in *Figure 6.5*. It causes recessive deafness. The gene product, connexin 26, acts as a hexamer to form gap junctions that allow ions to pass between cells, which is important in the inner ear. Certain full-length variant connexin 26 proteins (p.W44C, p.R75Q) interfere with assembly of the hexameric gap junction in a dominant negative manner, and produce dominant deafness.

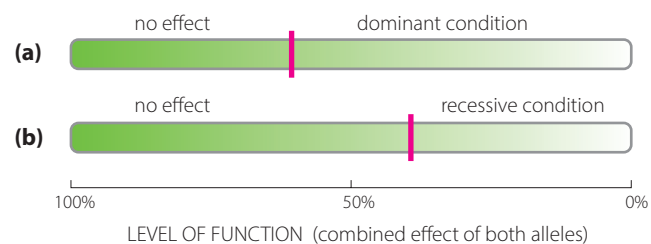


Figure 6.11 – Whether a loss of function mutation causes a dominant or a recessive condition depends on how sensitive the organism is to loss of that function.

If the threshold for normal function is as shown in (a) a heterozygote with 50% function will be affected, and the condition will be dominant. This situation is called haploinsufficiency. If the threshold is as in (b) the condition will be recessive.

In *Disease box 5* we saw how variants in the same ion channel genes could result in either the dominant Romano–Ward or recessive Jervell and Lange–Nielsen syndromes – both with liability to cardiac arrhythmias, but the latter also with hearing loss. This shows how different cell types may have different sensitivities to the same loss of function. Only a severe overall loss of ion channel function affects hearing, but heart cells are sensitive to a lesser degree of loss.

In checking candidate genes for a patient's condition it is often important to know whether loss of function variants would show haploinsufficiency. For example, **Karol Kowalski (Case 3)** was heterozygous for a nonsense change in the *ARID1B* gene

– but that would only be a good candidate for causing Karol’s intellectual disability if the gene was haploinsufficient. The ExAC and GnomAD databases, mentioned earlier, provide a powerful tool for this purpose. For every gene a statistic (pLI, probability of loss of function intolerance) is calculated by comparing the frequency of loss of function variants in these healthy individuals with the expected frequency if heterozygosity for random loss of function changes was harmless. For *ARID1B* four variants were observed among the 125 748 subjects in the GnomAD database, compared to an expected figure of 91, confirming the haploinsufficiency.

One might ask why *any* gene should show haploinsufficiency. Why has natural selection not ensured a more robust level of expression? The answer in most cases is that the gene product is titrated against something else in the cell. Cells depend on innumerable interactions, between a receptor and its ligand, for example, or between a DNA-binding protein and its target, to switch some process on or off. Such switches depend on the relative concentrations of the two partners being just right, and so are sensitive to changes in the level of any one partner.

Understanding the phenotype

There is often a wide gulf between the biochemical action of a gene product within a cell and the clinical result of mutations in the gene. Many of the cases discussed throughout this book illustrate this. Even when we know both the biochemistry and the phenotype it is often difficult to explain how the one leads to the other. Why, as one example among many, should lack of the FMR1 RNA-binding protein cause intellectual disability, macro-orchidism and a long face in men with Fragile X syndrome? In some cases the connection is clear – why absence of dystrophin causes a slowly progressive neuromuscular disease, for example, or why sickle cell disease has the features it does. In many cases the connection is far from clear. This is especially true of variants that affect development. We simply don’t know enough about normal development and normal cell biology to hope to explain every case where it goes wrong.

It is important to avoid thinking naïvely about ‘the gene for...’. You would hopefully not describe your domestic freezer as a machine for ruining your stocks of frozen food – that is what happens when things go wrong. Similarly, we don’t have genes ‘for’ cystic fibrosis or muscular dystrophy. Because many human genes were discovered through studies of the diseases that result when they go wrong, there is a natural tendency to attach the disease name to the gene. It can be hard to avoid talking about, for example, ‘the Huntington disease gene’, but it is important not to be drawn into thinking that the disease defines the function of the gene.

Genotype–phenotype correlations

Establishing correlations that are accurately predictive between DNA variants and phenotypes is the Holy Grail of molecular pathology – highly desirable but seldom achieved. For the most part the chain of events between a DNA sequence change and a patient’s problems is just too long to allow neat correlations. Gene products do not act in a vacuum. They function in concert with other molecules in the cell, and in the context of the general biochemical milieu of the individual. Humans are much too varied, both in their overall genotypes and in their environments and lifestyles, to allow

the neat genotype–phenotype correlations that can often be seen in laboratory flies or mice.

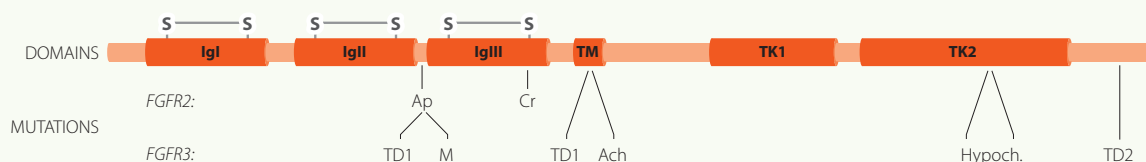
Even mendelian diseases are seldom simple when carefully examined. The mendelian diseases that have been the main topic of this book are simply the small subset of all phenotypes where the effect of a single genetic change happens not to be completely submerged by other genetic or environmental effects. But those other effects are seldom completely absent. Considering different variants within a gene, one might expect all variants causing a loss of function to have very similar effects. But the effects can differ, both between individuals with exactly the same variant and, more so, between different loss of function variants within a gene. Moreover, the loss of function is not necessarily total, and while one would expect a broad general correlation with the degree of loss, those other factors can affect the clinical consequences. An excellent article by Scriver and Waters (1999) dissects the reasons why the degree of loss of function of the phenylalanine hydroxylase enzyme in phenylketonuria, does not closely predict the degree of intellectual disability in an untreated phenylketonuric individual.

Gain of function conditions sometimes allow closer correlations. We have already seen (Figure 6.10) that the range of genotypes in such conditions is often very limited. If several different gain of function variants exist in the same gene, the effects might be distinct. The fibroblast growth factor receptors illustrate this (Box 6.2). We have already met the p.G380R variant in *FGFR3* in connection with achondroplasia (**Jenkins family, Case 14**), but other changes in the same gene or the closely related *FGFR2* produce distinctive phenotypes.

Genotype–phenotype correlation in the *FGFR* genes

The nine fibroblast growth factor proteins (FGFs) control the growth and differentiation of various mesenchymal and neuro-ectodermal cells. They act through four cell surface receptors encoded by the *FGFR1–4* genes. Each receptor consists of an extracellular portion with three immunoglobulin-like domains, a transmembrane segment, and an internal portion with two tyrosine kinase domains and a C-terminal transactivation region (see Box figure 6.1). On activation by the ligand, receptors dimerize. This triggers conformational changes that activate the intracellular tyrosine kinase. This in turn leads to activation of Stat1 signaling molecules and ultimately to cell cycle arrest. The four *FGFR* genes each encode at least 12 alternative splice isoforms. Receptors can heterodimerize as well as homodimerize, and they have different affinities for the nine FGFs. As a result they can mediate a large variety of subtly tuned responses to different combinations of FGFs in different cell types.

Several *FGFR* variants cause a gain of function. The variant receptor becomes constitutionally active and transmits its signal, to a greater or lesser extent, even in the absence of ligand, thus triggering inappropriate cell cycle arrest. Variants in the short linker between the second and third immunoglobulin domains affect the flexibility of the molecule and hence its dimerization potential. Each immunoglobulin domain is held together by an S–S bridge. Certain mis-sense changes either remove one of the cysteines involved or introduce an additional one, and it is likely that these too make the molecules more likely to dimerize even in the absence of ligand. Other frequent gain of function variants affect the transmembrane domain. Different variants in the *FGFR2* and *FGFR3* genes cause very specific syndromes with craniosynostosis or skeletal dysplasias (see Box figure 6.1 and Box table 6.1).



Box figure 6.1 – Structure of a fibroblast growth factor receptor and position of major pathogenic variants.

Ach, achondroplasia; Ap, Apert syndrome; Cr, Crouzon syndrome; Hypoch, hypochondroplasia; M, Muenke craniosynostosis; TD1 and TD2, thanatophoric dysplasia 1 and 2; TK, tyrosine kinase domain; TM, transmembrane domain. The figure shows a composite of the closely similar FGFR2 and FGFR3 proteins.

Box table 6.1 – Phenotypes and major mutations in the *FGFR2* and *FGFR3* genes

Gene	Disease (OMIM number)	Variants
<i>FGFR2</i>	Apert syndrome (101200)	p.Ser252Trp (65%), p.Pro253Arg (34%)
	Crouzon syndrome (123500)	p.Cys342Tyr or Arg (50%)
	Beare–Stevenson cutis gyrata (123790)	p.Ser372Cys (25%), p.Tyr375Cys (75%)
<i>FGFR3</i>	Achondroplasia (100800)	p.Gly380Arg (97%)
	Hypochondroplasia (146000)	p.Asn540Lys (50%), p.Asn540Thr, p.Ile538Val
	Thanatophoric dysplasia I (187600)	p.Arg248Cys (60%), p.Tyr373Cys (25%)
	Thanatophoric dysplasia II (187601)	p.Lys560Glu (100%)
	Muenke syndrome (see 134934.0014)	p.Pro250Arg (100%)

While the genotype–phenotype correlations among FGF receptor variants are unusually tight, taking a more global view reveals how different sets of related syndromes are caused by changes affecting specific multigene pathways. The concept of syndrome families has proved a productive way of exploring the relation between genotypes and phenotypes (Brunner and van Driel, 2004). Before it was known which genes were involved in specific diseases, clinicians classified diseases by clinical signs. The clinicians themselves could be classified as ‘splitters’ or ‘lumpers’. Splitters concentrated on the differences between conditions and divided them into sub-types; lumpers concentrated on the similarities and classified conditions with similar clinical signs together, arguing that there would be a single underlying mechanism. The latter approach was adopted in the 1980s by the German pediatrician J. Spranger for groups of skeletal dysplasias. He lumped many separately named syndromes together into a set of syndrome families based on the patterns of radiological and clinical findings, and not on their severity. For example, he classified hypochondroplasia (where affected individuals are short but otherwise healthy) and achondroplasia together with thanatophoric dysplasia which is lethal at birth because of severe shortening of bones and ribs. Other syndrome families he identified were the Stickler–Kniest family (characterized by varying degrees of shortening of limb bones, cleft palate and severe myopia) and the Oto–palato–digital and Larsen family (with joint deformities and dislocations and cleft palate). Now we know the molecular bases for these conditions, Spranger’s approach is seen to have been insightful. The Stickler–

Kniest family are all the result of *COL2A1* variants (see **Case 10, O'Reilly family**), while achondroplasia and related disorders are all due to changes in *FGFR3* (see *Box 6.2*).

It often happens that variants in one gene are found in most but not all patients with a particular syndrome. Where there is already some knowledge of the pathway involved, it is logical to seek the missing variants in other genes involved in the same pathway. Often it turns out that the phenotypic spectrum of patients with these new variants is somewhat different from the original cases, and this leads to clearer genotype–phenotype correlations. The overlapping clinical syndromes caused by genes in the RAS–MAPK pathway (see *Disease box 3*) are a prime example. For further discussion and examples see Donnai and Read (2003) and Brunner and van Driel (2004).

Next generation sequencing has led to a change of focus in genotype–phenotype correlations. In the past, after a researcher identified variants in a certain gene as causing a particular condition, that gene was sequenced almost exclusively in patients with that particular condition, either to explore the range of causative variants or for diagnosis. Now that we have exome or genome sequences on many thousands of individuals, it has become apparent that many variants previously thought always to cause a particular condition can be found in people who are either healthy or have a somewhat different phenotype from the ‘classical’ one. We are moving from a phenotype-led to a genotype-led view of the consequences of a variant. This has shown that quite often genotype–phenotype correlations are less specific than previously thought. A variant may be seen to cause a range of related phenotypes, depending perhaps on the genetic background, or the penetrance may be lower than previously thought. It remains true that many variants cause specific mendelian phenotypes with high penetrance, but in other cases genotype–phenotype correlations are proving looser than previously thought. The old view that variants that cause mendelian conditions have highly specific effects, while those underlying complex disorders have more variable consequences is giving way to a more nuanced view.

By way of summary, *Table 6.4* lists the extent of genotype–phenotype correlations for the clinical cases considered so far in this book.

Table 6.4 – Genotype–phenotype correlations in the clinical cases discussed so far

Case	Condition	Extent of genotype–phenotype correlation
1	Huntington disease	All cases have >36 CAG repeats in the <i>HTT</i> gene. The repeat size correlates statistically with the age of onset. Large repeats reproducibly cause juvenile-onset Huntington disease which has a rather different phenotype.
2	Cystic fibrosis	Little correlation with severity in patients with classical cystic fibrosis, but certain ‘mild’ mutations are seen in related conditions such as congenital absence of the vas deferens, nasal polyps, etc.
3	Intellectual disability	Enormous genetic heterogeneity. Mutations in any of thousands of genes can cause brain malfunction.
4	Muscular dystrophy	Frameshifting deletions almost always produce Duchenne muscular dystrophy, while in-frame deletions nearly always produce the milder Becker form.
5	Chromosomal imbalance	Phenotype depends on the size and gene content of the region involved.
6	Leber optic neuropathy	Very little correlation, even when heteroplasmy is taken into account.

Case	Condition	Extent of genotype–phenotype correlation
7	22q11 deletion	Little correlation between size of deletion and severity of phenotype.
8	Trisomy 21	Phenotype is readily recognizable but quite variable. Mosaic cases are likely to be less severe.
9	Turner syndrome	It has been claimed that behavioral problems depend on whether the single X chromosome is of maternal or paternal origin (see <i>Chapter 11</i>).
10	Stickler syndrome	Nature of <i>COL2A1</i> mutation correlates fairly well with position along the spectrum of chondrodysplasias.
11	Fragile X	Large expansions (>200 repeats) cause the classic syndrome in males; effects in females are variable. Premutation alleles (50–200 repeats) may cause tremor-ataxia syndrome, especially in males, and primary ovarian insufficiency in females.
12	Chromosomal microdeletion	Probably eventually correlations will be established; insufficient cases at present.
13	β -thalassemia	Good correlations between mutation type and β^0 or β^+ phenotype. Clinical result modified by variable persistence of fetal hemoglobin.
14	Achondroplasia	Almost perfect correlation with G380R variant of FGFR3

How do mutations arise?

Mutations arise through DNA damage or replication errors. DNA is a fairly stable molecule, but it is not immune to chemical change. Cytosine bases are liable to deaminate (lose the top -NH₂ group, see *Box 3.5*) spontaneously – the consequences of this are described in *Chapter 11*. Reactive oxygen species generated within cells as part of normal oxidative metabolism cause chemical modifications of the bases. Strand breaks are happening all the time as a result of chemical damage or natural radiation. Tobacco smoke and industrial or agricultural chemicals can act as mutagens, but the great bulk of all this damage is unrelated to industrial pollution, nuclear power plants or any other human activity. Cells have enzymes that are able to repair many types of DNA damage, so that much damage passes unnoticed, but they are not infallible. If damage is limited to one strand of the double helix, the complementary strand can be used as a template to make a correct repair. Double strand breaks, however, pose more problems, and often the repair process leaves errors in the sequence.

DNA replication is also a fallible process. The likelihood of the polymerase incorporating a mis-paired base into a growing chain is simply a function of the relative binding energies of correctly paired and mispaired bases. That thermodynamic calculation suggests an error rate orders of magnitude higher than the observed rate. The higher accuracy is the result of proof-reading and mismatch detection mechanisms. The polymerase checks the newly synthesized DNA for mispaired bases. If one is detected, the polymerase backs up, degrades a short stretch of the DNA and tries again. Interestingly, when mice were engineered to abolish the proof-reading capacity, but not the polymerase activity, of one minor DNA polymerase (specialized for replicating the mitochondrial DNA), they showed many features of accelerated aging (see Trifunovic *et al.*, 2004). One theory of aging ascribes it to accumulation of errors. After the polymerase has moved on, special enzymes excise and repair any remaining mismatched bases in newly replicated DNA. In

Chapter 7 we will see what can happen when this mismatch repair mechanism fails. Even with all this, occasional errors persist. Runs of identical bases are particularly liable to lose or gain a base by replication slippage, as described in relation to the c.35delG mutation in the *GJB2* gene (*Figure 6.5*).

Mutations can affect any piece of DNA at any time. We have focused on inherited variants in coding sequences, but non-inherited somatic mutations are increasingly being identified as causes of certain clinical conditions (see *Disease box 6*). They are also central to our understanding of cancer, as described in the following chapter. Mutations in non-coding sequences are much more numerous than those affecting coding sequences: the average human has over 70 new mutations overall, but only an average of 1.7 in coding sequence. Very little of the non-coding sequence is conserved between humans and other species, implying that it could mutate over evolutionary time without the mutants suffering any selective disadvantage. As we lack the knowledge to interpret most changes of one or a few nucleotides in non-coding sequence, it is fortunate that we can safely assume that the great majority are clinically irrelevant.

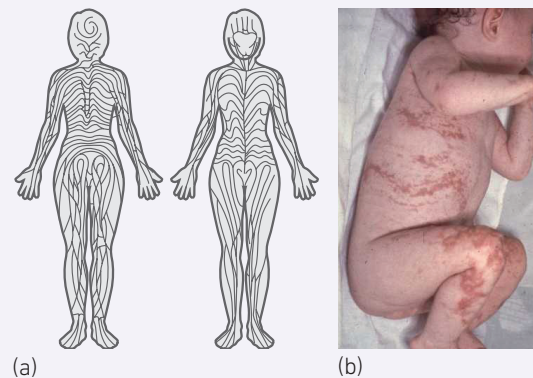
Mosaicism in clinical genetics

Mosaicism is a universal feature of humans and other multicellular organisms. As mentioned in *Chapter 1*, given the number of cell divisions required to produce and maintain the adult human body, and the likely error rate of DNA replication, every one of us must be mosaic for innumerable genetic variants. In clinical genetics mosaicism is significant in at least four situations.

1. Chromosomal mosaics, as described in *Chapter 2*. Patients can survive with chromosomal aberrations such as trisomies in mosaic form that would be lethal in full constitutional form.
2. Individuals who are mosaic for a mendelian condition that is normally seen in constitutional form. This is particularly seen with severe autosomal dominant or X-linked conditions where many cases are caused by new mutations. The difficulties such cases pose for genetic counselors are illustrated in *Chapter 1*.
3. Females who are heterozygous for an X-linked variant. They are always mosaic because of X-inactivation (see *Chapter 11*).
4. The ability to perform whole exome sequencing has revealed that a number of hitherto unexplained non-heritable clinical conditions are caused by mosaicism for DNA changes that would be lethal in constitutional form. These are discussed below.

In addition, all cancers are formally mosaic conditions, because tumors always carry acquired mutations that are not present in the constitutional genome of the patient (see *Chapter 7*).

Mosaicism is to be suspected in patients with body asymmetry or patchy phenotypes such as patchy skin pigmentation or patchy overgrowth of body parts. Patients with a variety of mosaic chromosomal constitutions often show linear skin hypo- or hyperpigmentation following Blaschko's lines (*Box figure 6.2*), as first described by the German physician Alfred Blaschko in 1901.

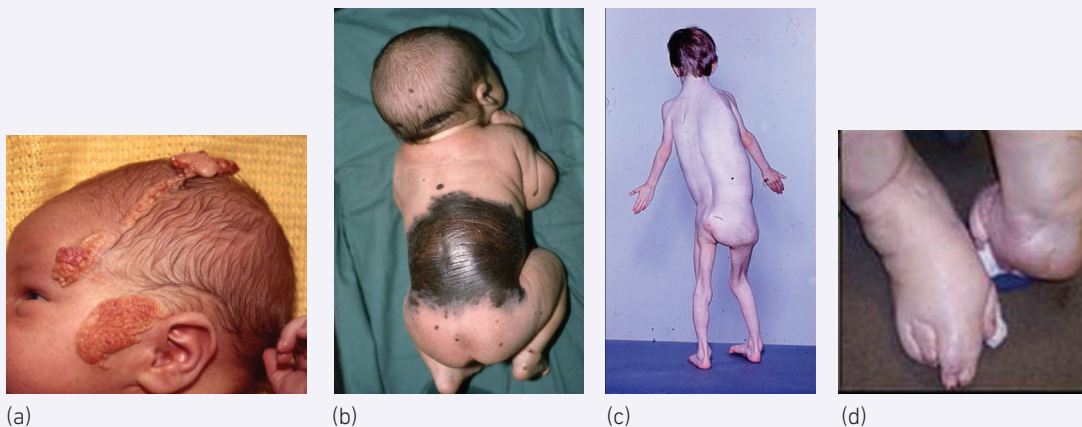


Box figure 6.2 – Mosaicism.

(a) Blaschko's lines. (b) Skin lesions following Blaschko's lines in a baby girl with incontinentia pigmenti (OMIM 308300). She is heterozygous for a mutation in the X-linked *IKBK* gene, and mosaic because of X-inactivation.

Patchy overgrowth is often the result of mosaicism for activating mutations in growth-promoting genes. The same genes, and often the same specific mutations, are frequently seen in cancerous tumors, for example (see *Box figure 6.3*):

- Patients with multiple congenital melanocytic skin nevi are often mosaic for activating mutations in the *NRAS* gene. The mutation is present in the abnormal skin, but not in normal skin. The patients may also have neurological abnormalities and again, the abnormal tissue carries the mutation.
- Non-melanocytic sebaceous nevi (Schimmelpenning syndrome) often have activating mutations in the *HRAS* or *KRAS* genes. The roles of these, and *NRAS*, in the RAS–MAPK growth-promoting signaling system was described in *Disease box 3*. Loss of function or mild gain of function mutations are compatible with life and are seen in the hereditary conditions described in *Disease box 3*. More strongly activating mutations, such as those here, can only survive in mosaic form. All three genes show frequent strongly activating mutations in a variety of tumors.
- Proteus syndrome, a non-heritable condition with overgrowth and asymmetry of limbs, epidermal nevi, connective tissue nevi, lipomas and vascular malformations is usually caused by mosaic activating mutations in the *AKT1* gene. A similar condition called CLOVES syndrome (congenital lipomatous overgrowth, vascular malformations, epidermal nevi and spinal/skeletal anomalies) is caused by mosaic mutations in the *PI3KCA* gene. The products of these genes form part of the PI3k–Akt–mTOR pathway by which growth-promoting signals from outside the cell activate gene transcription in the nucleus. Both genes are frequently mutated in cancer.



Box figure 6.3 – Patchy overgrowth due to mosaicism for an activating mutation in a growth-promoting gene.

(a) Sebaceous nevus, mosaic for a mutation in *HRAS*. (b) Melanocytic nevi due to mosaicism for a mutation in *NRAS*. (c) Overgrowth in Proteus syndrome due to mosaicism for a mutation in *AKT1*. (d) Overgrowth and deformation of parts of feet in CLOVES syndrome due to mosaicism for a mutation in *PIK3CA*.

Full constitutional forms of any of these mutations would likely be incompatible with life.

Sometimes mosaicism is present only in some tissues. This is the case with mosaic 12p tetrasomy (four copies of the short arm of chromosome 12) that causes Pallister–Killian syndrome (OMIM 601803; *Box figure 6.4*). The abnormal cells are seen in skin but not in standard blood samples (although they are detectable at very low levels in blood by deep sequencing). It is likely that in blood, which has a rapid cell turnover, abnormal cells are at a disadvantage and the normal cell lines come to predominate.

Tissue-limited mosaicism can be a particular problem in prenatal diagnosis. The procedure of chorion villus biopsy (*Chapter 14*) provides the cytogeneticist with a sample of the placenta for analysis. Although the placenta is a fetal tissue, if it shows mosaicism there is great uncertainty whether this might be confined to the placenta or whether the fetus itself might also be mosaic. If the fetus is mosaic, predicting the phenotype is difficult. It will lie somewhere along the spectrum between normal and the full constitutional phenotype, but exactly where along the spectrum is unpredictable, because it depends on the proportion of abnormal cells in different tissues and organs.



(a)



(b)

Box figure 6.4 – Pallister–Killian syndrome.

(a) Facial appearance; note lack of hair on temples, short upslanting palpebral fissures and short nose. (b) Karyotype; note small extra metacentric chromosome (arrowed) consisting of two extra copies of 12p.

DISEASE BOX 6 – continued

6.5. References

- Brunner HG and van Driel MA** (2004) From syndrome families to functional genomics. *Nat. Rev. Genet.* **5**: 545–551.
- Cleary JD and Ranum LPW** (2013) Repeat-associated non-ATG (RAN) translation in neurological disease. *Hum. Molec. Genet.* **22(R1)**: R45–R51.
- Donnai D and Read AP** (2003) How clinicians add to knowledge of development. *Lancet*, **362**: 477–484.
- Gordon CT and Lyonnet S** (2014) Enhancer mutations and phenotype modularity. *Nature Genetics*, **46**: 3–4.
- Goriely A, Hansen RMS, Taylor IB, et al.** (2009) Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. *Nature Genetics*, **41**: 1247–1252.
- Holbrook JA, Neu-Yilik G, Hentze MW and Kulozik AE** (2004) Nonsense-mediated decay approaches the clinic. *Nature Genetics*, **36**: 801–808.
- MacArthur DG, Balasubramanian S, Frankish A, et al.** (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**: 823–828.
- Manfredi C, Tindall JM, Hong JS and Sorscher EJ** (2019) Making precision medicine personal for cystic fibrosis. *Science*, **365**: 220–221.
- Perry GH, Dominy NJ, Claw KG, et al.** (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, **39**: 1256–1260.
- Scriver C and Waters PJ** (1999) Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet.* **15**: 267–272.

Snead MP and Yates JRW (1999) Clinical and molecular genetics of Stickler syndrome. *J. Med. Genet.* **36**: 353–359.

Tarpey PS, Smith R, Pleasance E, et al. (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nature Genetics*, **41**: 535–543.

Trifunovic A, Wredenberg A, Falkenberg M, et al. (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature*, **429**: 357–359.

Useful websites

For the nomenclature of mutations see <http://varnomen.hgvs.org/>

POLYPHEN (<http://genetics.bwh.harvard.edu/pph>) and PROVEAN (<http://provean.jcvi.org/index.php>) are web-based programs that attempt to marshal all available data to decide whether an amino acid substitution in a protein is likely to be damaging.

6.6. Self-assessment questions

- (1) Cystic fibrosis is caused by the absence from apical cell membranes of functional chloride ion channels that are encoded by the *CFTR* gene. Which of the following mutations might be a cause of cystic fibrosis?
 - (a) Deletion of the *CFTR* gene.
 - (b) A mutation in the gene encoding arginine-specific tRNA that causes the protein synthesis machinery to incorporate arginine in growing polypeptide chains in response to serine codons. Substitution of arginine for serine in the CFTR protein makes it non-functional.
 - (c) A mutation in the promoter of the *CFTR* gene that abolishes its ability to recruit the RNA polymerase machinery.
 - (d) A mutation in the coding sequence of the *CFTR* gene that replaces an essential serine with a non-functional arginine.
 - (e) A mutation in one of the small non-coding RNA molecules in the spliceosome that causes the splicing machinery to treat GA...AG as the signal for the start and end of introns, instead of GT...AG.
 - (f) A mutation in the RNA polymerase II gene that renders the polymerase non-functional.
 - (g) A mutation in the coding sequence of the *CFTR* gene that causes the ion channel to transport excessive quantities of chloride ions.
- (2) Box 6.3 shows the sequence of two exons (upper case), with flanking intron sequence (lower case), as amplified by PCR for mutation detection in the *PAX3* gene. For exon 1, only part of the 5'UT is included in the PCR product. Nucleotides are numbered as in the cDNA (with the first nucleotide of the initiator codon numbered +1), and the protein sequence is shown using single-letter codes (see Box 3.6).

For the following eight mutations, a short sequence is given, with the number of the first nucleotide shown. The changed nucleotide (or the first changed one if

several change) is underlined. For each mutation, give the correct nomenclature (a) as a DNA change (b) (where appropriate) as a protein change.

```

c.15   CGGCGCTGTGGCCAGGATGATGC
c.43   GGCCCGGGGTAGAACTACCCGCG
c.78   GCTGGAAGTTAAGGGAGGGCCTC
c.86   TGTCCACTCCACTCGGCCAGGGC
c.121  CTCGGCGGCGTTTTATCAACGGC
c.130  GTTTTGATCAACGGCAGGCCGCT
c.180  GGAGAGGCCCACCACGGCATCC
c.248  TCTCCGAGATCCTGTGCAGGTAC
c.283  TCCATTCCTGGTGCCATCGGCGG

```

- (3) Referring to the *PAX3* sequence in Box 6.3, write out the mutant sequence of each of the following, formatted as in Question 2:

```

p.N47H
c.247_248ins(C)
c.185_202del
p.E61X
c.85+6G>T
c.86-2A>G
p.V29M

```

Partial sequence of *PAX3* gene

Exon 1 (451 nt)

```

..CCGTTTCGC CCTTCACCTG GATATAATTT CCGAGCGAAG TGCCCCCAGG
1  ATG ACC ACG CTG GCC GGC GCT GTG CCC AGG ATG ATG CGG CCG GGC CCG GGG
1  M   T   T   L   A   G   A   V   P   R   M   M   R   P   G   P   G
52 CAG AAC TAC CCG CGT AGC GGG TTC CCG CTG GAA Ggtaaggagg gcctcagcgc..
18 Q   N   Y   P   R   S   G   F   P   L   E

```

Exon 2

```

..tgacttttcc cttgcttctc tttttcacct tcccacag
86 TG TCC ACT CCC CTC GGC CAG GGC CGC GTC AAC CAG CTC GGC GGC GTT TTT
29 V   S   T   P   L   G   Q   G   R   V   N   Q   L   G   G   V   F
136 ATC AAC GGC AGG CCG CTG CCC AAC CAC ATC CGC CAC AAG ATC GTG GAG ATG
46 I   N   G   R   P   L   P   N   H   I   R   H   K   I   V   E   M
187 GCC CAC CAC GGC ATC CGG CCC TGC GTC ATC TCG CGC CAG CTG CGC GTG TCC
63 A   H   H   G   I   R   P   C   V   I   S   R   Q   L   R   V   S
238 CAC GGC TGC GTC TCC AAG ATC CTG TGC AGG TAC CAG GAG ACT GGC TCC ATA
80 H   G   C   V   S   K   I   L   C   R   Y   Q   E   T   G   S   I
289 CGT CCT GGT GCC ATC GGC GGC AGC AAG CCC AAG gtagcgggc gggccttgcc..
97 R   P   G   A   I   G   G   S   K   P   K

```

- (4) Referring to the *PAX3* sequence in Box 6.3, for each of the following mutations, select one of the following options:
- (a) Synonymous
 - (b) Mis-sense
 - (c) Nonsense
 - (d) Frameshift
 - (e) Non-frameshifting insertion / deletion
 - (f) Splice-site mutation
 - (g) Initiator codon
 - (h) Terminator codon
 - (i) Intronic
- (1) c.85+1G>A
 - (2) c.86T>A
 - (3) c.86-18T>G
 - (4) c.101insGCC
 - (5) c.118C>T
 - (6) c.172_173delAA
 - (7) c.216C>G
 - (8) c.270C>G
- (5) The effects of a mutation can be studied at the protein level as well as by DNA sequencing. If a suitable antibody is available, mutations can be classified into CRM⁺ and CRM⁻. CRM means cross-reacting material. Classify each of the types of mutation in the list on page 146 in this way, commenting on cases where the result is hard to predict.
- (6) A student wrote the following answer to a question about the genetics of cystic fibrosis:
- ‘The gene for cystic fibrosis is recessive. If you have two copies of the gene you have cystic fibrosis, but if you only have one copy you are just the same as somebody who doesn’t have any copy’.

Comment on this.

[Hints on questions 1, 2, 3 and 4 are provided in the *Guidance* section at the back of the book.]

07

Is cancer genetic?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe tumorigenesis as an evolutionary process within an individual
- Define oncogenes and tumor suppressor genes, giving examples
- Describe the types of genomic instability found in cancer cells and the roles of cell cycle checkpoints in avoiding these
- List the essential capabilities of malignant tumors and describe the types of somatic genetic change that lead to their development, including activation of oncogenes or inactivation of tumor suppressor genes by somatic mutation, deletions leading to loss of heterozygosity, and chromosomal rearrangements leading to fusion genes
- Describe at least two inherited cancer syndromes and discuss their relationship to common sporadic cancers
- Describe the roles of genetics in diagnosis, treatment and prevention of cancer

7.1. Case studies

CASE 15 TIERNEY FAMILY

- 4-year-old boy, Jason
- Pale with extensive bruising and tachycardia
- ? Acute lymphocytic leukemia

175

190

261

395

Jason is a previously fit and well 4-year-old boy who presented with a two-week history of extensive bruising and a pain in his back. His mother took him to the family doctor who noted that he was pale and tachycardic (a fast heart rate) without obvious fever. Blood tests revealed a low hemoglobin level (anemia), a raised white blood cell count ($90 \times 10^9/l$, normal level $4-11 \times 10^9/l$) and low platelet level. Analysis of the blood count showed a high level of lymphocytes. These results raised the possibility of a diagnosis of childhood acute lymphocytic leukemia.

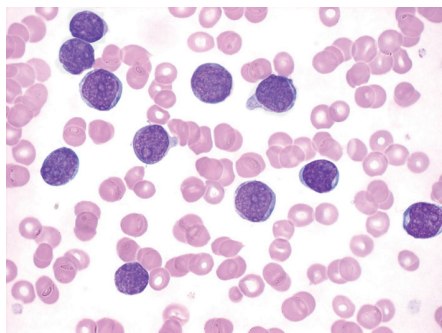


Figure 7.1 – Typical appearance of acute lymphocytic leukemia.

Small blasts with high nuclear-cytoplasmic ratio, some with prominent nucleoli. Photo. courtesy of Dr John Yin, Manchester Royal Infirmary.

CASE 16 WILSON FAMILY

- Family history of breast cancer

176**191****395**

Wendy Wilson saw a television program about breast cancer running in families. It raised her concerns about her own family history. She had mentioned her worries some years before to her family doctor but had been reassured. Her mother Wanda had developed the disease at the age of 42 years and sadly died when she was 44 years old. Her mother's sister Amy who lived in New Zealand also developed breast cancer in her 40s, but after surgery and chemotherapy was well 7 years later. The previous Christmas Wendy had received a card from an elderly great-aunt in which she mentioned that one of her grandsons was undergoing treatment for breast cancer as well, and expressed her shock that a man could be affected. Wendy got in touch with her brother William and sister Veronica and they decided they should look into things in more detail. Veronica was the family genealogist and before long had contacted several relatives they had lost touch with. They found one of Wanda's cousins had died at a young age of breast cancer. The television program had mentioned that tests for familial breast cancer were available through genetic clinics, so Wendy made an appointment with her family doctor to ask for a referral. At the surgery Wendy gave as many details as possible to the doctor and he consulted the online guidelines provided by the genetic center. Wendy's family fulfilled the criteria for a high risk and so a referral was made. Before the appointment the genetic counselor from the center contacted Wendy to draw up a family tree. She also asked Wendy for details of her mother and where she was treated so that her medical records could be obtained to confirm the details of her illness.

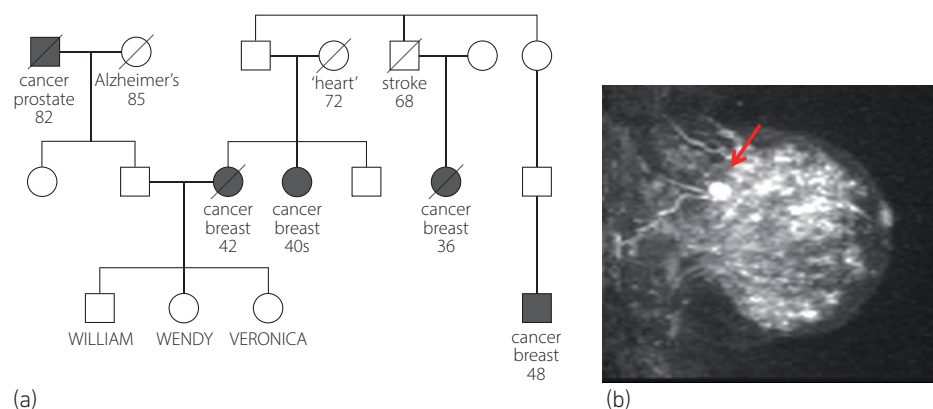


Figure 7.2 – (a) Pedigree of the Wilson family, showing types of cancer and age at diagnosis. (b) Carcinoma of the breast detected in a 40-year-old woman by magnetic resonance imaging. This lady had a *BRCA1* mutation. Photo. courtesy of Dr Gareth Evans, St Mary's Hospital, Manchester.

CASE 17 XENAKIS FAMILY

- Family history of bowel problems
- ? Familial adenomatous polyposis

176**195****395**

Christos was one of three children born in Cyprus in the 1960s to Xavier and Demi Xenakis. Xavier started having bowel symptoms at the age of 41 but didn't go to see his doctor until he became really unwell. At the hospital bowel cancer with liver metastases was diagnosed and only palliative treatment was possible. He died later that year, and soon afterwards Christos moved to Seattle to open his own restaurant with his wife and young son and daughter. Demi came to live with them. Life was very busy but the restaurant did well. At an insurance medical examination some years later, Christos mentioned

that he had recently noticed some rectal bleeding which he thought might be due to hemorrhoids. Because of the family history the doctor recommended that Christos have a sigmoidoscopy, and to everyone's concern this revealed multiple polyps.

The surgeon explained that this indicated Christos had familial adenomatous polyposis (FAP) and that the only sensible treatment was to remove the colon, since it was inevitable that cancer would develop in one or more of the polyps. While Christos was still in hospital, the surgeon recommended he be referred to the genetic clinic when he was recovered to find out about risks for his children, and also about screening tests that they could have in due course. After a few months the family attended the genetic clinic. The counselor explained dominant inheritance, and they worked out that each child was at 50% risk of having inherited the causative gene. They learned that genetic testing would be possible if a pathogenic change was found in Christos and that, for children who were at risk or who were found to be carriers of the variant, regular sigmoidoscopies were offered from 10 years of age.

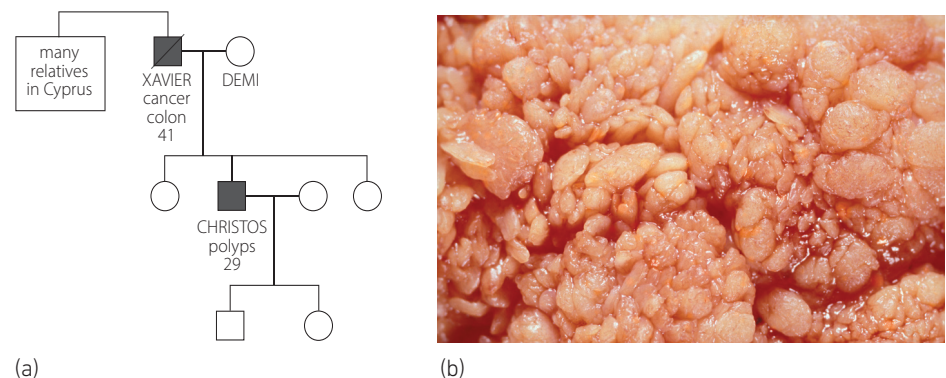


Figure 7.3 – Familial adenomatous polyposis coli.

(a) Pedigree of the Xenakis family. (b) Part of a surgically resected colon with polyps. Photo. courtesy of Medical Illustration Department, Manchester Royal Infirmary.

7.2. Science toolkit

Natural selection and the evolution of cancer

Imagine an isolated wood in which a population of voles lives. One vole acquires a heritable mutation that makes it able to reproduce faster than the other voles. When you visit the wood 100 vole generations later you would expect to find most of the voles were descendants of that one faster breeding mutant. Exactly the same simple Darwinian argument applies to the cells of your body. Cell division is under genetic control. If one cell acquires a mutation that makes it divide faster than others then, all things being equal, its descendants will take over your body. Thus cancer is not a specific disease with one cause, and one cure waiting to be discovered, it is simply the natural end-point of evolution within the body of any multicellular organism.

Fortunately, all things are not equal. Multicellular organisms could not survive if they did not have mechanisms to control and repress the evolution of their somatic cells, at least until their reproductive life is over. Over the billion years through which they have been evolving, multicellular organisms have developed sophisticated and many-layered defenses against rogue cells. Cell growth is tightly controlled. Most somatic cells do not have the capability of dividing indefinitely. Any cell that behaves inappropriately and whose antisocial behavior cannot be remedied is made to kill itself (**apoptosis**). Classical studies of the age-specific incidence of the common epithelial cancers suggested that four to seven independent events were needed to generate a malignant tumor. This is consistent with the idea that several independent control systems need to be inactivated before a cell can become malignant.

Overcoming the defenses

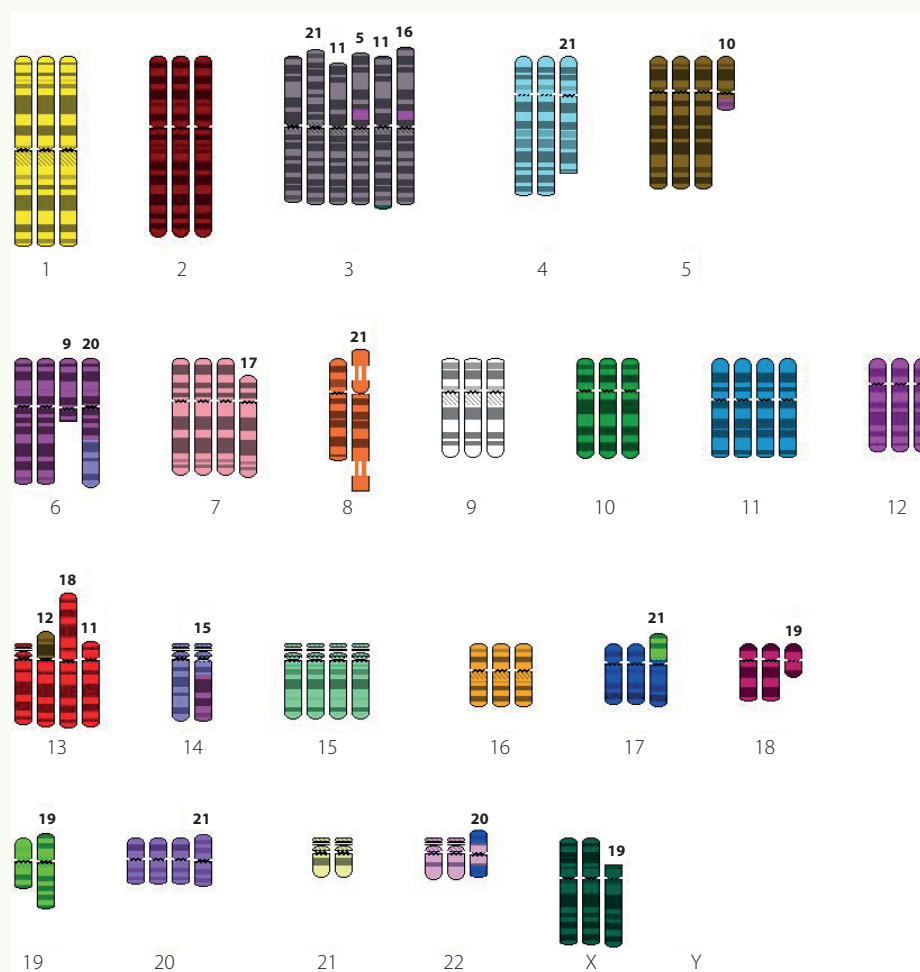
Given typical mutation rates, the chance of any cell in a person's body sequentially acquiring six specific mutations would appear to be negligibly low. Assuming a mutation rate of 10^{-6} per gene per cell, the chance of a cell picking up six successive specific mutations is 10^{-36} . There are only of the order of 10^{14} cells in a human body, so it would appear that the defenses against cancer are impregnable. Since we know that one person in three will nevertheless develop cancer, there has to be some way round the defenses.

The trick is that the mutations early in the process must somehow greatly increase the probability of later mutations. They can do this in either of two ways:

- They can give the cell a growth advantage. If the mutant cell can generate 1000 mutant daughter cells, the chance that one mutant cell will acquire the next mutation has increased 1000-fold. Pathologists have long known that tumors develop through stages that are marked by increasing growth potential, and that rapidly dividing tissues are the ones most likely to develop tumors. It is believed that the founding cells of many tumors have stem cell-like properties, so that they already have unusual growth potential.
- They can increase the general mutation rate by destabilizing the genome. Genomic instability is a key feature of almost all tumor cells (*Box 7.1*). Most tumor cells have bizarre karyotypes with grossly abnormal numbers of chromosomes and many structural rearrangements. Possible causes include erosion of telomeres (see below). An extreme form ('chromothripsis') in which one particular chromosome shows dozens of rearrangements is seen in 2–3% of cancers. The instability extends to the DNA sequence and regulatory systems.

As a result of these changes, the early stages of tumorigenesis are likely to produce a growing population of cells with a great variety of random mutations, making fertile ground for subsequent developments. One of the great challenges in understanding tumorigenesis is to distinguish **driver mutations**, that are causally implicated in the process, from the many incidental but irrelevant **passenger mutations**. Genes housing driver mutations can be classified into oncogenes and tumor suppressor (TS) genes. Oncogenes act to promote cell growth, TS genes restrain growth. Tumorigenesis is driven by activation of oncogenes (gain of function) or homozygous inactivation of TS genes (loss of function), as described below.

Cancer cells typically have grossly abnormal karyotypes. Gains, losses and rearrangements of chromosomes are the result of breakdown of one or other of the systems that should ensure chromosomal integrity. Changes may be drivers (*Table 7.2*) or passengers in the evolution of tumors. *Box figure 7.1* shows a typical example, analyzed using multicolor chromosome painting (*Section 4.4*). Nowadays the analysis would be based on a whole genome sequence but, as in *Chapter 2*, we show a karyotype here to make the changes more obvious.



21 cells were analyzed. The chromosome number varied between 67 and 71. The numbers above individual chromosomes are the number of cells having each specified abnormality. Reproduced from https://ftp.ncbi.nlm.nih.gov/sky-cgh/DATA/SKY-images/T_Ried_HT_29_karyotype_5.gif

Tumor genomes also usually show large numbers of small-scale sequence changes reflecting a high rate of mutation. Sequencing the genomes of individual tumors often reveals one or more characteristic signatures of particular mutational processes, for example, of tobacco carcinogenesis in lung tumors or UV mutagenesis in skin melanomas. Defective proof-reading of newly replicated DNA is revealed by microsatellite instability (see below).

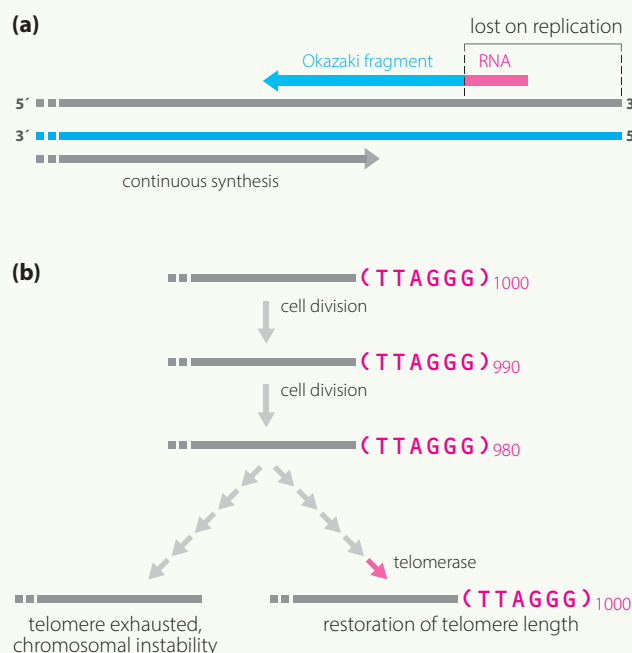
A third level of instability is epigenetic instability. Tumors show highly abnormal patterns of gene expression, because mutations have affected genes that control the regulatory systems described in Chapter 11.

BOX 7.1 – continued

Living for ever: the importance of telomeres

Cancer cells are immortal. HeLa cells, a standard laboratory cell culture workhorse, have been growing vigorously in culture ever since their unfortunate donor, Henrietta Lacks, died of cervical cancer in 1951 (see Skloot, 2010). Ordinary human cells won't do this. They grow in culture for a few dozen divisions, but then they stop growing, a condition called senescence. Cells with certain cancer-related mutations (in *TP53* and *RB1*, see below) are able to avoid senescence. After some further divisions they reach a stage called crisis. Crisis is marked by death of the great majority of the cells. The few survivors have acquired extensive chromosomal rearrangements and the immortality of cancer cells.

These events are related to a problem with replicating chromosome ends. Recall that the two strands of the double helix are anti-parallel, and that DNA strands can only grow in the 5'→3' direction (Box 3.2). One strand of the double helix has a free 5' end. When this strand is used as a template for synthesis of a complementary strand, the polymerase moves towards the end of the strand and there is no problem extending the new strand up to the end of the template (Box figure 7.2a). The other, with a free 3' end, has problems. When this strand is used as a template, the polymerase moves in the direction away from the end towards the interior of the template, against the direction of movement of the replication fork. The new strand is made discontinuously in 100–200 nt segments (**Okazaki fragments**). Each fragment starts with a short (10 nt) RNA primer. The last primer does not necessarily start on the very last nucleotide of the chromosome, and even if it did, the 10 nt corresponding to the primer would be lost as the primer is removed when the Okazaki fragments are processed and ligated together. Thus a linear DNA double helix inevitably loses some sequence at the 3' end each time it is replicated.



Box figure 7.2 – (a) DNA replication mechanisms cannot replicate the extreme 3' end of a molecule. (b) Telomeres of human chromosomes contain tandemly repeated TTAGGG units. Some repeats are lost each time a cell replicates its DNA. Continued loss in cultured cells leads to chromosomal instability. In the germ line and in cancer cells telomerase can restore telomeres to full length.

BOX 7.2

Bacteria and mitochondria solve this problem by having circular chromosomes; eukaryotic cells use **telomeres**. Human chromosomes end with about 10 kb of repetitive sequence, (TTAGGG)_n. With each round of cell division the telomere shortens by 50–100 nt. Within limits, this doesn't matter because the telomeric repeats contain no genetic information. Repeated division will lead to complete loss of telomeres. One function of the telomeres is to protect chromosome ends against DNA repair mechanisms that recognize broken DNA ends and try to join them together (*Box figure 7.2b*). This is what happens to cells in the post-senescence crisis, and is a likely cause of the abnormalities shown in *Box figure 7.1*.

Certain cells possess an enzyme, **telomerase**, that can restore telomeres to full length. Telomerase adds TTAGGG units using its own inbuilt RNA template, and is therefore not dependent on an external template. Telomerase restores telomere length in the germ line of each generation. Most normal somatic cells lack telomerase (the gene is there but not expressed), but most cancer cells possess it. Reactivation of telomerase is an important step in acquiring the capability to divide indefinitely. Equally, of course, telomerase looks like a promising target for an anticancer drug. The results of initial trials have been disappointing, but some are still ongoing.

Oncogenes

Oncogenes were first discovered in acute transforming retroviruses isolated from various animal tumors. Normal retroviruses have an RNA genome containing just three transcription units: *gag* (group-specific antigen), *pol* (polymerase) and *env* (envelope), each of which encodes several proteins. When the genomes of acute transforming retroviruses were analyzed in the 1970s they were seen to contain an extra gene, the oncogene. Initial excitement that all cancer might be the result of infection by oncogene-carrying viruses, and thus perhaps preventable by vaccination, died down with the discovery that viral oncogenes were copies of normal cellular genes that had been accidentally incorporated into the viral genome. Clearly there must be some difference between the oncogene in the virus, which transforms cells (causes cells in culture to acquire some of the properties of tumor cells), and the normal cellular version, which does not. Viral oncogenes are activated versions of the cellular genes (strictly called **proto-oncogenes**, but usually called simply oncogenes).

Several dozen oncogenes were identified through studies of such viruses, and named after the animal tumors from which they were isolated (see *Table 7.1*). A breakthrough in molecular understanding of carcinogenesis was achieved in the early 1980s when the normal functions of cellular oncogenes were identified. As the table shows, the normal, non-activated versions of these genes have roles in the control of cell proliferation. Many are tyrosine kinases. A frequent method of regulating the activity of proteins that act in cell signaling systems is reversible phosphorylation. Specific kinase enzymes activate signaling systems by attaching phosphate groups to hydroxyl groups on tyrosine, or sometimes serine, residues in signaling proteins (*Figure 7.4*). Given their normal function, it is entirely understandable that pathogenically activated versions of these genes should be oncogenic.

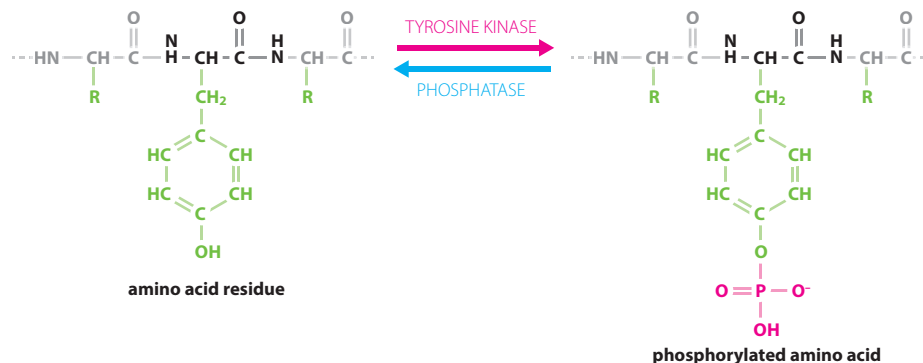


Figure 7.4 – The activity of many proteins is regulated by reversible phosphorylation of specific tyrosine residues. Many oncogenes are tyrosine kinases.

Table 7.1 – Viral oncogenes and their cellular counterparts

Gene	Animal source	Cellular proto-oncogene	
		Location	Function
<i>ABL1</i>	Abelson murine leukemia	9q34	Signal transduction (tyrosine kinase)
<i>ERBB2</i>	Avian erythroblastic leukemia	17q21	Signal transduction (receptor tyrosine kinase)
<i>FES</i>	Feline sarcoma virus	15q26	Signal transduction (tyrosine kinase)
<i>FMS</i>	Friend murine leukemia	5q33	M-CSF (receptor tyrosine kinase)
<i>HRAS</i>	Harvey rat sarcoma	11p15.5	Signal transduction (small GTPase)
<i>KRAS</i>	Kirsten mouse sarcoma	12p12	Signal transduction (small GTPase)
<i>MYB</i>	Avian myeloblastosis	6q22	Transcription control (nuclear protein)
<i>MYC</i>	Avian myelomacytosis	8q24	Transcription control (nuclear protein)
<i>SIS</i>	Simian sarcoma virus	22q12	Platelet-derived growth factor B chain
<i>SRC</i>	Rous chicken sarcoma	20q12	Signal transduction (tyrosine kinase)

See *Disease box 3* for more detail on the function of the RAS family of oncogenes.

Activation involves a gain of function. This can be achieved in a variety of ways.

- *Point mutations* – as always, a gain of function requires a specific mutation. For example, bladder cancers may have the mutation p.Gly12Val in the *HRAS* oncogene. The three human *RAS* genes (*KRAS*, *HRAS*, *NRAS*) encode small GTPases that activate the RAS–MAPK intracellular signaling cascade (see *Disease box 3*). Gain of function mutations produce hyperactive molecules that trigger excessive expression of the target genes.

- *Amplification* – some tumors contain many extra copies of oncogenes, sometimes in the form of small extra chromosomes, sometimes as duplications within a chromosome. The *MYC* oncogene, for example, is frequently amplified in tumors.
- *Chromosomal rearrangements* can bring together exons of two distant genes to make a novel chimeric gene. As mentioned above, cancer cells usually have many chromosomal abnormalities. Much painstaking research has been dedicated to identifying changes that are specific to particular tumor types, and distinguishing them from the large numbers of random changes. *Table 7.2* gives examples (mostly from leukemia, which is easier to study than solid tumors), and *Disease box 7* illustrates one well-known case. These rearrangements are exceptionally interesting because sequencing the breakpoints reveals the chimeric genes, and this has been a route to discovery of many oncogenes. Some genes are involved in many different rearrangements – the *MLL (KMT2A)* gene at 11q23 has been noted with over 30 different fusion partners in leukemia patients. Tests for specific chromosomal rearrangements are an important part of molecular diagnosis of cancer (see *Disease box 7*). Often defining the chimeric oncogene provides a guide to prognosis and treatment, as we will see in the case of **Jason Tierney (Case 15)**. Research into the function of the chimeric gene also provides an important entry into understanding the biology of the tumor.
- *A chromosomal rearrangement can up-regulate expression of an oncogene* by moving it into a transcriptionally highly active region of chromatin. The classic case is Burkitt's lymphoma, a childhood tumor affecting particularly the jaw and found mainly in Africa. Incidence is associated with infection by malaria and Epstein–Barr virus. Tumor cells have a characteristic somatically acquired balanced reciprocal translocation, $t(8;14)(q24;q32)$, *Figure 7.5*. The effect of the translocation is to move the *MYC* oncogene from chromosome 8 to the neighborhood of the *IGH* immunoglobulin heavy chain gene on chromosome 14. Unlike most tumor-specific rearrangements, the move does not create a chimeric gene, but it places the *MYC* gene under the influence of a powerful B-lymphocyte-specific enhancer (see *Section 6.2* for a description of enhancers). Thus B lymphocytes, but not other cells with the translocation, greatly over-express *MYC*. The result is a lymphoma. Sometimes an alternative translocation moves *MYC* to the immunoglobulin light chain gene regions on chromosomes 2 or 22.

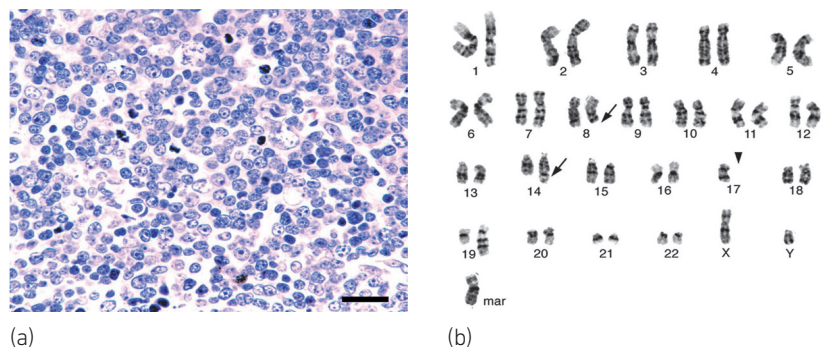


Figure 7.5 – Burkitt's lymphoma.

(a) Histology, and (b) a karyotype showing the characteristic 8;14 translocation. Additional chromosome abnormalities are also present, as is usually the case in neoplasia. Reproduced from *Molecular Cancer*, **2**: 30; © 2003 Duensing *et al.*; licensee BioMed Central Ltd.

Table 7.2 – Examples of tumor-specific balanced chromosomal rearrangements that create chimeric genes

Rearrangement	Genes	Disease
t(1;22)(p13;q13)	<i>RBM15 / MKL1</i>	Acute megakaryoblastic leukemia (FAB–M7)
t(2;13)(q35;q14)	<i>PAX3 / FKHR</i>	Alveolar rhabdomyosarcoma
t(3;8)(p21;q12)	<i>PLAG1 / CTNNB1</i>	Pleomorphic salivary gland adenoma
inv(3)(q21q26)	<i>RPN1 / EVI1</i>	AML without maturation (FAB–M1)
t(4;11)(q21;q23)	<i>MLL / AFF</i>	ALL/lymphoblastic lymphoma
t(6;11)(q27;q23)	<i>MLL / MLLT4</i>	AMML (FAB–M4)
t(9;11)(p22;q23)	<i>MLL / AF9</i>	ALL/lymphoblastic lymphoma
t(11;19)(q23;p13)	<i>MLL / MLLT1</i>	ALL/lymphoblastic lymphoma
t(7;11)(p15;p15)	<i>NUP98 / HOXA11, HOXA13, HOXA9</i>	AML with maturation (FAB–M2)
t(9;22)(q34;q11)	<i>BCR / ABL1</i>	Chronic myeloid leukemia
t(11;14)(q13;q32)	<i>IGH / CCND1</i>	Chronic lymphocytic leukemia, Mantle cell lymphoma
t(15;17)(q22;q12)	<i>PML / RARA</i>	Acute promyelocytic leukemia (FAB–M3)
t(12;16)(q13;p11)	<i>FUS / DDIT3</i>	Liposarcoma
inv(16)(p13q22)	<i>CBFB / MYH11</i>	AMML (FAB–M4)
t(X;18)(p11;q11)	<i>SS18 / SSX1,SSX2,SSX4</i>	Synovial sarcoma
t(14;18)(q32;q21)	<i>IGH / BCL2</i>	Follicular lymphoma
t(12;21)(p13;q22)	<i>ETV6 (TEL) / RUNX1 (AML1)</i>	ALL/lymphoblastic lymphoma
t(8;21)(q22;q22)	<i>RUNX1 / ETO</i>	AML with maturation (FAB–M2)

See <https://mitelmandatabase.isb-cgc.org> for a large database of rearrangements established by Dr Felix Mitelman. Rearrangements have been particularly defined in leukemias and lymphomas because in these conditions the cells are usually a single clone and are more amenable to cytogenetic analysis than those in solid tumors. Whole genome sequencing now allows similar analyses in solid tumors. ALL, acute lymphoblastic leukemia; AML, acute myeloblastic leukemia; AMML, acute myelomonocytic leukemia.

Tumor suppressor genes

Tumor suppressor (TS) genes were discovered through studies of the rare forms of cancer that are familial. Their existence was hypothesized in 1971 by Alfred Knudson, based on his studies of retinoblastoma (RB), a rare childhood tumor of the retina. RB can be sporadic or familial (autosomal dominant). Knudson's analysis of the age dependence of incidence led him to hypothesize that the early rate-determining steps of carcinogenesis required the founder cell of a tumor to suffer two 'hits' – these might be simple mutations or some other genetic change. In sporadic cancers both hits were chance events, each with a low probability. In the familial version one hit was inherited: every cell of a susceptible person carried one hit, so it only required one cell in the target tissue to suffer one further hit for the tumor to develop (*Figure 7.6*). The familial susceptibility was inherited from one parent as a dominant trait, but the cellular phenotype that allows a cell to found a tumor is recessive, requiring two hits (this is a useful reminder that dominance and recessiveness are properties of phenotypes and

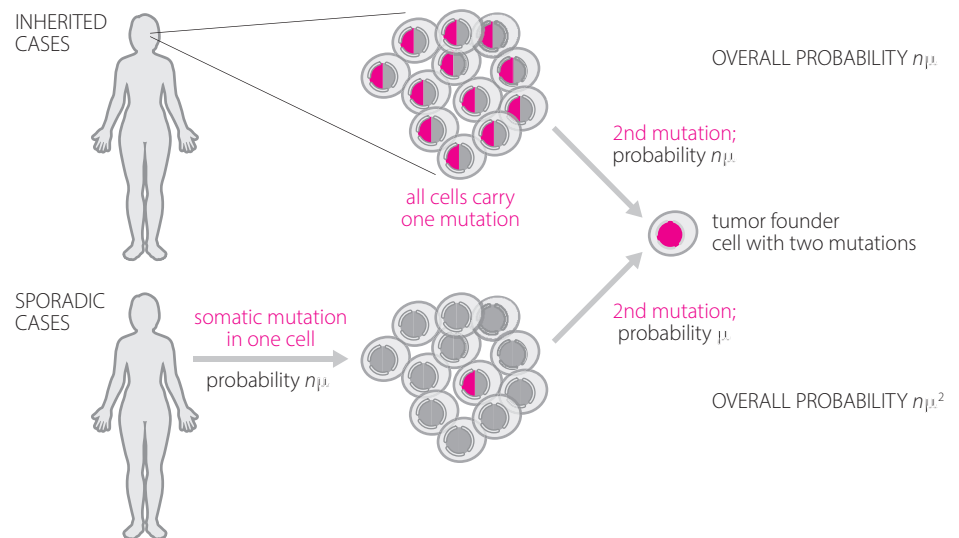


Figure 7.6 – The relation between sporadic and inherited forms of the same tumor.

The target tissue contains n cells and the chance of one cell suffering a loss of function mutation in the tumor suppressor gene is μ . For example, if the target cell population comprised 100 000 retinoblast precursor cells, and if the mutation rate was 10^{-5} per cell, then an infant heterozygous for a constitutional mutation would be highly likely to develop the family tumor, whereas only 1 in 100 000 infants without the inherited susceptibility would develop it (note that these calculations are to illustrate the principle and are not intended to reflect all the complexities of the real situation).

not of genes or variants). Molecular work in the 1980s confirmed Knudson's hypothesis. In familial RB second (somatic) hits were identified, and they always worked so as to delete or mutate the allele inherited from the unaffected parent.

This work confirmed the existence of TS genes, explained how cancers could be familial, and suggested two ways of identifying TS genes.

- Identifying chromosomal regions with deletions in sporadic tumors. Frequently the second hit is deletion of the chromosomal segment containing the wild-type allele of the TS gene.
- Using family linkage studies to identify the chromosomal location of the genes mutated in familial cancers, as described in *Chapter 8*.

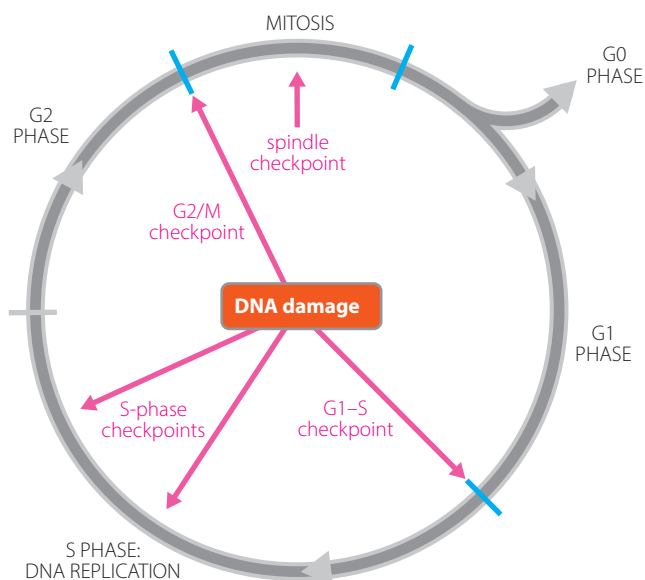
The search for deletions turned out to be problematic. Fully developed tumors have innumerable chromosomal changes including many deletions, but these are mostly passenger mutations rather than drivers of tumorigenesis. In contrast, the family linkage approach quickly led to identification of the causative gene in many familial cancer syndromes. Once the gene responsible has been identified, sporadic tumors can be checked for mutations in the same gene. In this way a significant number of tumor suppressor genes have been identified (*Table 7.3*).

Table 7.3 – Examples of familial cancer syndromes with inherited mutations in tumor suppressor genes

Syndrome	OMIM number	Gene	Location
Retinoblastoma	180200	RB1	13q14
Familial adenomatous polyposis coli	175100	APC	5q21
Lynch syndrome (hereditary non-polyposis colon cancer)	120435	MSH2	2p22
	120436	MLH1	3p21
Familial breast cancer	113705	BRCA1	17q21
	600185	BRCA2	13q12
Li–Fraumeni syndrome	151623	TP53	17p13
Gorlin syndrome	109400	PTC	9q22
Ataxia-telangiectasia	208900	ATM	11q23
Neurofibromatosis 1	162200	NF1	17q11
Neurofibromatosis 2	101000	NF2	22q12
Von Hippel–Lindau disease	193300	VHL	3p25
Multiple endocrine neoplasia 1	131100	MEN1	11q13
Multiple endocrine neoplasia 2	171400	RET	10q11
Familial melanoma	155601	CDKN2A	9p21

The normal functions of tumor suppressor genes

TS genes are particularly involved in ensuring the integrity of the genome. They form part of the systems that detect and repair DNA damage, that prevent DNA replication until all damage is repaired, that check and correct replication errors, that ensure chromosomes segregate correctly in mitosis, and that force recalcitrant cells to commit suicide (apoptosis). One particular protein, p53, the product of the *TP53* gene, is centrally

**Figure 7.7 – The cell cycle and its checkpoints.**

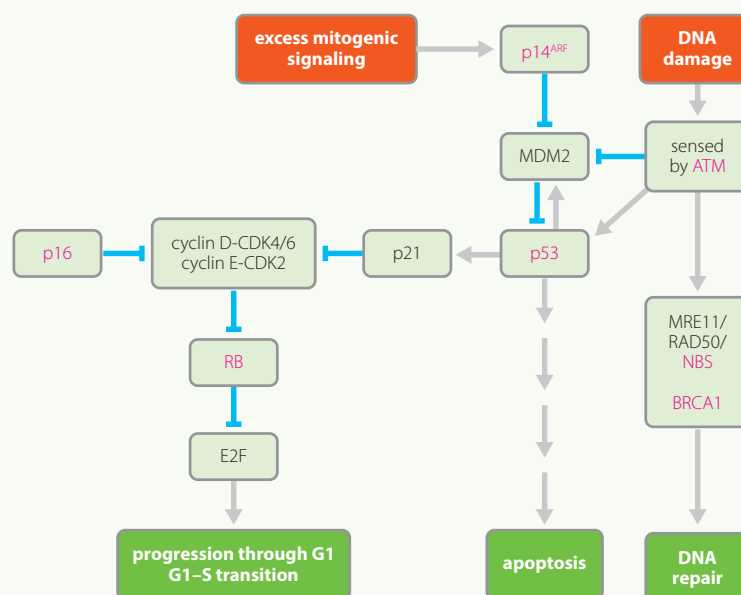
In a set of continuously growing cells the cell cycle can be divided into four phases. Checkpoints control progression through the cycle. Cells that are terminally differentiated and will not divide again enter G0 phase and remain there.

involved in so many of these processes, and is so frequently mutated in tumors, that it has acquired the label 'guardian of the genome' – but many other TS gene products also play important roles.

Detecting and repairing DNA damage such as double-strand breaks involves the ATM, BRCA1 and BRCA2 proteins (see *Table 7.3*), among many others. Replication of damaged DNA is prevented by a series of checkpoints in the cell cycle (*Figure 7.7*). Checkpoint mechanisms are strongly conserved through evolution, and much of our understanding of human cell cycle controls comes from studies of yeast. The 2001 Nobel Prize for Medicine was awarded to Leland Hartwell, Tim Hunt and Paul Nurse for their work unraveling these controls (you can read their accounts of how they did it at www.nobelprize.org/prizes/medicine/2001/hunt/lecture/ (and, similarly, [.../nurse/lecture/](http://www.nobelprize.org/prizes/medicine/2001/nurse/lecture/) and [.../hartwell/lecture/](http://www.nobelprize.org/prizes/medicine/2001/hartwell/lecture/)). The G1–S checkpoint in particular involves an impressive cast of TS genes (*Box 7.3*). A separate check in G2 (the decatenation checkpoint) prevents cells entering mitosis until the chromosomes are fully disentangled from one another, while, during mitosis, the spindle checkpoint prevents the separation of chromatids in anaphase until every chromosomal centromere is attached to spindle fibers.

The G1–S checkpoint

Progression through the cell cycle is primarily governed by the availability of specific cyclin proteins. These activate cyclin-dependent kinases (CDKs). The kinases act as effectors by phosphorylating a range of downstream targets; they are regulated by the availability of their cognate cyclins and by various inhibitors. Progression through G1 phase and into S phase is governed by cyclins D and E with their kinases. A complex network of upstream controls modulates their activity. *Box figure 7.3* shows only part of the network.



Box figure 7.3 – Some of the controls and interactions governing G1–S progress.

TS genes identified through familial cancers are shown in red. —| shows an inhibitory action, —> shows a stimulatory action.

Mismatch repair and microsatellite instability

DNA replication errors are minimized by various proof-reading mechanisms. One of these is the mismatch repair (MMR) system. This specifically checks for errors made by DNA polymerase when it replicates repetitive sequences like homopolymer runs (see *Figure 6.5*) or **microsatellites** – short tandem repeats of a 2-, 3- or 4-nucleotide unit such as (CA)_n or the repeats described in *Disease box 4*. The human genome contains maybe 150 000 such repeats. If the polymerase temporarily detaches from the template strand while replicating such a repeat, when it re-associates it may skip one or two repeat units, or replicate one or two twice, so that the new DNA strand has fewer or extra repeat units. Hence, although these repeats are mostly entirely harmless, they are hotspots for replication errors, and a special set of enzymes is dedicated to detecting and repairing this class of error. Loss of function of any of these enzymes leads to a proliferation of errors in certain types of cancer cells.

Six genes (*MSH2*, *MLH1*, *MSH6*, *MLH3*, *PMS1* and *PMS2*) are involved in the human MMR machinery. Homozygous loss of function, especially of *MSH2* or *MLH1*, produces **microsatellite instability**, in which microsatellites across the genome randomly acquire length variants. Instability of non-coding microsatellites is not pathogenic, but it alerts diagnostic laboratories to the risk that the instability may also affect coding sequences. For example, exon 3 of the *TGFR2* (TGFβ receptor 2) gene contains a run of 10 consecutive A nucleotides (*Figure 7.8*). MMR-deficient cells are liable to insert or skip one or more As. This creates a frameshift and renders the new copy non-functional. The TGFβ receptor 2 relays signals from transforming growth factor B, a strong inhibitor of cell proliferation in the colorectum. One survey found somatic *TGFR2* mutations in the tumors of 100/111 cases of colon cancer with microsatellite instability. Several other relevant targets have been noted – that is, genes with relevant functions whose coding sequences contain microsatellite-like or homopolymer sequences. Microsatellite instability is particularly seen in colon, endometrial and gastric tumors.

When all else fails, a cell is made to ‘commit suicide’ by apoptosis. Apoptosis is a specific active process that is triggered by various abnormal cellular states as well as being an important part of normal development (for separating the fingers of the hand in a developing embryo, for example). The mechanism involves activation of a cascade of proteolytic enzymes called caspases. Caspases can be activated through perturbations of mitochondria, with release of cytochrome c, or through the actions of FAS protein in conjunction with so-called death receptors. Such potent suicide machines naturally need to be kept under tight control, and the regulatory circuitry for apoptosis is extremely involved. Among the conditions that can trigger apoptosis are irreparable DNA damage and excessive growth signaling. The p53 protein is a central link in these processes. It is activated by phosphorylation of specific residues and/or inactivation of its inhibitor MDM2. As shown in *Box figure 7.3*, p53 stimulates expression of the p21 protein which acts

```

743  TGC  ATT  ATG  AAG  GAA  AAA  AAA  AAG  CCT  GGT  GAG  ACT  TTC
120  Cys  Ile  Met  Lys  Glu  Lys  Lys  Lys  Pro  Gly  Glu  Thr  Phe

```

Figure 7.8 – Part of the sequence of exon 3 of the *TGFR2* gene.

A run of 10 consecutive As makes replication of the gene vulnerable to defects in mismatch repair.

to promote cell cycle arrest. p53 also stimulates transcription of several proteins that are involved in both the mitochondrial and the FAS-mediated pathways of apoptosis.

It is noteworthy that several TS genes encode extremely large proteins. Examples include the proteins encoded by *APC* (2843 amino acid residues), *ATM* (3056 residues), *BRCA1* (1863 residues), *BRCA2* (3418 residues) and *NF1* (2839 residues). There are counter-examples: the two products of the *CDKN2A* gene are a mere 156 and 173 residues long, while some of the largest known proteins are muscle structural proteins like dystrophin, 3685 residues, and titin, 19 946 residues, that have no role in cancer. Nevertheless, the list in *Table 7.3* suggests that loss of function of genes encoding very large non-structural proteins is a frequent initiating event in tumorigenesis. Such proteins do in fact often have a structural role at the molecular, though not the microscopic, level. They interact with many other proteins and serve as scaffolds for assembling the large multi-protein machines that carry out many cellular tasks. Loss of function at such nodal points in the networks of interactions within cells may be a powerful way of disrupting normal cellular functions.

The multistage development of cancer

Pathologists have long known that malignant tumors develop through stages marked by increasing growth and de-differentiation. Familial adenomatous polyposis (FAP, the condition affecting **Christos Xenakis, Case 17**) offers exceptional opportunities for studying the process. When the colon is surgically removed from a FAP patient, it often contains lesions showing every stage of tumor development (*Figure 7.3b*). Many years ago, Vogelstein and colleagues analyzed such series, together with sporadic colon tumors, for mutations or deletions in candidate genes. They found some alterations that were often present in the early stages, and others that only appeared in later stages. Putting their observations together, they proposed the developmental scheme shown in *Figure 7.9*.

Later genome-wide analyses (*Section 7.4*) have greatly extended and complicated this picture, but it still demonstrates two useful points.

- The earliest stages are the most critical, because these mutations must cause instability or a growth advantage in a relatively normal cell that still has most of its defenses intact. Probably there are only a small number of ways of achieving this. Once a pre-tumor cell has acquired genomic instability the range of

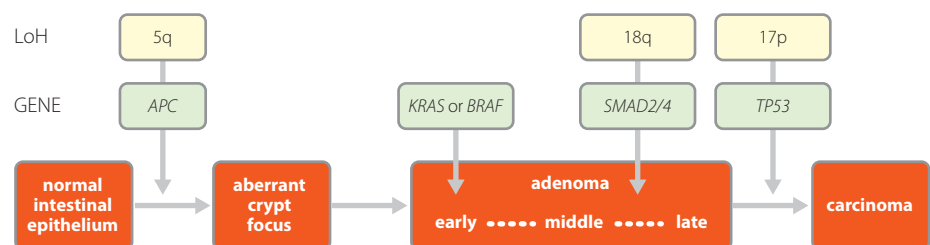


Figure 7.9 – A possible common pathway for the development of colon cancer.

The scheme (see Kinzler and Vogelstein, 1996) summarizes the way that some genes are often mutated in early lesions, while others are mutated only in late-stage lesions. There is no suggestion that every colon cancer has developed in exactly this way. LoH, loss of heterozygosity: chromosomal regions with no heterozygous variants compared to the same region in non-tumor cells of the patient, suggesting one copy of that region has been lost in the tumor cells.

possibilities expands and fewer generalizations are possible. Cells with mismatch repair defects may have a wider range of possible evolutionary paths towards cancer, but many of the same players are involved.

- Tumorigenesis concerns pathways, not individual genes. Tumors lacking *APC* mutations often have mutations in other components of the same pathway, for example, β -catenin or axin. Tumors lacking *RAS* mutations commonly have mutations in the *BRAF* gene, which has similar functions (see *Disease box 3*). As mentioned above, TGF β signaling is an important negative regulator of cell growth in the colon. Within the cell, the TGF β receptor signals by phosphorylating serine or threonine residues in SMAD proteins. This pathway is commonly subverted in colorectal cancer: in cancers with MMR defects by *TGFR2* mutations, or in MMR-competent cancers by loss of chromosome 18 where the *SMAD2* and *SMAD4* genes are located.

Some combination of these lesions is necessary for cancer to develop, but it is apparently not sufficient. Studies of normal tissue from healthy older adults have documented the presence of cells and clones of cells with many of the same mutations as are found in cancer cells – yet they do not develop into tumors (Martincorena *et al.*, 2015, 2018). It seems that some specific accumulation of mutations or some sort of favorable cellular or biochemical context is needed for a mutant cell to establish a tumor.

7.3. Investigations of patients

CASE 15 TIERNEY FAMILY

- 4-year-old boy, Jason
- Pale with extensive bruising and tachycardia
- ? Acute lymphocytic leukemia
- Diagnosis of ALL confirmed with *TEL-AML1* fusion gene

175

190

261

395

Childhood acute lymphoblastic leukemia (cALL) is a neoplasm of B cell precursors or stem cells. About 25% of cases have a balanced reciprocal 12;21 translocation, with breakpoints at 12p13 and 21q22.3 (*Table 7.2*). Testing of blood and bone marrow by fluorescence *in situ* hybridization (FISH) showed that Jason's abnormal cells had this translocation. At the translocation junction, the 5' portion of the *TEL* (*ETV-6*) gene and almost the entire coding sequence of the *AML1* (*RUNX1* or *CBFA2*) gene are brought together to create a fusion gene. Loss of function experiments in mice show that both *TEL* and *AML1* are critical genes for hematopoiesis. Each gene encodes a transcription factor. *AML1* encodes the alpha subunit of core binding factor, a master regulator of the formation of hematopoietic stem cells. The fusion inhibits normal *AML1*-mediated transcriptional activity, resulting in alteration of the capacity of hematopoietic stem cells to self-renew and to differentiate. The *TEL-AML1* fusion appears to be unique to B cell progenitor ALL, but both genes are found as fusion partners with a variety of other genes encoding kinases or transcription factors in both lymphoid and myeloid leukemias.

The positions of the *TEL/AML1* breakpoints in the pale-staining subtelomeric chromosome regions make the translocation hard to spot by conventional karyotyping. Diagnosis is based on fluorescence *in situ* hybridization (FISH). Interphase cells from Jason were hybridized to differently colored FISH probes for the *TEL* and *AML1* genes. One of the colored spots marking the *TEL* genes always lay adjacent to one of the *AML1* spots, confirming the presence of the *TEL-AML1* fusion gene (*Figure 7.10*). This was good news, because individuals with *TEL* rearrangements usually show an excellent response to chemotherapy. Jason's story is continued in *Chapter 10*.

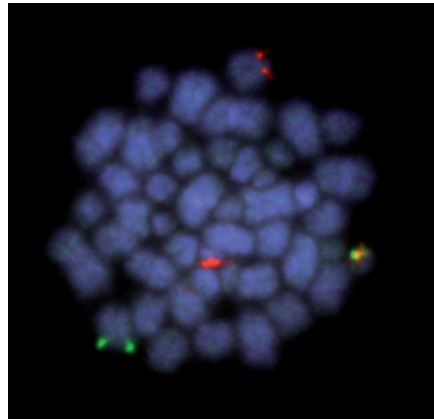


Figure 7.10 – Metaphase with *TEL-AML1* fusion.

The green signal is on the normal chromosome 12, one red signal is in the normal chromosome 21 and one is on the derived chromosome 12. The yellow *TEL-AML1* fusion signal is on the derived chromosome 21. Photo. courtesy of Dr Christine Harrison, University of Newcastle upon Tyne.

CASE 16 WILSON FAMILY

176 191 395

- Family history of breast cancer
- Options for genetic testing
- Family *BRCA2* mutation identified
- Implications for relatives
- Possibilities for therapy

Since one woman in eight in the USA or UK is likely to develop breast cancer at some time in her life, it is not surprising that many families have more than one case. Up to 20% of affected women have an affected first- or second-degree relative. Many of these represent chance coincidences, but statistical analysis suggests that in 5–10% of women with breast cancer the condition is truly familial. In 1990, linkage analysis in a large collection of multicase families pinpointed a possible susceptibility locus for early-onset breast cancer at chromosome 17q21. After four more years of intensive work the *BRCA1* gene was identified at that location. A further round of analysis in *BRCA1*-negative families led to identification of the *BRCA2* gene on chromosome 13.

The lifetime risk for a *BRCA1/2* mutation carrier has been variously estimated between 60% and 85%. The high initial estimates were made in the large multi-case families in which the mutations were first found. But these families were selected for having many affected cases, and so the initial estimates of risk are from a biased set of families. Population-based surveys show a lower risk. In fact historical studies in Iceland, based on genealogical links to present-day mutation carriers, suggest that in the early 20th century the risk may have been as low as 30%, showing the strong influence of lifestyle and environmental factors. Nevertheless, the present-day risk is still substantial compared to the risk for people who do not carry the mutations, and possibly particularly so in women with a strong family history. Thus there is considerable demand for mutation screening.

Checking for *BRCA1/2* mutations is a lot of work. Cancer susceptibility is conferred by loss of function, so mutations might be anywhere in either gene, and both genes have very large coding sequences. *BRCA1* has 24 exons encoding a 1863 amino acid protein, and *BRCA2* has 28 exons encoding 3418 amino acids (Figure 7.11). For Ashkenazi Jews, three founder mutations are very frequent (see Disease box 9), allowing easy screening (although a negative screen would not remove the risk that some other *BRCA1/2* variant, or a variant in another gene, may be putting the person at high risk). Other populations, for example French-Canadians, Icelanders and Pakistanis, also have their own particular founder mutations. Wendy Wilson's family do not come from such a population, so before offering gene sequencing it was necessary to decide whether the chance of a *BRCA1/2* mutation was high enough to justify the work.

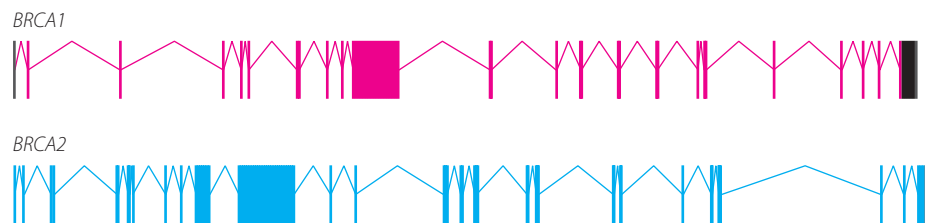


Figure 7.11 – Exon structure of *BRCA1* and *BRCA2* genes.

Both genes have one very large central exon (3426 and 4932 bp, respectively). The two genes and proteins share no homology, and are involved in different aspects of detecting and repairing DNA damage in cells.

Markers of *BRCA1/2* mutations include:

- cases with unusually early onset
- bilateral cases (compare with the RB example, *Figure 7.6*)
- cases with male breast cancer (particularly a feature with *BRCA2* variants)
- families with both breast and ovarian cancer (particularly a feature with *BRCA1* variants).

None of these features is completely specific to *BRCA1/2* breast cancer, but scoring systems have been developed based on these factors. *Box 7.4* shows one such system (Evans *et al.*, 2017). The scores allow physicians to assess the probability that any given case is likely to have a *BRCA1/2* variant.

A scoring system for assessing the likelihood of a *BRCA1/2* mutation

A score is calculated for the patient and all blood relatives who are on the side of the family with the suspect family history; stop counting when the history passes through two unaffected females who are aged >60. For the patient, but not affected relatives, adjust for the breast cancer pathology as shown; for each affected woman with ovarian cancer, add 2 to the score if the cancer was high-grade serous at age <60. Add together the scores for the patient and affected relatives. See Evans *et al.* (2017).

Scores for female breast cancer:

Age at diagnosis (years)	<30	30–39	40–49	50–59	>59
Score	11	8	6	4	2

Epithelial ovarian cancer (any grade): Age <60 13 points; age >59 10 points

Other cancers:

Male breast cancer: Age <60 13 points; age >59 10 points
 Pancreatic cancer 1 point
 Prostate cancer: Age <60 2 points; age >59 1 point

Adjust for breast cancer pathology (index case only):

Grade 1 / 2 / 3 –2 / 0 / 2 points
 ER +ve / –ve –1 / 1 point
 Triple negative 4 points
 HER2 +ve –6 points

In a UK-based population, a score of 15 or higher equates to a *BRCA1/2* mutation probability of >10%.

Many multicasé breast cancer families do not have mutations in either of the *BRCA1/2* genes. Sustained searches have failed to reveal any ‘*BRCA3*’ gene that could account for any significant proportion of the remaining cases. It seems likely that the remaining familial tendency is made up of the combined effects of several low-penetrance loci. Several such loci have been identified, either by targeted analysis of known interactors with the *BRCA1* or *BRCA2* proteins, or by large-scale association studies of the type described in Section 13.4 (Table 7.4).

Table 7.4 – Some inherited genetic variants associated with risk of breast cancer

Gene	Variant	Chromosomal location	Odds ratio
<i>ATM</i>	Loss of function	11q22.3	2.37
<i>BRIP1</i>	Loss of function	17q22	2.00
<i>PALB2</i>	Loss of function	16p12	2.30
<i>CHEK2</i>	c.1100delC	22q12.1	2.34
<i>FGFR2</i>	SNP rs2981582	10q26	1.26
<i>MAP3K1</i>	SNP rs889312	5q11.2	1.13
<i>TNRC9</i>	SNP rs51005538	16q12	1.11
<i>LSP1</i>	SNP rs1865582	11p15.5	1.07

Odds ratios are the odds of finding the specified variant in a woman with breast cancer compared to a healthy control (see Box 12.1). For *ATM*, *BRIP1* and *PALB2* the odds ratio refers to the overall frequency of loss of function mutations in cases compared to controls. For *FGFR2*, *MAP3K1*, *TNRC9* and *LSP1* the variants are single nucleotide polymorphisms, which probably do not themselves increase risk, but mark chromosomal segments that contain a risk factor, most probably a *cis*-acting control element of the gene listed in the first column. See Section 13.4 for further discussion of such data.

All these variants together explain only a small extra part of the familial tendency of breast cancer. The functional gene variants are rare, while the SNPs are common but confer very modest extra risks. The question whether genotyping women for any or all of these variants would have any clinical utility has been debated. Some researchers recommend calculating a **polygenic risk score** (an idea discussed further in Chapter 13). Briefly, a computer is used to analyze many thousands of random variants from large panels of women with breast cancer and healthy controls. Artificial intelligence is then used to devise the estimator that best predicted risk. The idea is controversial because the great majority of variants analyzed have no known association with breast cancer. Nevertheless, enthusiasts point out that in independent datasets the 10% of women with the highest risk scores have several times the risk of the 10% with the lowest scores. Arguably the main use of a polygenic risk score would be to fine-tune the age at which women are offered routine mammographic screening.

In the Wilson family, applying the scoring system of Box 7.4 to the pedigree (Figure 7.2a) gave a score of 33 (before adjusting for pathology). This suggested there was a high likelihood of a *BRCA1/2* mutation – most likely *BRCA2* given the case of male breast cancer.

Note that no score is given for the prostate cancer in Wendy's paternal grandfather, because if the family problem is caused by an inherited *BRCA1/2* variant, the pedigree shows that the problem is on Wendy's mother's side of the family. The score was amply high enough to prioritize the family for DNA analysis.

A section of Wanda's excised tumor was recovered from the pathology lab archive and DNA was extracted. Testing eventually revealed a single nucleotide deletion in exon 18 of *BRCA2* in the tumor. The deletion of nucleotide 8525 in codon 2766 created a frameshift, leading to codon 2776 being read as a stop codon. No second variant was identified in the tumor, although no check was made for loss of heterozygosity because there was no sample of Wanda's normal (blood) DNA for comparison. The presence of this variant in the tumor strongly suggested that this was indeed a *BRCA2* family, because sporadic tumors rarely have *BRCA2* mutations.

Before family members could be offered DNA testing it was necessary to discover whether the variant found in the tumor was a first or second hit – if it was the second hit, then the inherited variant remained unidentified. Amy in New Zealand was contacted and, in discussion with Wendy, agreed to provide a blood sample for testing. DNA was extracted by arrangement with Amy's local genetic service. Unlike blood, DNA is stable and can readily be sent around the world by normal mail. It was easy for Wendy's genetic center to test Amy's sample specifically for the c.8525delC variant, and the result was positive. Now that the family variant had been identified, Wendy, William and Veronica gave DNA samples for testing. The samples were tested just for the identified family variant. This showed that Veronica and William carried the variant but Wendy did not.

Wendy was naturally relieved at the news, but in counseling her it was important to make sure she understood that she was still at the population risk of developing sporadic breast cancer, and would be wise to take part in the standard mammography screening program when she reached the eligible age. This was also an opportunity to point out to her that breast cancer risk is strongly influenced by lifestyle, and that women (including carriers of *BRCA1/2* variants) can reduce their risk by around 40% by weight loss and moderate exercise. Veronica carried the variant and was therefore at high risk. Options for her included doing nothing, lifestyle changes, entering an enhanced surveillance program with annual mammography, taking Tamoxifen in the hope that this would reduce her risk, or the more radical step of prophylactic mastectomy. The result also implied that Wendy's great-aunt (the grandmother of the man with breast cancer) almost certainly carried the variant, and so was at high risk – though whether, at the age of 72, she would wish to do anything about that was something she needed to discuss with her GP. William was not at high risk himself of breast cancer because it is rare in males (his relative risk was very high, but the absolute risk was still low). On the other hand, although the relative risk of prostate cancer in *BRCA2* variant carriers is only moderate, the absolute risk is high, and he was advised to have regular screening. Any daughter he had would also need counseling about her substantial risk of breast cancer.

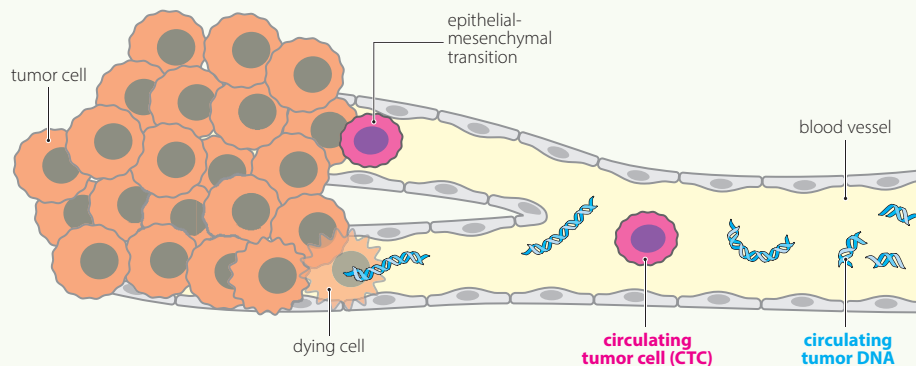
Wendy's aunt Amy, who had survived breast cancer, now knew she carried the family mutation, and she discussed with her oncologist in New Zealand whether her current monitoring program was still her best option. In particular they discussed the potential of liquid biopsies (Box 7.5) for monitoring the risk of recurrence.

Liquid biopsies

The peripheral blood of cancer patients contains both circulating tumor cells (CTC) and cell-free tumor DNA (ctDNA). Isolating and characterizing either of these would offer an attractive alternative to needle biopsies for monitoring the development or recurrence of a tumor. A particular advantage is that the cells or DNA could come from both the primary tumor and any metastases, thus potentially providing information not available from a needle biopsy of the primary tumor.

Isolating tumor-specific cells or DNA in peripheral blood requires great technical sophistication. CTC in particular may be present as only one CTC among 10^7 white blood cells in 1 ml of blood. Isolating ctDNA is somewhat less demanding, but the proportion of total cell-free DNA in a patient that is tumor-derived can be anything from 0.01% to 93%. Because of the great promise of liquid biopsies there is very intensive technical development and numerous trials. The reviews by Finotti *et al.* (2018) and Alimirzaie *et al.* (2019) give many details and references.

In Amy's case (**Case 16, Wilson family**) liquid biopsies would provide an excellent non-invasive way to check for minimal residual disease and to detect any further evolution or metastasis of the original tumor.



Box figure 7.4 – Tumors shed both intact cells and cell-free DNA into the circulation.

Adapted from Wyatt AW & Gleave ME (2015) *EMBO Mol. Med.* **7**: 878–894; and reproduced here under a Creative Commons Attribution CC BY 4.0 Licence.

CASE 17 XENAKIS FAMILY

- Family history of bowel problems
- ? Familial adenomatous polyposis
- APC mutation identified
- Risk to relatives
- How to manage his children?
- Possibilities for therapy

176

195

395

The *APC* gene (OMIM 175100) on chromosome 5q21–22 was identified as the cause of FAP through family linkage studies, guided by reports of an affected patient with a interstitial deletion of 5q. *APC* acts as a classical TS gene. Familial cases inherit one *APC* mutation and somatically inactivate the normal allele in their tumor (but see Albuquerque *et al.* (2002) for some complications). In addition, in contrast to the situation with *BRCA1/2* in breast cancer, mutation or loss of both alleles of *APC* is a common event early in the development of common sporadic colorectal cancers (see Figure 7.9).

The APC protein is another large (2843 amino acids) multifunctional protein. It appears to be involved in several different processes in the cell, including cell adhesion and interactions with the cytoskeleton. The function most clearly related to cancer is its role in the control of β -catenin levels. In the nucleus, β -catenin acts as a transcriptional

co-activator, promoting transcription of target genes including cyclin D1 and the *MYC* oncogene. APC protein forms part of a complex in the cytoplasm that degrades β -catenin. The APC protein has three β -catenin-binding modules and seven 20-amino acid modules that down-regulate β -catenin levels (Figure 7.12). Cell signaling through the Wnt pathway inhibits formation of the complex, freeing the β -catenin to enter the nucleus. APC mutations allow the level of β -catenin to rise independently of the Wnt signal although, interestingly, as explained in the caption to Figure 7.12, it seems that most mutations do not completely abolish all control (Albuquerque *et al.*, 2002). Perhaps complete loss would produce β -catenin levels that were sufficiently abnormal to trigger apoptosis. Colorectal tumors that lack APC mutations may have gain of function mutations in β -catenin or loss of function mutations in axin (another component of the APC complex) that presumably have a similar end result.

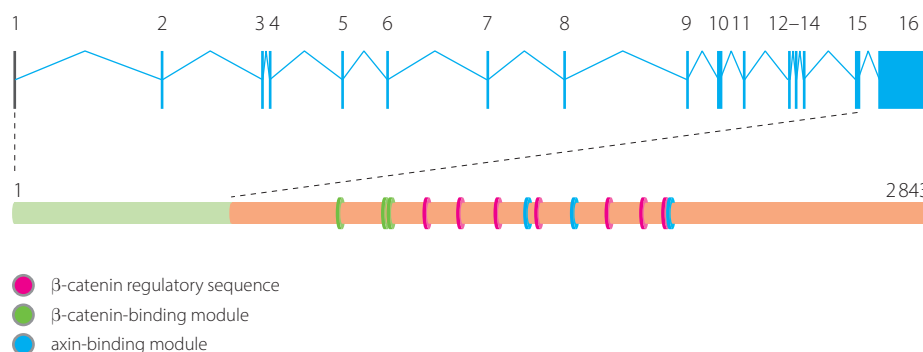


Figure 7.12 – The APC protein and its role in regulating β -catenin levels.

Three 15-amino acid modules bind β -catenin and seven 20-amino acid modules down-regulate the level. All these modules are encoded by the large 3' exon 16. The first 15 exons of the gene encode only 653 of the 2843 amino acids. Thus APC genes with a premature stop signal in any codon downstream of codon 640 can produce a truncated protein, because nonsense-mediated RNA decay does not occur for stop codons in the last exon of a gene or about 50 nt upstream of the last splice site (see Figure 6.4). Only truncations in the green shaded part of the protein would be eliminated by nonsense-mediated decay. Thus many mutations produce truncated APC proteins that can still bind β -catenin and have some limited ability to down-regulate it.

Christos Xenakis gave a blood sample, from which DNA was extracted for APC mutation testing. This showed that he was heterozygous for a nonsense change at codon 1309 (p.Glu1309ter), a frequent germ-line variant in FAP patients. It is highly penetrant, so that carriers have a near-100% risk of developing colon cancer if untreated. Identifying it allowed at-risk family members to be tested. Christos's mother Demi contacted several relatives of her late husband in Cyprus to let them know about the newly discovered high risk in the family, and gave them a contact number of the Seattle geneticist for use by their local genetics service if any of them opted for testing. Meanwhile there were Christos's two young children to consider. Each was at 50% risk, and they would need annual sigmoidoscopy examinations from age 10 to check for development of polyps. All this could be avoided if a DNA test showed that the child had not inherited the variant. The question arose whether to test them now or leave it until later. Testing now would mean that an at-risk child could grow up with the knowledge that he would need annual bowel screening and not have this suddenly sprung on him; postponing testing would

mean that the child, by age 10, would be more able to understand and consent to the procedure.

This family illustrates some of the ethical arguments about DNA testing in children. Normally it is felt improper to test children for genetic susceptibility. In the present case there are benefits for a child in being tested by age 10. For children at 50% risk of FAP, annual sigmoidoscopy surveillance needs to start around this time. It is clearly beneficial to a child to avoid this unpleasant procedure if he can be shown not to carry the family mutation. Only parental consent would be formally required for testing, but good practice would include discussing the issue with the child and obtaining consent as far as possible.

7.4. Going deeper...

Getting the complete picture: whole genome studies

Cancer research has moved decisively from studies of single oncogenes or TS genes to identifying the totality of somatic changes in tumor genomes. Many laboratories are using next-generation sequencing to define whole genome sequences of tumors. Whole genome sequencing, rather than exome sequencing, is appropriate in studies of tumors because it will reveal structural variants, which play important roles in cancer. If the normal genome of the host is also sequenced, a complete list of somatic mutations can be produced. In contrast to earlier work, the new studies give an unbiased picture of changes across the whole genome. The COSMIC database (<http://cancer.sanger.ac.uk/cancergenome/projects/census>) aims to catalog all genes mutated in cancer.

Complementing this are genome-wide studies of changes in gene expression. Initially these used expression arrays, microarrays carrying probes specific for a large number of cDNAs. These had the disadvantage of only reporting changes in the genes represented by the probes on the array. Newer studies use mass sequencing of cDNAs, which again gives an unbiased view. Large collaborative projects such as the Cancer Genome Anatomy Project (https://en.wikipedia.org/wiki/Cancer_Genome_Anatomy_Project) sought to determine the gene expression profiles of normal, pre-cancer, and cancer cells of many different types. Genome-wide surveys of epigenetic modifications, microRNAs and proteomics – so-called multi-omics studies (Box 7.6) complete the picture.

Multi-omics approaches to cancer

The biggest change of the last 30 years, across the whole of biomedical research, has been the movement from focused, hypothesis-driven approaches to hypothesis-free data-driven approaches. The move from genetics to genomics (from 'fishing with a line' to 'fishing with a net' in the words of the geneticist Andrea Ballabio), has been just one of a series of parallel developments. We now have transcriptomics, proteomics, epigenomics and a whole range of 'omics' technologies, all amassing huge quantities of data in a hypothesis-free way for subsequent analysis.

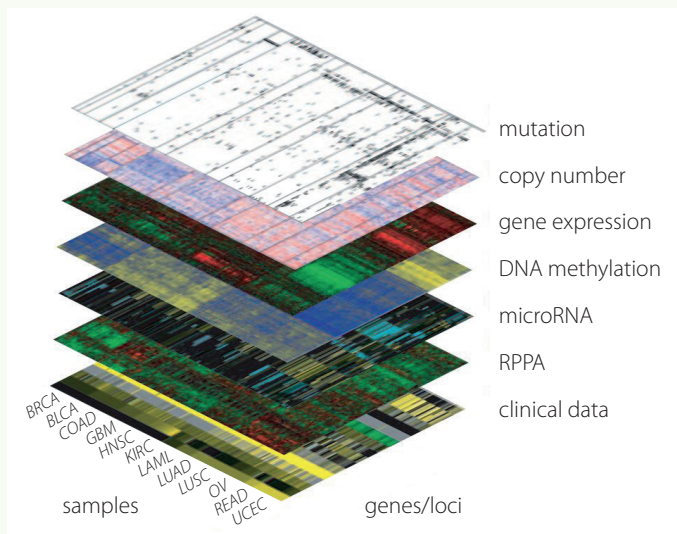
In cancer research, Box figure 7.5 shows the multi-omics approach used by one big collaborative study, the Cancer Genome Atlas (TCGA; see www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Starting in 2006 the project molecularly characterized over 20 000

BOX 7.6

primary cancers and matched normal samples spanning 33 cancer types. As well as the obvious layers of investigation, other layers include:

- RPPA (reverse-phase protein arrays), a proteomic approach.
- microRNAs – these 20–25-nt small non-coding RNAs must have been seen by anybody who ran an RNA gel, but they were dismissed as degradation products until the 1990s, when Andrew Fire and Craig Mello realized their significance (for which they won the 2006 Nobel prize in Physiology or Medicine – you can read their accounts of how they came to their discoveries at www.nobelprize.org/prizes/medicine/2006/mello/lecture/ (and .../fire/lecture)). MicroRNAs fine-tune gene expression by hybridizing to mRNAs, usually to sequences in the 3' untranslated region. This usually represses translation. One microRNA may interact with many different mRNAs, and several microRNAs may regulate a single mRNA. Humans have around 1900 microRNAs (cataloged at www.miRBase.org). They form a pervasive network of gene regulation, which is typically disturbed in tumor cells.
- DNA methylation is a major regulator of transcription. Methylation is an epigenetic mechanism – that is, a mechanism that alters gene expression without altering the DNA sequence. See *Chapter 11* for details.

Details of this and other large collaborative projects, most recently the Pan Cancer Analysis of Whole Genomes Project (Campbell *et al.*, 2020), can be accessed through the US National Cancer Institute's Office of Cancer Genetics (<https://ocg.cancer.gov/>) and the International Cancer Genome Consortium (<https://icgc.org/>).



Box figure 7.5 – Multi-omics analysis of tumors.

The 2012 data freeze of The Cancer Genome Atlas showed data on 5074 tumor-normal pairs for 12 cancer types (*BRCA*, breast; *BLCA*, bladder; *COAD*, colon; *GBM*, glioblastoma; *HNSC*, head and neck; *KIRC*, kidney; *LAML*, leukemia; *LUAD*, lung adenocarcinoma; *LUSC*, lung squamous; *OV*, ovary; *READ*, rectum; *UCEC*, endometrial). Reproduced from The Cancer Gene Atlas Research Network (2013; *Nature Genetics*, **45**: 1113); with permission from Nature Publishing Group.

Analysis of individual tumors gives a daunting impression of complexity. The mass of data obtained can be represented by a so-called Circos plot, as in *Figure 7.13*. Collating changes in many tumors of the same type allows a first attempt to distinguish drivers from passengers.

A first step in reducing the complexity is to think in terms of genes rather than individual variants: does a variant cause a loss, gain or change of function of a gene? To further reduce the complexity, think in terms of pathways rather than individual genes. For example, *Figure 7.14* shows how many different gene mutations in the brain tumor glioblastoma multiforme all act by dysregulating one signaling pathway.

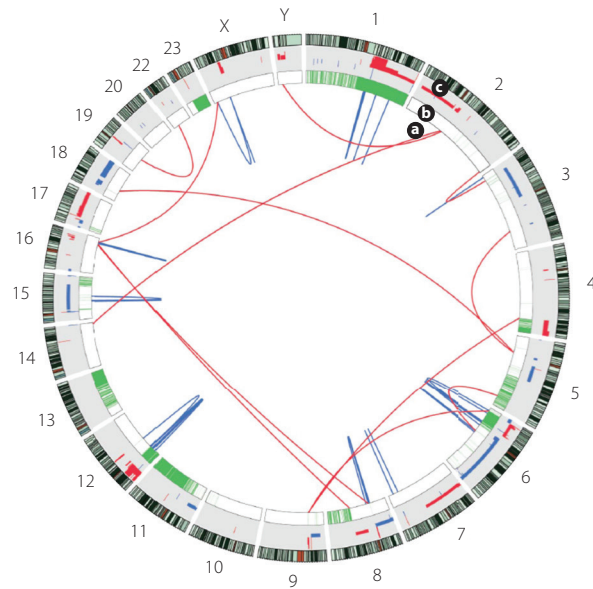


Figure 7.13 – A Circos plot showing somatic changes in a primary non-small cell lung tumor from a cigarette smoker.

The outer circle shows the chromosomes with their banding patterns. The innermost lines (labeled a) show structural rearrangements, red for interchromosomal and blue for intrachromosomal changes. Circle b shows regions of loss of heterozygosity and allelic imbalance, marked in green. Circle c shows copy number changes (red = gains, blue = losses). In addition to the changes shown here, 50 675 point mutations were identified. Data of Lee *et al.* (2010; *Nature* **465**: 473–477), reproduced with permission from Nature Publishing Group.

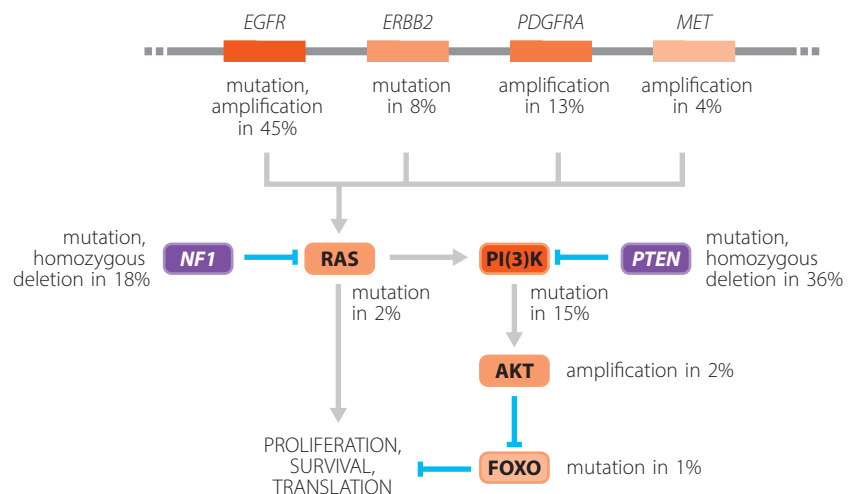


Figure 7.14 – Altered signaling in glioblastoma multiforme cells.

In different tumors different mutations all have the same effect of deregulating the responses of cells to signals at the cell surface receptors EGFR, ERBB2, PDGFRA and MET. Reproduced from The Cancer Gene Atlas Research Network (2008; *Nature*, **455**: 1061–1068), with permission from Nature Publishing Group.

Recently great progress has been made by single-cell analyses. All the analyses described above used pooled DNA from large numbers of cells; however, tumors are not homogeneous clones of identical cells like colonies of bacteria on an agar plate. Even apart from obvious features like blood vessels, the genomic instability means that tumors consist of multiple clones related by complex branching mutational histories. Techniques for single-cell genomics have advanced considerably in the past few years, so that both DNA and RNA (cDNA) analysis of single cells have become routine, at least in big-budget research laboratories. Single cell analyses allow mutational histories to be unraveled. For treating cancer, the key target must be those cells that have the potential to metastasize (to seed secondary tumors elsewhere in the patient's body). These are only a very small subset of all the cells in most tumors, and their specific properties and vulnerabilities may be hidden when pooled cells are studied.

In fact, it is now possible to step back and take a fully global overview of cancer.

- First, consider what it is that makes a cancer cell different. A highly recommended review by Hanahan and Weinberg (2000, updated in 2011) suggests any cancer cell needs to acquire six specific capabilities:
 - ability to divide independently of external growth signals
 - ability to ignore external anti-growth signals
 - ability to avoid apoptosis
 - ability to divide indefinitely without senescence
 - ability to stimulate sustained angiogenesis
 - ability to invade tissue and establish distant secondary tumors.In their 2011 update the authors suggested two further 'emerging' hallmarks might be re-programming of energy metabolism to support continuing cell proliferation and the ability to evade immune surveillance.
- Secondly, consider how these capabilities might be acquired. In 2013, Vogelstein and colleagues suggested that all the many oncogenes and TS genes acted in just 12 signaling pathways (*Figure 7.15*).
- Finally, consider the processes by which the many necessary DNA sequence changes are acquired. Apart from chromosomal rearrangements, a global analysis of small-scale sequence changes reveals a limited number of mutational signatures that help identify the processes at work (see Alexandrov *et al.*, 2013).

Genomics-based classifications of tumors

The objective of classifying tumors is to ascertain the prognosis and identify the optimal treatment. Traditional schemes use the tissue of origin and histological appearance. Antibody staining may allow a finer classification – for example, breast tumors can be divided into molecular subtypes that correlate with the presence or absence of estrogen or progesterone receptors and Her2 (ERBB2) amplification. Analysis of 500 tumors in the TCGA project confirmed that there are four main breast cancer classes, luminal A, luminal B, Her2-enriched and basal-like or triple-negative. Within each class there was considerable heterogeneity, but overall the data robustly showed that breast cancer should be split into four different diseases. Previous studies had shown substantial differences in prognosis, and several commercial profiling kits use these and similar findings to guide treatment.

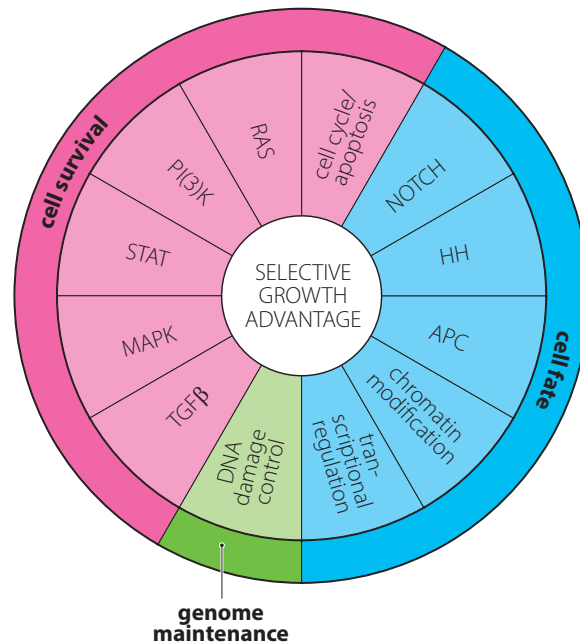


Figure 7.15 – Twelve cell signaling pathways that are the targets for oncogenic changes in oncogenes and tumor suppressor genes.

Reproduced from Vogelstein *et al.* (2013; *Science* **339**: 1546–1558), with permission from American Association for the Advancement of Science.

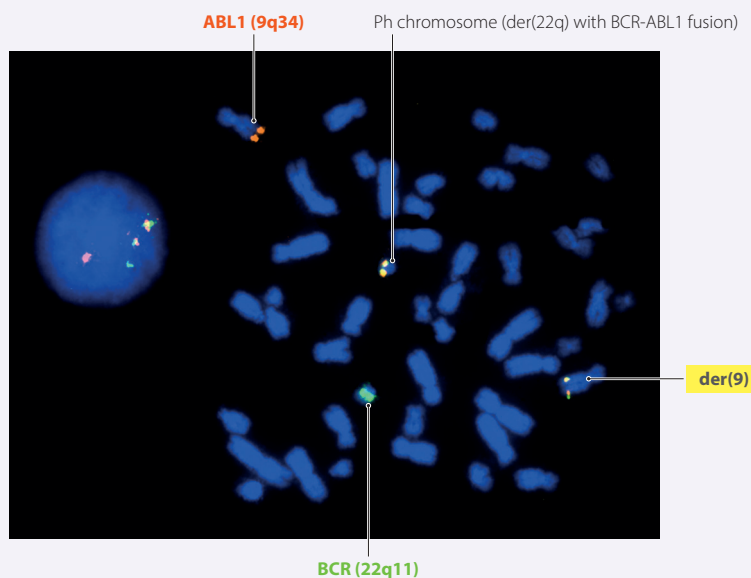
Other examples of molecular classification include colorectal tumors, which have long been known to divide into those with microsatellite instability (*MSH2/MLH1*–driven) and those without (*APC*–driven). Similarly, ovarian tumors split into two groups with different expression profiles. Tumors with *BRCA1* mutations fall into one group, those with *BRCA2* mutations in the other. Sporadic tumors lacking *BRCA1/2* mutations fall into one or the other group. Thus cancer of the ovary comprises two different diseases that may respond to different optimal therapies.

As well as splitting cancers of a given tissue into a number of different conditions, molecular studies also allow cancers of different tissues to be lumped together. Ciriello *et al.* (2013) used statistical clustering schemes on the TCGA data from 3299 tumors of the 12 types in *Box 7.5*. Their results showed that it is often more useful to classify tumors by molecular signatures than by the tissue of origin. The tumors could be divided into two broad categories, those mainly driven by somatic mutations and those primarily driven by copy-number changes. Below these main groups, about 30 signatures could be identified that mostly ran across tissue boundaries. Tumors with a given signature tended to share some common mutational events, although there was substantial heterogeneity in the detail. Some of the shared events were known to be responsive to specific drugs, thus patients might hopefully share responses to targeted therapy. All these advances in understanding have underpinned new developments in therapy, which are briefly described in *Disease box 7* below, and covered in more detail in *Chapter 10*.

Chronic myeloid leukemia

Chronic myeloid leukemia (CML) typically presents with rather non-specific symptoms of fatigue, exercise intolerance and maybe anorexia or weight loss in an elderly person (mean age at diagnosis 65). Clinical examination usually reveals an enlarged spleen and greatly elevated white cell count in the blood. Untreated, the symptoms worsen progressively over a year or two and median survival is 3–5 years after diagnosis. Over the years CML has been a model, both for understanding the etiology of cancers and for developing effective treatment.

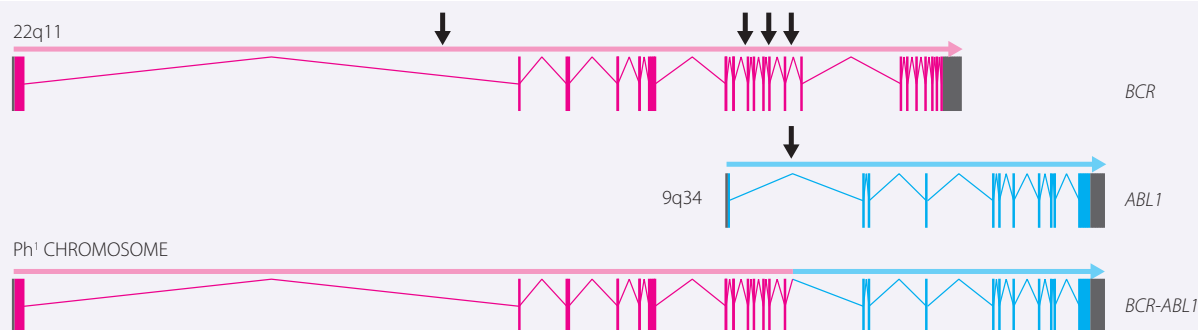
Lymphocytes from 90% of patients with CML contain an abnormal small chromosome, the Philadelphia (Ph¹) chromosome. It is a sufficiently constant feature of the disease to make the diagnosis of CML questionable in its absence. Back in 1960 the Ph¹ chromosome was shown to be one product of a balanced reciprocal 9;22 translocation, t(9;22)(q34.1;q11.2) (*Box figure 7.6*). Many other specific chromosomal rearrangements were subsequently described in other leukemias and cancers (see *Table 7.2*). It was later shown that the translocation junction creates a novel chimeric gene on the Ph¹ chromosome by joining the 5' part of the *BCR* gene from chromosome 22 on to the 3' part of the *ABL1* gene from chromosome 9. The resulting *BCR-ABL1* gene always contains exon 1 of *BCR* and usually the next 10 or so exons, joined to exons 2–11 of *ABL1* (*Box figure 7.6*). The chimeric gene is transcribed and translated, producing a functional protein.



Box figure 7.6 – Detecting the *BCR-ABL1* fusion gene by FISH.

Left: interphase FISH shows one red (*ABL1*) and one green (*BCR*) signal plus two mixed signals. Right: metaphase FISH identifies the chromosomes involved. Reproduced from the Atlas of Haematological Cytology (www.leukemia-cell.org/atlas) with permission.

ABL1 is a known oncogene (see *Table 7.1*). Its product is a tyrosine kinase, a common class of signaling molecule that exerts its effect by phosphorylating tyrosine residues in its target proteins, in this case controlling cell growth (see *Figure 7.4*). Activity of the kinase is tightly controlled, in part by an N-terminal domain that is lost by the translocation. The chimeric gene is transcribed and translated to produce a tyrosine kinase but, unlike the *ABL1* kinase, the *BCR-ABL1* kinase is constitutively active, transmitting its growth-promoting signal without the normal regulation. Such a gain of function in an important growth control could clearly tip cells towards uncontrolled proliferation.



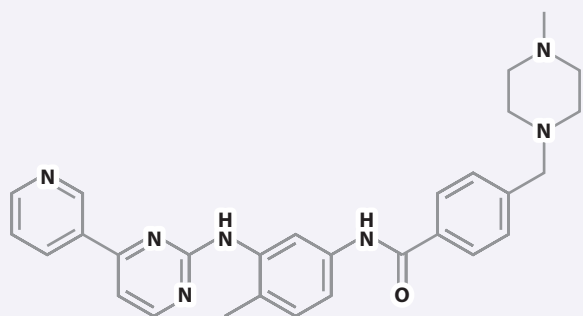
Box figure 7.7 – Joining the *BCR* and *ABL1* genes to make the chimeric *BCR-ABL1* gene.

The breakpoint in *BCR* can be in any of several different introns.

Treatment of CML was revolutionized in 2001 by the release of Imatinib (Gleevec, Box figure 7.8). Gleevec is a specific inhibitor of the BCR-ABL1 kinase. It induced remission in 70% of Ph¹-positive patients, in many cases allowing these elderly individuals to enjoy a normal lifespan and die *with* their disease rather than *of* it.

Imatinib set the pattern, and the hopes, for development of many subsequent targeted anticancer drugs (described further in Chapter 10). It also revealed the main limitation of these drugs: the tumor cells, with their genomic instability and rapid evolution, inevitably sooner or later develop resistance to the drug. A second-line drug is then needed to specifically target the resistant clones, and the whole scenario repeats.

In recent years, a number of germ-line mutations not associated with disease outside the hematopoietic system have been discovered, defining new leukemia predisposition syndromes including CML. These can have important clinical implications for management of the leukemia, as well as genetic counseling. *ANKRD26* variants are known to be associated with increased risk of thrombocytopenia, but case and family studies reveal histories that show increased risk of leukemia including CML. Germ-line mutations in *DDX41* define another disorder limited to malignant risk in the hematopoietic system with generally later onset. Malignancies include myelodysplastic syndrome, AML, CML, Hodgkin lymphoma, and non-Hodgkin lymphoma.



Box figure 7.8 – Chemical structure of imatinib (Gleevec).

7.5. References

- Albuquerque C, Breukel C, van der Luijt, R, et al.** (2002) The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. *Hum. Molec. Genet.* **11**: 1549–1560.
- Alexandrov LB, Nik-Zainal S, Wedge DC, et al.** (2013) Signatures of mutational processes in human cancer. *Nature*, **500**: 415–421.
- Alimirzaie S, Bagherzadeh M and Akbari MR** (2019) Liquid biopsy in breast cancer: a comprehensive review. *Clin. Genet.* **95**: 643–660.
- Beroukheim R, Murmel CH, Porter D, et al.** (2010) The landscape of somatic copy-number alterations across human cancers. *Nature*, **463**: 899–905.
- Campbell PJ, Getz G, Korbel JO, et al.** (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**: 82–93.
- Cancer Genome Atlas Research Network** (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, **45**: 1113–1120.
- Ciriello G, Miller ML, Aksoy BA, et al.** (2013) Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, **45**: 1127–1133.
- Evans DG, Harkness EF, Plaskocinska I, et al.** (2017) Pathology update to the Manchester Scoring System based on testing in over 4000 families. *J. Med. Genet.* **54**: 674–681.
- Finotti A, Allegretti M, Gasparello J, et al.** (2018) Liquid biopsy and PCR-free ultrasensitive detection systems in oncology (Review). *Int. J. Oncol.* **53**: 1395–1434.
- Hanahan D and Weinberg RA** (2000) The hallmarks of cancer. *Cell*, **100**: 57–70.
- Hanahan D and Weinberg RA** (2011) Hallmarks of cancer: the next generation. *Cell*, **144**: 646–674.
- Jansson MD and Lund AH** (2012) MicroRNA and cancer. *Mol. Oncol.* **6**: 590–610.
- Kinzler KW and Vogelstein B** (1996) Lessons from hereditary colorectal cancer. *Cell* **87**: 159–170.
- Martincorena I, Roshan A, Gerstung M, et al.** (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, **348**: 880–886.
- Martincorena I, Fowler JC, Wabik A, et al.** (2018) Somatic mutant clones colonize the human esophagus with age. *Science*, **362**: 911–917.
- Shen H and Laird PW** (2013) Interplay between the cancer genome and epigenome. *Cell*, **153**: 38–55.
- Skloot R** (2010) *The Immortal Life of Henrietta Lacks*. Crown Publishing Group, New York.
- Vogelstein B, Papadopoulos N, Velculescu VE, et al.** (2013) Cancer genome landscapes. *Science*, **339**: 1546–1558.

Useful websites

COSMIC database of genes mutated in cancer: <http://cancer.sanger.ac.uk/census>

International Cancer Genome Consortium: <https://icgc.org/>

Office of Cancer Genetics: <https://ocg.cancer.gov/>

7.6. Self-assessment questions

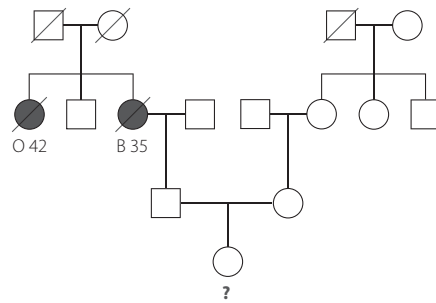
- (1) In relation to tumorigenesis, for each of the following statements decide if it is:
- (a) true of both oncogenes and tumor suppressor genes
 - (b) true of oncogenes but not tumor suppressor genes
 - (c) true of tumor suppressor genes but not oncogenes
 - (d) true of neither tumor suppressor genes nor oncogenes
- (1) May show nonsense mutations in sporadic cancers
- (2) May encode proteins involved in cell cycle regulation
- (3) Frequently mutated in familial and sporadic cancers
- (4) Often involved in loss of heterozygosity in sporadic cancers
- (5) Frequently mutated in familial but not sporadic cancers
- (6) May indirectly act to inactivate telomerase
- (7) Often involved in chromosomal rearrangements in sporadic cancers
- (8) May show inherited mis-sense mutations in familial cancers
- (9) Frequently mutated in sporadic but not familial cancers
- (10) May show gene amplification in familial and sporadic cancers
- (2) Blood and tumor DNA was extracted from two unrelated children with retinoblastoma and typed for a 2-allele DNA marker that maps close to the *RB* gene. One child has a family history of retinoblastoma while the other has a sporadic unilateral tumor. The genotypes were:

	Child A	Child B
Blood	Heterozygous 2–1	Homozygous for allele 1
Tumor	Homozygous for allele 2	Homozygous for allele 1

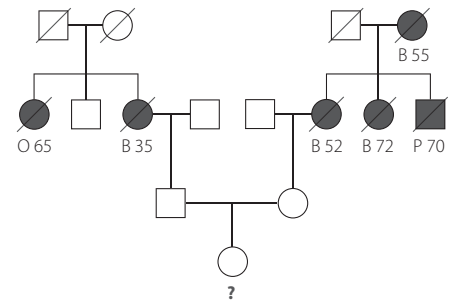
Mark each of the following statements as true or false:

- (a) the result indicates that Child B has the sporadic form of the disease
 - (b) the result indicates that Child A has the familial form of the disease
 - (c) the result indicates that the tumor in Child B could be either homozygous or hemizygous for the marker
 - (d) the result indicates that if Child A has the inherited form of the disease, he inherited it on the chromosome that carries allele 2 of the marker
- (3) Neurofibromatosis 2 (OMIM 101000) is the result of inherited and/or somatic mutations in the *NF2* tumor suppressor gene. Inherited *NF2* shows 90% penetrance. Assuming a mutation rate of 2×10^{-5} per gene per cell, what is the size of the target cell population? What is the expected incidence of sporadic *NF2*?

(a)



(b)



(4) The *Figure* shows two families in which there are several cases of cancer (B, breast; O, ovarian; P, prostate; numbers are the age at diagnosis). Which of the two women marked with a question mark would you prioritize for *BRCA1/2* mutation screening?

[Hints on questions 1, 3 and 4 are provided in the *Guidance* section at the back of the book]

08

How do researchers identify genes for mendelian diseases?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe the use of microsatellites and SNPs as genetic markers
- Describe the principles underlying positional cloning and autozygosity mapping
- Describe how whole exome or whole genome sequencing is used to identify disease genes
- Describe ways of picking the correct gene underlying a condition from a list of candidates
- Describe methods for functional testing of candidate genes

8.1. Case studies

CASE 18 CHOUDHARY FAMILY

- Baby girl Nasreen, healthy but deaf
- Multiply consanguineous family

207

216

240

395

Nasreen is the first child born to young parents (Aadnan and Mumtaz Choudhary) who are first cousins. The pregnancy and birth were normal. Nasreen was a healthy active baby, but she failed her newborn hearing screening test (*Figure 8.1*). This checked for otoacoustic emissions – sounds made by the hair cells of the cochlea in response to clicking noises played into the baby's ear. A different test, checking the auditory brainstem response, failed to pick up any neural response to sounds. She was referred to the audiology clinic where more refined versions of the two tests were used. These confirmed the results of the screening tests. Nasreen thrived; she smiled, held up her head and rolled over all at the normal times, but repeat audiological tests suggested she had a severe to profound bilateral hearing loss. The audiological team prepared to fit her with hearing aids and raised the possibility of giving her cochlear implants later on. Mumtaz and Nasreen were offered the option of talking to a geneticist, which they accepted gladly because, as discussed below, there were family issues.

The genetics clinic visit turned out to be a family affair because Aadnan's sister Benazir and Mumtaz's brother Waleed also asked to come, together with Aadnan, Mumtaz and Nasreen. Waleed was deaf, but was skilled at lip-reading, and with careful attention to seating and lighting he was able to play his full part in the consultation. The pedigree revealed an extensive family with several consanguineous marriages (*Figure 8.2*). In the UK pedigrees of this type are most often seen in people originating from the Middle East or the Indian subcontinent. Aadnan and Mumtaz are first cousins. Mumtaz's parents

(who are also first cousins) have four children in addition to Mumtaz. Two boys, Waleed and his brother Mohammed, are deaf and attend a college for deaf students where they are doing well academically. Aadnan and Mumtaz want to know the risk of deafness in any children they might have in the future. Mumtaz has had two miscarriages, and she is anxious about prospects for future children.

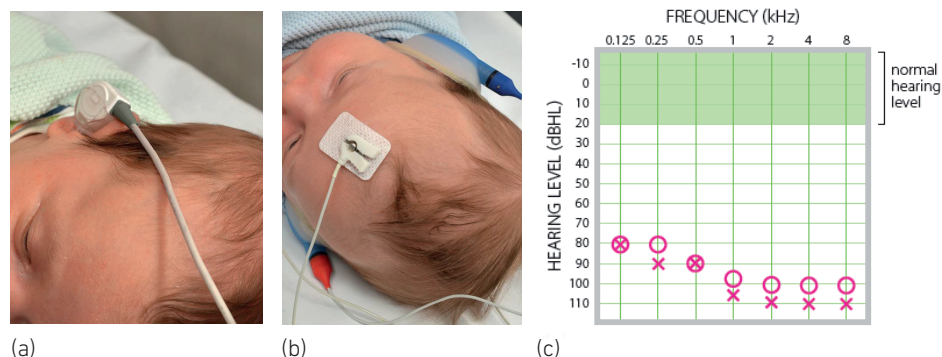


Figure 8.1 – Newborn hearing tests.

(a) Testing a baby's hearing by checking for otoacoustic emissions. (b) Testing the auditory brain stem response. Note the soft earphones. (c) Audiogram showing bilateral severe–profound hearing loss. The horizontal axis shows the frequency and the vertical axis the hearing threshold in decibels. Different symbols are used for readings from the two ears. 0–20 dB is normal hearing; hearing loss is defined as 20–40 dB (mild), 40–70 dB (moderate), 70–95 dB (severe), over 95 dB (profound). Photos (a) and (b) reproduced from the NHS leaflet *Screening tests for you and your baby* under the Open Government Licence v3.0.

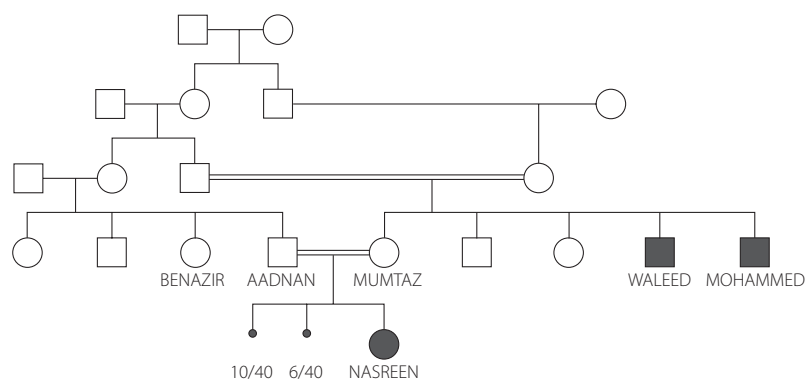


Figure 8.2 – Pedigree of Nasreen Choudhary's family; two early miscarriages are indicated.

8.2. Science toolkit

In this chapter we will focus on the question of identifying genes and validating them as the loci of variants that cause disease. In some branches of medicine there is a very sharp distinction between the day-to-day activity of diagnosis and the work of researchers investigating the causes of a condition. In genetics the distinction was always less clear. Clinical geneticists deal with a plethora of extremely rare and poorly understood

conditions, and they share with researchers a common interest in identifying the genes underlying genetic conditions. In the era of next-generation sequencing the distinction is particularly blurred: the laboratory procedure is identical whether seeking the cause of a patient's problem, as with **Karol Kowalski (Case 3)** or investigating a novel syndrome. The only difference comes once a likely gene has been identified: the researcher must show why that novel gene is a convincing candidate, while diagnosis can rest on established precedent.

Associating a phenotype with a DNA sequence variant

Genes can be defined in two very different ways:

- Genes inferred from the mendelian pedigree pattern shown by a character, usually a disease. Fifty years ago clinical geneticists were fully familiar with 'the Huntington disease gene', 'the cystic fibrosis gene', etc., and could offer their patients counseling based on the pedigree risk.
- Genes as functional DNA units performing a biochemical role in the cell.

The task of the researcher is to demonstrate a convincing association between a mendelian phenotype and a DNA sequence variant, and then to demonstrate why the biochemical action of that variant should lead to the phenotype. The first part of this task has become immeasurably easier over the past decades; the second part often remains challenging.

Over the years, many different strategies have been used to identify the gene mutated in a mendelian condition (*Table 8.1*). We will discuss each strategy in turn, but describe some only briefly because they are of historical interest only.

Table 8.1 Strategies used for gene identification

Strategy for identifying a gene	Period used
Via the gene product	Pre-1985
Through a gross chromosomal abnormality	1980 –
Through an animal model	1990 –
By positional cloning	1985–2005
By autozygosity mapping	1994 –
By exome or genome sequencing	2010 –

Identifying a gene through its product

In the very early days of molecular genetics, studies of proteins were more advanced than studies of DNA. Once the genetic code had been deciphered in the 1960s it was possible to work backwards from the amino acid sequence of a protein to the DNA sequence of the gene that encoded it. A probe could then be synthesized to isolate the gene or mRNA by hybridization. However, given the degeneracy of the genetic code (most amino acids can be encoded by more than one codon) any probe had to consist of a cocktail of possible sequences. Other possible approaches included raising an antibody to the protein and using that to precipitate ribosomes synthesizing the protein,

together with the associated mRNA – that is how Savio Woo and colleagues isolated the phenylalanine hydroxylase gene in 1983 (Woo *et al.*, 1983). Sometimes a mRNA could be isolated directly if a certain cell type mainly made the target protein in large quantities – globins in reticulocytes, for example, or ovalbumin in oocytes.

Identifying a gene through a chromosomal abnormality

Sometimes a patient has a *de novo* case of a dominant or X-linked condition, together with a *de novo* chromosome deletion, translocation or inversion. That might just be coincidence but, alternatively, the deletion might take out the relevant gene, or one of the breakpoints of a chromosomal rearrangement might disrupt it. Two competing research groups managed to clone the dystrophin gene in 1985, one through a deletion in an affected boy and the other using a translocation (Hoffman, Brown & Kunkel, 1987; Ray *et al.*, 1985). At the time these were groundbreaking achievements. Since then a number of other high-profile genes have been identified by this route, and it remains an important way by which an alert clinician can contribute to research. For example, Kurotaki and colleagues (2002) identified the gene for Sotos syndrome (OMIM 117550) at chromosome 5q35 by cloning the breakpoints of a *de novo* reciprocal translocation 46XX t(5;8)(q35;q24.1) in a patient with *de novo* Sotos syndrome. Having identified that the *NSD1* gene was disrupted in this particular patient, they proved that variants in this gene caused the syndrome by identifying *NSD1* deletions or point mutations in a series of unrelated Sotos patients.

Identifying a gene through an animal model

Most human genes have an exact counterpart in mice and other animals, and it is easier to identify genes in laboratory animals. Controlled breeding means genes can be mapped (for positional cloning as described below) far more easily and accurately than in humans, and techniques like mutagenesis by chemicals or radiation can be used. Having identified a mouse gene, the DNA can be used as a probe to isolate the corresponding human gene. The main challenge for human gene identification by this route is relating animal phenotypes to specific human conditions – they do not always correspond closely. A successful example is the identification of *SOX10* as the gene responsible for Type 4 Waardenburg syndrome (OMIM 613266), based on study of the *Dominant megacolon* mouse mutant (see Pingault *et al.*, 1998).

Identifying a gene by positional cloning

This was the main strategy by which the majority of the genes underlying mendelian conditions were identified, mostly in the 1990s. The overall process is illustrated in *Figure 8.3*. The near-disappearance of such family studies (apart from autozygosity mapping, see below) from current research is not so much because they have been rendered obsolete by advances in sequencing, but more because virtually every condition where extensive



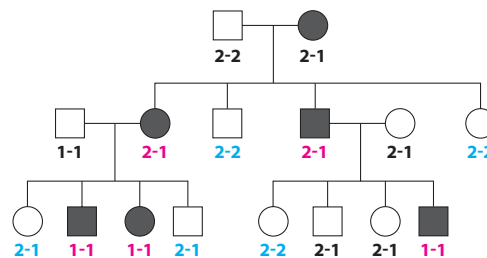
Figure 8.3 – The principle of positional cloning.

pedigrees suitable for linkage analysis were available had already been studied and the disease gene identified. What remained were the highly heterogeneous, excessively rare or sporadic conditions where family linkage studies were impossible.

The essential step, having found a suitable family and obtained DNA from as many family members, both affected and unaffected, as possible is to type the family members for a large panel of DNA markers, looking for a marker that exactly tracks the unknown disease gene through the family. If this happens to a degree that is too significant to be just coincidental, the unknown disease locus must lie close to the marker locus on the same chromosome. The markers can be any polymorphic locus that has a known chromosomal location and more than one common allele, such that there is a reasonable chance that a randomly selected individual would be heterozygous (Box 8.1). Figure 8.4 shows how the process works.

Figure 8.4 – Part of a large kindred where an autosomal dominant disease and a two-allele marker are segregating.

See text for discussion.



The figure shows part of a hypothetical large kindred in which an autosomal dominant disease (shaded characters) is segregating. Family members and their spouses have been typed for a two-allele marker, a non-pathogenic variant where individuals could have genotypes 1–1, 2–1 or 2–2. Individuals whose genotypes are marked in red can be seen to have inherited marker allele 1 with the disease. Those marked in blue inherited marker allele 2 with the normal (non-disease) allele at the disease locus. No deduction can be made for those whose genotypes are shown in black. Note that individuals III-6 and III-7 could have inherited marker allele 2 from their mother and allele 1 from their father, or *vice versa*. Thus although we know they inherited the non-disease allele at the disease locus from their father, we cannot tell which marker allele they inherited from him. There are 10 cases where we can work out the co-segregation (colored genotypes), and in each case either marker allele 1 has segregated together with the disease allele, or marker allele 2 has segregated with the non-disease allele. This data suggests the two loci may be linked, and hence lie together on some particular chromosome segment. A statistical test, the lod score (see www.scionpublishing.com/NCG4 “Resources” for more details about this, or *Chapter 17* of Strachan & Read, 2019) is used to confirm linkage. Knowing the location of the marker (readily established, for example, by FISH), we have identified the chromosomal location of the disease gene.

Note that it is not the case that all affected individuals have the same marker genotype, while all unaffected have a different genotype. What matters is the way alleles segregate through the pedigree. This is an example of **linkage analysis**. Readers may find the reasoning above quite subtle. Because linkage analysis, though fascinating, is no longer in the mainstream of human gene identification, we shall not spend more time on it here. Interested readers can find an extended discussion of a real example, from the paper by Miyamura *et al.* (2003) on www.scionpublishing.com/NCG4 – see “Resources” for access to the text.

Genetic markers

Genetic markers are used to track the transmission of a chromosomal segment through a pedigree or a population. A useful marker must be polymorphic, that is it must have common allelic variants, and mendelian, i.e. determined by variation at a single chromosomal location. In the distant past blood groups, serum enzyme variants and tissue types were tried as markers. DNA variants replaced these, being more numerous, easier to genotype by uniform procedures, showing codominant inheritance and having readily determined chromosomal locations. Commonly used DNA markers are of two types.

- **Microsatellites** are short tandem repeats (for example, (CA)_n sequences) that are present in everybody at the same specific chromosomal locations, but where the number of repeat units varies from person to person. Around 150 000 microsatellites have been documented in the human genome. They are genotyped by PCR-amplifying a segment of DNA containing the microsatellite and determining its size by gel electrophoresis (often using fluorescent labeling and a DNA sequencing machine). They have the advantage of having many possible alleles (that is, people might have 5, 6, 7... etc. repeats), increasing the likelihood that a meiosis will be informative (i.e. allow the parental origin of each allele to be determined).
- **Single nucleotide polymorphisms (SNPs)** (also termed single nucleotide variants, SNVs) are less informative, because there are almost always only two alternative nucleotides at a polymorphic site. However, they are abundant (over 10 million common SNPs are catalogued in the dbSNP database) and they can be typed on SNP chips (see *Chapter 4*), allowing SNPs covering the whole genome to be genotyped in a single operation. Some SNPs create or abolish a recognition site for a restriction enzyme, thus creating a restriction fragment length polymorphism (RFLP, see *Figure 5.8*). RFLPs were the original DNA markers, superseded by microsatellites and SNPs for general mapping, but still useful because they are easy and convenient to type in a small laboratory.

(a)

ctctcacagt	agc	ca	ca	ca	ca	ca	ac	cc	gctgc	ac	ag	cggcct	n=5	
ctctcacagt	agc	ca	ca	ca	ca	ca	ac	ca	ccgct	gc	ac	agcggc	ct	n=6
ctctcacagt	agc	ca	ca	ca	ca	ca	ac	ca	ca	ccg	ctgc	acagcg	gcct	n=7

(b)

tttccttcc	atgggtgat	a	ttgcttcttg	aaatacggac	A
tttccttcc	atgggtgat	c	ttgcttcttg	aaatacggac	C

Box figure 8.1 – Common types of DNA polymorphism. (a) A (CA)_n microsatellite, (b) a A/C SNP.

Markers are used in disease studies to identify chromosome segments shared by relatives or by apparently unrelated affected people in order to map an unknown disease gene or genetic susceptibility factor, as described here and in *Chapter 13*. They can also be used to track the segregation through a family of a chromosomal segment known to carry a pathogenic allele, in order to provide risk estimates for relatives ('gene tracking') – a method obsolete for general risk estimation now that we can directly test for pathogenic variants for most mendelian diseases, but that still has some applications. One such was shown in *Chapter 4* (see *Figure 4.15*) where gene tracking was suggested as a way to identify fetuses of **John and Joan Ashton (Case 1)** that are at risk of Huntington disease, without revealing whether or not John, who is at 50% risk from the pedigree, has actually inherited the disease allele from his affected father. Gene tracking is also sometimes used in pre-implantation diagnosis (*Box 14.4*). Rather than identify the specific pathogenic change in an affected or at-risk parent, and then have to develop a specific assay to test for it in a single cell taken from a pre-implantation embryo, a laboratory may prefer to rely on gene tracking using a set of well-validated markers for the family disease.

Further uses of markers for identifying a person's ancestry or finding unknown relatives are beyond the scope of this book, but one such application, for catching criminals, is briefly discussed in *Section 9.4*.

Mapping the locus for a disease (finding its chromosomal location) is the first step in positional cloning (*Figure 8.3*). How closely a locus could be mapped depends partly on the size of the family collection, and partly on luck. A large collection (ideally one extremely large family, but alternatively a collection of independent smaller families where the clinician is confident they all have the same condition) allows the segregation of the disease gene and the marker to be studied in many meioses. If the lod score confirms beyond reasonable doubt that the marker and disease loci are linked, attention then moves to the rare recombinant meioses where they do not co-segregate. These can be used to narrow down the candidate region. For interested readers, see www.scionpublishing.com/NCG4 “Resources” to show this process at work. Typically, family studies of this sort might define a candidate region of 1–5 Mb.

Having identified the chromosomal location of the disease gene it is necessary to clone it. In the early days this was a major challenge, taking years of work by highly skilled postdoctoral scientists to identify all the genes in the candidate region. With the availability of the Reference Human Genome it became immeasurably easier – a list of candidate genes can now simply be downloaded. They can then be prioritized for mutation testing as described in *Section 8.4*. Ultimately, successful identification of the disease gene depends on demonstrating that this gene is mutated in all or most of a panel of unrelated individuals affected by the same genetic condition.

Identifying a gene by autozygosity mapping

Linkage analysis, as in *Figure 8.4*, works well for dominant or X-linked conditions, provided suitably large families can be found. For autosomal recessive conditions large extended pedigrees are rarely available. A modified approach, **autozygosity mapping**, has proved very effective in inbred families or communities, and because it does not require large extended kindreds, it is still widely used today. *Figure 8.5* shows the principle.

Autozygosity means homozygosity for alleles identical by descent. The (hypothetical) autosomal recessive disease in *Figure 8.5* is rare. It is therefore highly likely that the three affected children all inherited both their disease alleles from one or other of their great-grandparents in generation I, and that the individuals marked with dots are all heterozygous carriers. If that is the case, all three affected individuals will be homozygous

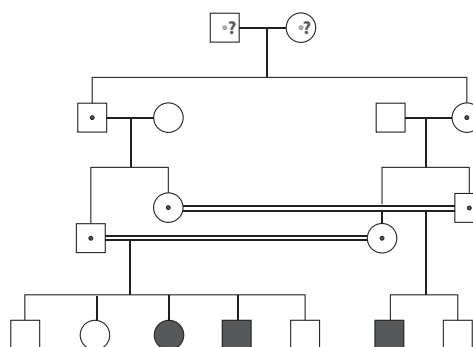


Figure 8.5 – In this inbred family individuals IV-3, IV-4 and IV-6 all suffer from the same rare autosomal recessive disease.

for the same disease allele. They will also be homozygous for all the non-pathogenic markers carried on that same chromosomal segment. If the three are typed for a panel of polymorphic markers spanning the whole genome, the disease locus can be mapped by looking for a region in which all three are homozygous for the same marker alleles. As with positional cloning, having mapped the disease locus to a specific chromosomal region, it is then necessary to check genes in that region for mutations, in the affected children or in unrelated children with the same disease.

Autozygosity mapping requires several, preferably fairly distant, affected relatives. The offspring of first cousins would on average have 187 Mb of homozygosity (1/16th of the genome – see *Chapter 9* for the calculation) by virtue of the consanguinity, plus any other coincidental runs of homozygosity. That is far too much to pinpoint the location of any gene, but comparison with any other affected relatives would narrow down the field.

An alternative strategy is to compare several unrelated individuals, each the product of a consanguineous marriage and each with the same rare recessive condition. Though they might well have different pathogenic variants, those should all be in the same gene, and so all should have a run of homozygosity at the same chromosomal location. Being unrelated and with independent mutations, they would probably not share the same marker alleles at the relevant location, but each affected person should be homozygous across this region for whichever marker alleles they carry. If they are unrelated, there should be few other homozygous regions that they all share by chance, so it should be relatively easy to pick out the relevant chromosomal segment. The key limitation is that the same gene must be responsible for the condition in each affected person. Sometimes the same clinical condition can be caused by loss of function in any one of several genes – the clinical condition might be the result of failure of a multi-gene pathway, and look the same regardless of which gene in the pathway had caused it to fail. If there were such **locus heterogeneity** among the test cases this autozygosity mapping strategy would fail.

Identifying a gene by exome or genome sequencing

We saw in *Chapter 5* how next-generation sequencing allows a researcher to list all the DNA sequence variants present in a patient. This is now the default pathway for identifying disease genes. The principle has already been set out in the case of **Karol Kowalski (Case 3)**. Although that case was about diagnosis, and this chapter is more about research, in this area the two follow identical procedures. The only difference is that in Karol Kowalski's case the geneticists were confident that the correct causative gene had been identified when they found that other similarly affected cases had been reported who had variants in the same gene. If the research procedure implicates a novel gene, so that there is no precedent, it will require extra evidence, probably in the form of functional studies, to convince a skeptical world that the right gene has been found. *Figure 8.6* shows some of the strategies used by researchers to try to identify a novel disease gene from exome or genome sequencing data.

- Strategy A uses a linkage approach. The example shown uses a multi-case inbred family to identify the cause of an autosomal recessive condition. One could equally use families in which a dominant or X-linked condition was segregating, as described above. If it is possible, one way or another, to map

the unknown gene to a specific chromosomal location, that might reduce the number of candidate genes from 20 000 to maybe 20. Much less stringent supporting evidence is needed to make the case for a candidate gene selected from a field of 20 than for one selected from a field of 20 000.

- Strategy B involves showing that the same gene is mutated in independent cases of the same recessive disease. Patients might be homozygotes or compound heterozygotes (two different variants in the same gene); either way they have both copies of the same gene mutated.
- Strategy C shows a similar procedure being used for a dominant condition. Hoischen and colleagues in 2010 identified *SETBP1* as the gene responsible for the rare dominant Schinzel–Giedion syndrome (OMIM 269150 – note that numbers starting with 2 in OMIM normally denote a recessive condition, but the work of Hoischen and colleagues showed that this suspected recessive condition is actually dominant) by demonstrating heterozygous mutations in 12 unrelated cases (see Hoischen *et al.* (2010)).
- Strategy D shows the approach used in the case of **Karol Kowalski (Case 3)**. It has become apparent that many sporadic conditions are caused by *de novo* dominant mutations. Across the whole genome a person might have around 70 *de novo* mutations, but typically there would only be 1–2 that changed a protein. Thus parent–child trios are a very powerful resource for gene hunters.

The reviews by Boycott and colleagues (2019) and Shendure and colleagues (2019) summarize some of the approaches and issues.

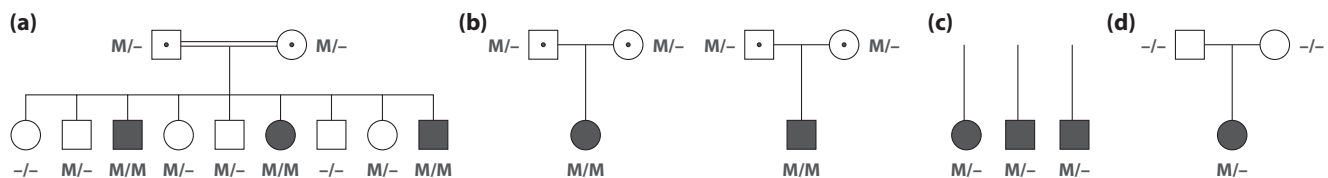


Figure 8.6 – Four strategies to increase the chance of identifying the gene underlying a mendelian condition.

M signifies the pathogenic variant; – the corresponding normal allele. See text for discussion. Reproduced from Strachan & Read (2019) *Human Molecular Genetics* 5e, with permission from Garland Science/Taylor & Francis LLC.

Demonstrating why a variant causes a phenotype

Once a candidate variant has been identified, it is necessary to show a plausible reason why it might cause the patient's condition. Where the nature of the condition is closely linked to a biochemical abnormality – with inborn errors of metabolism or hemoglobinopathies, for example – this is straightforward, but for most genetic diseases the connection may be far from obvious. In very many cases the biochemical and physiological story is still incomplete, while with behavioral phenotypes the work has scarcely begun. Often, as with **Case 3 (Karol Kowalski)**, the only line of evidence is that variants in the same gene can be seen in unrelated patients with the same clinical phenotype. Matchmaker Exchange (www.matchmakerexchange.org/) is a very useful resource for finding such cases worldwide. Possible ways of moving forward are discussed in Section 8.4.

8.3. Investigations of patients

CASE 18 CHOUDHARY FAMILY

- Baby girl Nasreen, healthy but deaf
- Multiply consanguineous family
- Autozygosity mapping
- Exome sequencing
- A second recessive condition?

207

216

240

395

As it turned out, all the problems in this family were due to known genes. However, the procedures used for the investigation would equally have revealed any novel gene involved. The only difference would come after the initial gene identification where, as described above, efforts would be needed to try to find other affected cases, and functional and other studies would be needed to make the case convincing.

Congenital deafness can be due to environmental factors such as maternal rubella or birth trauma. The geneticist had checked Mumtaz's notes to make sure there was no evidence of such problems with Nasreen, but she noted the need to check with Mumtaz's mother and her family doctor that there was no evidence of such factors with Waleed or Mohammed. The pedigree is most simply interpreted as showing autosomal recessive deafness. About two-thirds of congenital deafness has such a cause. If this interpretation is correct, then each subsequent child of Aadnan and Mumtaz has a 1 in 4 risk of being deaf. The extensive inbreeding is an additional pointer to a recessive condition. Although consanguineous marriage roughly doubles the risk of abnormal children (depending on the degree of consanguinity – see *Chapter 9*), the increase is only from around 2–2.5% to around 4–4.5%. In other words, the chance of a healthy child is only reduced from 98% to 96%. In many traditional societies there are good social reasons to marry a relative rather than a stranger. Mumtaz's two miscarriages are unlikely to be related to the deafness. They might be evidence of another, much more severe recessive condition in the family, but there are many non-genetic causes of miscarriage. The geneticist reassured Mumtaz that two early miscarriages is quite a common history that usually has no sinister import (see *Box 14.1*).

The presence of Waleed and Benazir at the clinic visit was explained by the fact that they have been introduced and a wedding is planned. Benazir was worried that if a child of theirs was deaf, it might relate exclusively to Waleed and she would be cut out of the family circle. Waleed, however, took a much more relaxed view. The geneticist had to be sensitive to the undercurrents and careful not to try to impose her own views. Consanguineous marriage within a family with a known recessive condition does carry specific risks. On the recessive hypothesis, Benazir's brother Aadnan (Nasreen's father) was a carrier. Therefore one of Benazir's and Aadnan's parents was a carrier – presumably their mother, since she was the link into the other side of the family. Thus Benazir's carrier risk would be 1 in 2 and the overall risk that a child of hers and Waleed's would be deaf was 1 in 4. This calculation showed that there was a large uncertainty in the risk – high if the family condition was genetic and recessive, much lower if not. The family were keen to know if a DNA test was available to confirm the interpretation.

At present it was still only a hypothesis that Nasreen's deafness was autosomal recessive, or even that it was genetic at all, since there are so many different causes of hearing impairment. The fact that she had two affected uncles, and that all were offspring of consanguineous marriages, strengthened the case for it being autosomal recessive, but the only way to prove it would be to demonstrate a pathogenic variant. A standard resource, the Hereditary Hearing Loss Homepage (<https://hereditaryhearingloss.org/>) lists 75 genes where variants can cause autosomal recessive non-syndromic hearing loss. How best to approach such heterogeneity?

Although any one of 75 genes – or a novel gene – might be responsible, in most populations mutations in one gene, *GJB2*, are much the most frequent cause. This gene encodes the connexin 26 protein that forms gap junctions between cells in the inner ear. These are important for recycling potassium ions, which enter the hair cells when they fire in response to sound. *GJB2* is a very simple gene, having just two exons, and all the coding sequence is in exon 2. In specific populations particular variants are common – c.35delG in Europeans (see *Figure 6.5*) and c.235delC in Chinese, for example. A simple targeted test can check for the commonest variant in the patient's ethnic group, and it would not be difficult to sequence the whole small gene, so this is often performed as a first test. It provides a definite answer in up to half of all cases. Thus as a first approach the whole *GJB2* coding sequence was sequenced in Nasreen's DNA, but no mutation was found in either copy.

Two alternative strategies were possible to identify the family variant: autozygosity mapping or exome sequencing. Here we show both approaches to illustrate the processes.

Identifying the gene by autozygosity mapping. As described in the previous section, this rests on the assumption that one or other of Nasreen's great-great-grandparents had carried a pathogenic variant in heterozygous form, and that Nasreen, Waleed and his brother Mohammed had each inherited two copies of that original variant as a result of the consanguinity in the family. If they did indeed all have two copies of that ancestral chromosome segment, they should all be homozygous for the same set of marker alleles on the relevant chromosomal region. High resolution SNP chips (see *Figure 4.13*) provide a convenient tool for autozygosity mapping. As shown in *Figure 8.7*, samples from Nasreen, Waleed, Mohammed and two unaffected sibs were genotyped on SNP arrays, revealing a shared region of autozygosity at chromosome 3p21.

The Hereditary Hearing Loss Homepage provides lists of the locations of all known genes where variants cause hearing loss. Loci *DFNB1–DFNB108* host variants causing non-syndromic autosomal recessive hearing loss and so are possible candidates for the Choudhary family condition (there are only 75 genes: some DFNB numbers are duplicates or are reported loci where the gene responsible has not been identified). The *DFNB6* locus mapped to chromosome 3p21, the location identified by the autozygosity mapping.

The gene responsible for *DFNB6* hearing loss had already been identified as *TMIE* (transmembrane inner-ear expressed gene; OMIM 607237). *TMIE* deafness is quite frequent in inbred kindreds in southeastern Turkey, where affected individuals are homozygous for the same pathogenic variant, p.R84W, presumably inherited from a remote common ancestor. Otherwise only a handful of families had been described with deafness caused by *TMIE* mutations, and without the evidence from the autozygosity mapping, *TMIE* would have been low on the list of candidate genes to test. The four exons of this gene were sequenced in Nasreen's DNA. She was homozygous for a frameshifting change, an insertion of four nucleotides in exon 2. Having identified the family variant, the geneticist was now in a position to offer definitive testing to Waleed, Benazir and any other family members who wished it.

Other than in Turkey only five families had previously been shown to have *TMIE* mutations, all from south India, and one of them had the same sequence change as Nasreen. If DNA from an affected member of that family were tested, it would be possible to work out whether both families had inherited the variant from an unknown common ancestor,

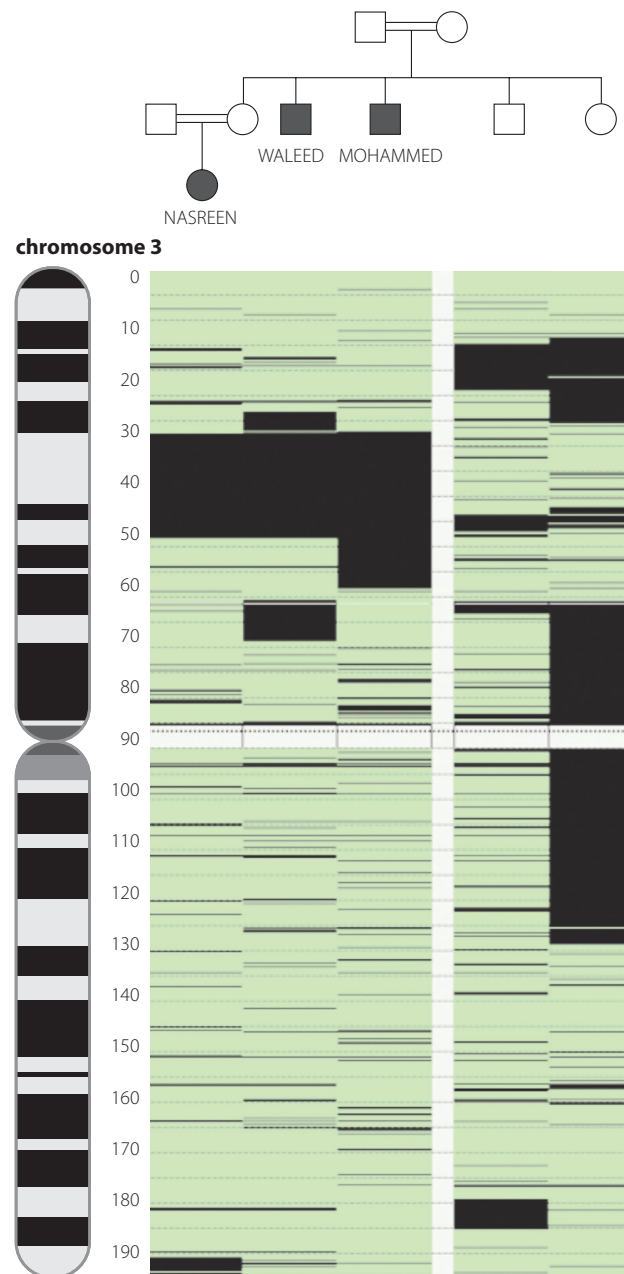


Figure 8.7 – Autozygosity mapping in the Choudhary family.

The three blacked-in individuals have profound hearing loss, presumed to be autosomal recessive (see *Figure 8.2*). DNA from these three and two unaffected sibs was genotyped on high-density SNP arrays. The graphic shows the results for chromosome 3. SNPs are represented by horizontal lines at the appropriate chromosomal location. Heterozygous SNPs are shown in green, homozygous ones in black. The white area marks the centromere. The three affected individuals are all homozygous for a contiguous set of SNPs across the 3p21 region, while the unaffected sibs are heterozygous for most of these SNPs. No other region of shared homozygosity was seen on other chromosomes. It is likely that this region contains the gene responsible for their hearing loss. Figure courtesy of Jill Urquhart, St Mary's Hospital, Manchester.

or whether there had been two independent mutations. If they had the same ancestral chromosome segment, they would be expected to have the same haplotype of SNPs immediately surrounding the *TMIE* gene.

Identifying the mutation by exome sequencing. The procedure here follows that described for **Karol Kowalski (Case 3)**, see *Section 5.3*). Briefly, a sample of Nasreen's DNA was randomly fragmented by sonication, adaptors ligated to the fragments and exons captured. The amplified exons were sequenced to an average depth of 80x on a next-generation machine. The raw reads were aligned to the human Reference Sequence and some 20 000 variants that passed quality control measures were listed in a file. At this point a different analytical strategy was used. Only variants that might be responsible for Nasreen's hearing loss were picked out for analysis. Genomics England had defined a virtual gene panel of 356 genes associated with hearing loss (<https://panelapp.genomicsengland.co.uk/panels/>), but in Nasreen's case the analysis was limited to just the *DFNB* genes listed by the Hereditary Hearing Loss Homepage (<https://hereditaryhearingloss.org>) as harboring variants causing autosomal recessive non-syndromic hearing loss. This simplified analysis rapidly identified the 4 nucleotide insertion in exon 2 of the *TMIE* gene. Using a gene panel like this limits the analysis to known candidate genes and would miss any case due to variants in a novel gene. However, had the result been negative it would have been simple to extend the search to the whole exome: this was a *virtual* gene panel. The whole exome sequence was present in the laboratory computer; just the analysis had been restricted to those 75 genes.

Nasreen was given cochlear implants and over time she acquired a useful degree of hearing. However, when she was 4 years old her parents had a new concern. Her teeth had now erupted but instead of the gleaming white teeth Mumtaz and Aadnan expected they were a yellowish brown color, and they seemed to be very sensitive. Nasreen cried and said her teeth hurt when she ate ice cream or anything that was too hot. They had taken her to the local dentist, who said that her tooth enamel was softer than normal. He had referred Nasreen to the pediatric dental service who found she had cracks in the enamel of several teeth, and on X-ray there was a lack of contrast between enamel and dentine. The pediatric dentist said this was a condition called amelogenesis imperfecta (*Figure 8.8*).

The pediatric dentist wondered whether there was a link between Nasreen's deafness and her dental abnormality, because this tooth condition is often seen with other disorders. Dental genetics is a subspeciality in its own right, because variants in many different genes



Figure 8.8 – Amelogenesis imperfecta.

Photo by Peter JM Crawford and reproduced here from Wikipedia under a Creative Commons CC BY 2.0 licence.

can affect tooth development, so she called on a specialist colleague for help. Their first question was whether Nasreen's deaf uncles, Waleed and Mohammed, had poor teeth. Mumtaz produced family photos showing fine white teeth. She did comment that one distant relative, Walid, had terrible teeth rather like Nasreen's, but she had always assumed that was just due to poor dental care. The dentists checked the reports of individuals previously shown to be deaf due to *TMIE* variants, but the limited information did not indicate any dental problems. The *TMIE* gene had been knocked out in mice as part of the systematic International Mouse Knockout Consortium (described below); those mice had been systematically examined in many ways, and their dentition was entirely normal. None of this constituted definitive refutation of the idea that Nasreen's poor teeth were due to her *TMIE* mutation, but they made it unlikely. It thus seemed possible that there was an independent cause and, given that Walid had similarly poor teeth, maybe there was a second autosomal recessive condition segregating in the family.

If there was a second condition there should be a second region of homozygosity in Nasreen. However, due to the consanguinity, Nasreen had extensive areas of homozygosity, and without the comparison with Waleed and Mohammed it was not possible to identify the specific region involved. Instead the investigators turned to the exome data. Although the list of variants in Nasreen's exome had only been checked for variants known to be associated with hearing loss, the entire list was still available on the laboratory computer. The list was filtered for variants that would affect a protein (missense, nonsense or splice site variants) and that were present in homozygous form. One candidate variant caught their attention. This was a nonsense change, p.W153X in the *KLK4* (kallikrein 4) gene at chromosome 19q13. Homozygous variants had been reported to cause a very similar type of amelogenesis imperfecta (OMIM 204700) in three families. As expected, neither Waleed nor Mohammed was homozygous for this variant (variants at 19q13 would segregate independently of those at 3p21 linked to the *TMIE* gene). With this result in hand, Walid was approached, agreed to be tested, and was shown to be homozygous for the same variant. He was pleased to be free of insinuations that his poor teeth were because he had been neglecting his dental hygiene. His result confirmed that two different recessive conditions were segregating in the family. This is not an uncommon finding in consanguineous families, and can cause difficulty in disentangling the phenotypes.

8.4. Going deeper...

When variants in a novel gene are proposed as the cause of a genetic condition, some extra evidence is needed beyond the initial gene identification. This is especially the case for genes identified through exome or genome sequencing. When positional cloning or autozygosity mapping are used, a candidate gene is probably one of only a dozen or so at the relevant chromosomal location, so only a modest amount of supplementary evidence may be required to make a convincing case. A candidate gene proposed on the basis of exome or genome sequencing is one among 20 000. Very good evidence is necessary to convince the world that the correct gene has been identified. It would be very useful to show that independently ascertained unrelated individuals with variants in the same gene have the same phenotype. Matchmaker Exchange (www.matchmakerexchange.org/) could help ascertain such cases, but that will not always be possible. Even if other

cases can be identified it would add conviction to be able to suggest why variants in that gene should cause the condition in question.

Fortunately, sources of supplementary evidence have multiplied hugely in the past decades. Vast amounts of relevant information about any gene are available on the internet before there is any need to get one's hands dirty in the laboratory. Examples include:

- The ENSEMBL or UCSC genome browsers will have information on the exon–intron structure of the gene, a catalogue of all transcripts and splice isoforms and information on the chromosome region.
- The ClinVar database (www.ncbi.nlm.nih.gov/clinvar) is a freely available public archive of human genetic variants, together with interpretations of their significance to disease. For example, searching on 'ARID1B[gene]' brought up a list of 451 entries that have been submitted by a variety of laboratories or extracted from publications, each reporting a variant in the *ARID1B* gene, together with the submitter's assessment of whether it is benign or pathogenic, and a note of the condition of the subject, if affected. Searching on "cystic fibrosis"[dis] brought up 1458 records of variants reported in individuals with cystic fibrosis, again with assertions about the pathogenicity of each variant [searches conducted 31 May 2019].
- The GnomAD database (<https://gnomad.broadinstitute.org/>) lists variants found in 125 748 exomes and 15 708 genomes from unrelated individuals sequenced as part of various genetic studies. Individuals with severe pediatric disease have been excluded so that, at least for early onset disease, it can be used (with caution) as a database of ostensibly healthy individuals. For example, searching on ARID1B produced a list of 2605 variants, with a count of the frequency of each variant in the data, and some annotation. It also reported that only four loss of function variants were seen, compared to an estimate of 91 if such variants were non-pathogenic in heterozygotes and so occurred at random among the data subjects. That suggests that being heterozygous for a loss of function variant is probably incompatible with featuring in this database of more or less healthy individuals. For a recessive condition like Nasreen's, one would look for a deficiency of homozygotes, relative to the frequency in the database of heterozygotes for loss of function variants.
- Genecards (www.genecards.org/) provides a wealth of information on a gene and its product, including a free-text description of its function, a catalog of reported regulatory elements, description of the pathways in which it is involved, a list of animal orthologs, the subcellular localization of the protein and extensive data on the expression in a large number of tissues.

Naturally some genes will have fuller information than others, but because in recent years many experiments have been run on a genome-wide rather than a targeted basis, there is often quite a lot of information even on quite obscure genes. Using all these resources it should usually be possible to decide whether a candidate gene is a plausible contender for a patient's condition. At least the gene should be expressed in the tissues involved. It is more difficult to decide whether the known function of a candidate gene is compatible with the patient's condition, because very often we lack sufficient understanding of the meshwork of interactions and functions within cells.

If all these internet searches fail to provide a definite answer it may be necessary to produce one's own experimental data. Over the past decade or so our ability to manipulate genes and cells, both in tissue culture and in whole living organisms has seen very significant progress. Two key developments have been:

- **The development of a simple and efficient technique for gene editing.** What was once a difficult and laborious process suddenly became very much cheaper and simpler with the development of the CRISPR-Cas system (see Charpentier & Doudna, 2013). Using this technique any desired change can be made to the genomic DNA of living cells. The technique is not fully mature, and attempts to use it clinically to edit the human germline are strongly deprecated at this stage of our understanding, but it has revolutionized the scope of experiments in tissue culture or model organisms.
- **The development of iPSC (induced pluripotent stem cells).** Suppose one wished to study a certain type of brain cell from a patient carrying a specific pathogenic variant: until recently it would have been impossible to obtain such cells, assuming a brain biopsy was out of the question. Now there are two ways of obtaining the cells. If cells of the right type (but lacking the desired genetic variant) could be obtained, perhaps from an autopsy or from a patient undergoing brain surgery, CRISPR-Cas might be used to create the variant. Alternatively, the patient might agree to donate cells of a readily obtainable type such as fibroblasts. These could be manipulated in cell culture to produce iPSC – a fairly standard procedure – which might then be induced to differentiate into brain cells of the desired type (see *Chapter 14* for more detail).

Such cells can be used to study how a candidate variant affects the expression of its gene or other genes with which the host gene interacts. The subcellular location of the gene product can be studied and any changed gene interactions or cellular responses noted. Further developments in cell culture have made it possible in many cases to develop organoids. These are three-dimensional clusters of cells that in some important respects mimic biopsies of whole organs such as kidney, intestine or brain. All these techniques are well beyond the remit of diagnostic laboratories, but are enabling researchers to ask in great detail about the precise effect of any variant.

Some effects can only be studied in intact organisms. Extensive resources of loss of function variants exist in several widely studied model organisms – for example, the International Mouse Knockout Consortium (see www.mousephenotype.org/about-impcc/about-ikmc) systematically inactivated 18 500 mouse genes, representing over 90% of all mouse protein-coding genes, and has established a bank of frozen embryonic stem cells with each inactivated gene. For most, though not all the genes, work progressed to creating living mice carrying each inactivated gene and went on to study the phenotype in detail.

In practice, laboratory work is more likely to be needed for investigating gain of function than loss of function variants. Gain of function variants are more specific and may need to be created in a suitable organism. Hopefully the mutants will have a phenotype that can be related to the phenotype of the patients studied. When deciding which model organism to use, a balance must be struck between the relevance to humans and the ease and cost of using any particular organism. Mice offer the closest homologies to humans, but are difficult and expensive to work with, quite apart from the ethical

issues of experimenting on mammals. *Drosophila* flies avoid these problems, but may be too different from humans for these purposes. Zebrafish (*Danio rerio*) are a popular compromise. Being vertebrates, they have more in common with humans than do fruit flies or worms, but they are much more convenient to work with than mice. They can be readily bred in large numbers and their free-living transparent embryos can be studied without the need for dissection.

The diversity of techniques for functional studies merits a book on its own. All these topics go well beyond the scope of this book. They are covered in rather more detail in the textbook by Strachan & Read (2019). Readers looking for more detail should seek out a good recent review. The overall message is that final proof (and publication) of a disease gene identification usually requires functional studies.

What happens if exome sequencing does not identify a candidate gene?

Reading published papers can give the impression that exome sequencing always identifies the causative gene. That is only because unsuccessful research does not get published. In reality the diagnostic yield from next-generation sequencing is typically in the range of 30–60%. So why is it not 100%? Possible explanations include:

- **The sought-for variant does not exist.** Maybe this person's condition is not caused by the sort of single high-penetrance variant that NGS can detect, but by some unlucky combination of several lesser variants, each with a relatively small effect, and/or environmental factors.
- **The variant was present in the raw exome sequence but was not included in the final list of variants.** The necessary quality control procedures may eliminate a true variant that was poorly supported in the raw data. Small insertions or deletions cause a sequence read to align poorly to the reference and risk being discounted; the same is true of coding sequences within highly repetitive regions of the genome. Maybe that exon was poorly captured or poorly sequenced. Micro-exons of only a dozen or so base pairs are frequent in genes important in the central nervous system, and they may not be included in captured exomes (Scheckel & Darnel, 2015).
- **The variant was identified but wrongly classified as non-pathogenic.** The programs such as POLYPHEN-2 and SIFT that are used to classify mis-sense variants as pathogenic or not are individually only around 80% accurate. Using a consensus of several programs improves the accuracy but is still not perfect. For example, the very first pioneering identification of a disease gene by NGS, Miller syndrome (OMIM 263750) in 2010 by Ng *et al.*, almost failed because one of the correctly identified variants was wrongly classed as non-pathogenic. A particular problem is if the replacement amino acid is found in some other species as the normal, wild-type amino acid in the corresponding protein. That would lead the programs to label it non-pathogenic – but because of other differences between the human and corresponding animal proteins, that amino acid may only be functional in the rather different context of the animal protein and not in the human protein (see Jordan *et al.*, 2015).
- **The variant is in non-coding sequence.** Because over 98% of genomic DNA is non-coding, it might be expected that exon sequencing would miss very

many pathogenic variants. In fact few changes in non-coding DNA cause a phenotype with high penetrance; most contribute only in a minor way, if at all, to any phenotype. They are the subject of the investigations into common non-mendelian conditions described in *Chapter 13*. However, there are counter-examples.

- o One common class are changes deep inside introns that activate a cryptic splice site (see *Section 6.2*). If only a single pathogenic exonic variant can be found in a patient with a known recessive condition, it is worth sequencing the genomic DNA of the gene in question and checking all intronic variants with one of the programs that looks at splice sites. Confirmation would come from sequencing cDNA and demonstrating the presence of an abnormal splice isoform (but nonsense-mediated decay might mean that transcripts from the mutant allele would not be present in the mature mRNA). The same procedure could act as a pointer to a novel candidate gene, though that would need strong confirmation.
- o Promoters are often included in the targets of exon capture procedures, so promoter variants close to the transcription start site should not be missed. However, variants further upstream of the transcription start site may well be missed. Functional studies would be needed to show that the level of gene expression was abnormal, and that the abnormal expression had pathogenic consequences.
- o 5' or 3' untranslated regions of genes may or may not be included in exon-captured sequence. If not, conditions such as Marie Unna hypotrichosis (OMIM 146550) or myotonic dystrophy 1 (*Disease box 4*) would be missed. Even if a variant has been correctly identified, deciding whether or not it is pathogenic is much more difficult than with a variant in coding sequence.
- o Large highly methylated expanded repeats (*Disease box 4*) are not picked up by NGS. Baratela–Scott syndrome (OMIM 615777) is an interesting example – the heavily methylated DNA was completely invisible by all normal techniques; it could not be amplified or sequenced (LaCroix *et al.*, 2019).
- o A few mendelian conditions are caused by changes in non-coding RNAs – for example, cartilage-hair hypoplasia (OMIM 250250 – see Ridanpaa *et al.*, 2001) or Taybi–Linder syndrome (OMIM 210710 – see Edery *et al.*, 2011).
- o Small structural rearrangements (deletions, duplications or inversions) entirely located in non-coding sequence may affect the way regulatory sequences such as enhancers relate to their target genes. This may cause an enhancer to lose its influence over its target, or sometimes to transfer its influence from its normal target to some other nearby gene. The result is mis-expression, which may be pathogenic. The paper by Lupiáñez and colleagues (2015) gives examples.

The Deciphering Developmental Disorders (DDD) Study: a model of translational research and national collaboration for gene identification and patient benefit



The Deciphering Developmental Disorders (DDD) study (www.ddduk.org) was established in the UK in 2011 to identify the causes of developmental disorders*. Doctors in the 24 Regional Genetics Services throughout the UK and Republic of Ireland, collaborated with scientists at the Wellcome Trust Sanger Institute to collect DNA and clinical information from 13 500 undiagnosed children and adults with severe developmental disorders, and their parents. Major achievements are highlighted in bold text below.

Initially microarrays were used to detect copy number variants, followed by exome sequencing and latterly whole genome sequencing. All results reported were verified in an accredited diagnostic laboratory. At first, analysis was restricted to genes where mutations were known to be associated with developmental disorders. Although all cases had previously been examined by a clinical geneticist without the disorder being diagnosed, many mutations were found, demonstrating that:

- **the phenotypic spectrum for many disorders is much wider than previously thought.**

As the analyses progressed and as the results of similar large-scale studies were published, many new disorders were identified and delineated.

- **Up to 2019, 49 new disorders have been published from the DDD study.**
- **Diagnoses have been established in ~42% of cases.**

Before the results of large-scale studies were published using new sequencing technologies, only very vague recurrence risks could be given to families with an affected undiagnosed child where there was no other family history. DDD and other studies showed that:

- **The major cause of severe developmental disorders (where there is no consanguinity) is a *de novo* dominant mutation. It is estimated in the DDD study that 42–48% of the cohort can be explained by a *de novo* coding mutation, meaning recurrence risks are very low[#].**

Families with a child with a newly described condition really need more information and support, and collaboration between DDD, clinicians and patient organisations such as Unique have been invaluable in developing information resources (www.rarechromo.org/disorder-guides).



Image reproduced with the permission of Unique – www.rarechromo.org.

Reanalysis of data continues to yield further results and new diagnoses for families, including the identification of new genetic mechanisms (such as the contribution of **retrotransposition** to developmental disorders). International collaborations between DDD and other major large-scale projects are underway, and many new disorders are being identified.

**The DDD study is jointly funded by the Health Innovation Challenge Fund – a parallel funding partnership between Wellcome and the UK Department of Health – and the Wellcome Trust Sanger Institute, and is supported by the NHS National Institute for Health Research.*

Some cases of mosaicism have been described in a parent, and other studies of inbred populations show a higher risk of recessive disorders.

8.5. References

- Boycott KM, Hartley T, Biesecker LG, et al.** (2019) A diagnosis for all rare genetic diseases: the horizon and the next frontiers. *Cell*, **177**: 32–37.
- Charpentier E and Doudna JA** (2013) Rewriting a genome. *Nature*, **495**: 50–51.
- Edery P, Marcaillou C, Sahbatou M, et al.** (2011) Association of TALS developmental disorder with defect in minor splicing component *U4atac* snRNA. *Science*, **332**: 240–243.
- Hoffman EP, Brown EP and Kunkel LM** (1987) Dystrophin: the protein product of the Duchenne muscular dystrophy locus. *Cell*, **51**: 919–928.
- Hoischen A, van Bon BWM, Gilissen C, et al.** (2010) *De novo* mutations of *SETBP1* cause Schinzel–Giedion syndrome. *Nature Genetics*, **42**: 482–485.
- Jordan DM, Frangakis SG, Golzio C, et al.** (2015) Identification of *cis*-suppression of human disease mutations by comparative genomics. *Nature*, **524**: 225–229.
- Kurotaki N, Imaizumi K, Harada N, et al.** (2002) Haploinsufficiency of *NSD1* causes Sotos syndrome. *Nature Genetics*, **30**: 365–366.
- LaCroix AJ, Stabley D, Sahraoui R, et al.** (2019) GGC expansion and exon 1 methylation of *XYLT1* is a common pathogenic variant in Baratela–Scott syndrome. *Am. J. Hum. Genet.* **104**: 35–44.
- Lupiáñez DG, Kraft K, Heinrich V, et al.** (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**: 1012–1025.
- Miyamura Y, Suzuki T, Kono M, et al.** (2003) Mutations of the RNA-specific adenosine deaminase (DSRAD) gene are involved in dyschromatosis symmetrica hereditaria. *Am. J. Hum. Genet.* **73**: 693–639.
- Ng SB, Buckingham KJ, Bingham AW, et al.** (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, **42**: 30–35.
- Pingault V, Bondurand N, Kuhlbrodt K, et al.** (1998) *SOX10* mutations in patients with Waardenburg–Hirschsprung disease. *Nature Genetics*, **18**: 171–173.

Ray PN, Belfall B, Duff C, et al. (1985) Cloning of the breakpoint of an X:21 translocation associated with Duchenne muscular dystrophy. *Nature*, **318**: 672–675.

Ridanpaa M, van Eenennaam H, Pelin K, et al. (2001) Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell*, **104**: 195–203.

Scheckel C and Darnel RB (2015) Microexons – tiny but mighty. *EMBO J.* **34**: 273–274.

Shendure J, Findlay GM and Snyder MW (2019) Genomic medicine – progress, pitfalls, and promise. *Cell*, **177**: 45–57.

Woo SLC, Lidsky AS, Guttler F, et al. (1983) Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. *Nature*, **306**: 151–155.

General background

Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd edn. Johns Hopkins Press. A highly authoritative exposition of the basis of human genetic mapping.

Strachan T and Read AP (2019) *Human Molecular Genetics*, 5th edn. CRC Press. Covers the material of this chapter in greater depth.

Useful websites

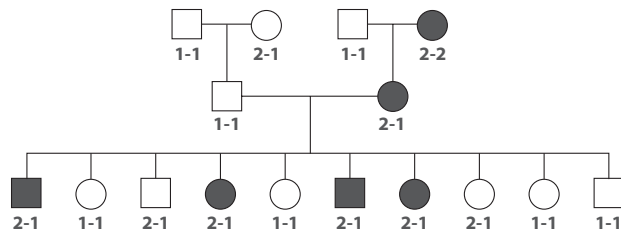
dbSNP: www.ncbi.nlm.nih.gov/snp/

Hereditary Hearing Loss Homepage: <https://hereditaryhearingloss.org>

Zebrafish Information Network: <http://zfin.org>

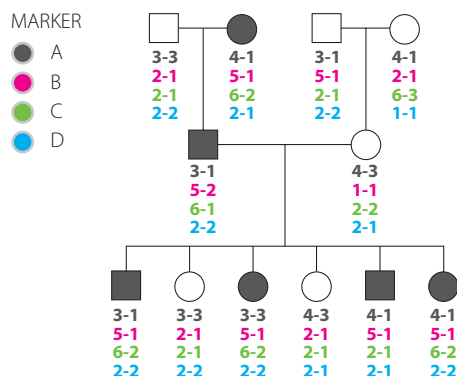
8.6. Self-assessment questions

- (1) [This question relates to material in the “Resources” tab on www.scionpublishing.com/NCG4]. The pedigree shows a family with a fully penetrant autosomal dominant disease. Genotypes are shown for a DNA marker that has alleles 1 and 2.

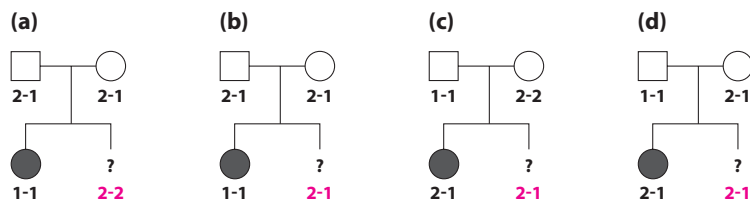


- Mark each meiosis as definitely non-recombinant, recombinant, or uninformative.
- What is the best estimate of the recombination rate?
- Test the significance of the deviation from the null hypothesis of 50% recombinants (no linkage) using χ^2 . Is the result significant?

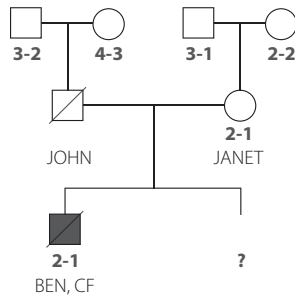
- (d) In what proportion of such pedigrees would you expect to see the observed pattern of markers in generation III if there were no linkage between the marker and the disease? Call this L1.
- (e) In what proportion of such pedigrees would you expect to see the observed pattern of markers in generation III if there were linkage (recombination fraction θ) between the marker and the disease? Call this L2 (L2 is a function of θ , of course).
- (f) Tabulate the values of L1, L2, L2/L1 and $\log_{10}(L2/L1)$ for $\theta = 0, 0.05, 0.1, 0.15, 0.2 \dots 0.5$.
- (g) $\log_{10}(L2/L1)$ is the lod score. What is the maximum lod score? Is this significant? Comment.
- (2) In the family below, a fully penetrant autosomal dominant disease is segregating. The disease has been mapped to chromosome 2q35 in previous studies. The pedigree shows genotypes for four DNA polymorphisms from within the candidate region, arranged in chromosomal order. Can you narrow down the disease locus using these data?



- (3) Four couples each have a child with the same severe autosomal recessive disease. They each request pre-implantation diagnosis for their next child. After *in vitro* fertilization, a single cell is removed. Rather than attempt to develop and validate specific tests for the 8 variants involved across the four families, the lab types them for a genetic marker closely linked to the disease locus. Genotypes are shown. For each family, report the prediction.



- (4) Janet was pregnant with their second child when her husband John and 14 year old son Ben were killed in a car crash. Ben had cystic fibrosis. Janet desperately wants to have the baby, but feels she cannot cope with a second tragedy in the form of another child with CF. It was only after the double funeral that she raised the issue with her physician. He pointed out that prenatal diagnosis required DNA samples



from Ben and John, who had been cremated. John's elderly parents were contacted and they agreed to give mouthwash samples for mutation testing. A week later the laboratory requested blood samples, because the mouthwash DNA was too poor quality for sequencing to identify the CF mutation that one of them must have passed to John. However, by this time they were uncontactable, having embarked on a world cruise to forget their grief. A search through old files hit on a Guthrie blood spot card that had been used for Ben's routine neonatal screening. The sample was too degraded to be used for sequencing, but the laboratory was able to type it and the mouthwashes from John's parents for a microsatellite polymorphism known to be tightly linked to the CF locus. Janet and her parents were typed for this marker, with the results shown. What are the possible genotypes and implications for the fetus if a prenatal test is carried out?

- (5) Linkage studies have mapped the gene causing a mendelian disease to a 2 Mb region. Database searching shows this region contains genes encoding the following products:
- an enzyme involved in detoxification reactions in the liver
 - a phosphate transporter
 - a component of the large ribosomal subunit
 - a nuclear-encoded component of the mitochondrial electron transport chain
 - a ubiquitously expressed protein of unknown function containing a variable length run of glutamines
 - an enzyme catalyzing one step in the tricarboxylic acid cycle
 - a protein of unknown function expressed in the adult central nervous system and retina
 - a transcription factor expressed only in restricted regions of early embryos
 - a ubiquitously expressed transcription factor
 - a component of the DNA damage repair pathway

What would be your first candidate for mutation testing if the disease in question was:

- (a) an adult-onset neurodegenerative disease showing anticipation
- (b) a rickets-like skeletal malformation
- (c) absence of the pituitary gland
- (d) an accelerated aging syndrome
- (e) a combination of deafness and diabetes

[Hints on questions 1, 2 and 5 are provided in the *Guidance* section at the back of the book.]

09

Why are some conditions common and others rare?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Define allele frequency, inbreeding, founder effect, coefficient of relationship
- Use the Hardy–Weinberg formula to calculate carrier frequencies for autosomal and X-linked recessive conditions
- Describe qualitatively the effects of inbreeding and use Sewall Wright’s path coefficient method to calculate gene sharing by relatives
- Describe the consequences of heterozygote advantage and give examples
- Explain why it is difficult to change population allele frequencies by medical interventions

9.1. Case studies

CASE 19 ULMER FAMILY

- Hannah, 6-month-old baby girl, Ashkenazi Jewish background
- Normal at birth but then increasing problems
- ? Tay–Sachs disease
- Enzyme test confirms diagnosis

231

239

316

395

Rachel and Uzi Ulmer’s families originally came from Eastern Europe and are of Ashkenazi Jewish origin, though neither family was religious, and neither Rachel nor Uzi attended Jewish schools or the synagogue. Hannah was their third child; they had two healthy children, aged 5 and 3 when Hannah was born. Hannah seemed very well as a young baby and smiled and held her head up at the normal time. However, at her 6 month check, the doctor noted her head control was not good and arranged a further appointment 4 weeks later. He was more concerned then because her head control was even worse and she seemed less responsive. He undertook a full neurological examination and found that she had cherry red spots in the maculae of both eyes. He felt fairly sure that Hannah had Tay–Sachs disease (OMIM 272800) and arranged for blood samples to be sent to measure levels of hexosaminidase A. Sadly the levels were extremely low and the diagnosis was confirmed. Rachel and Uzi and their families were devastated at the poor prognosis they were given for Hannah, but resolved to try and make her life as comfortable as possible. Over the next 3 years Hannah further deteriorated, her limbs became spastic and she lost her vision and hearing. She died at the age of four and a half years.

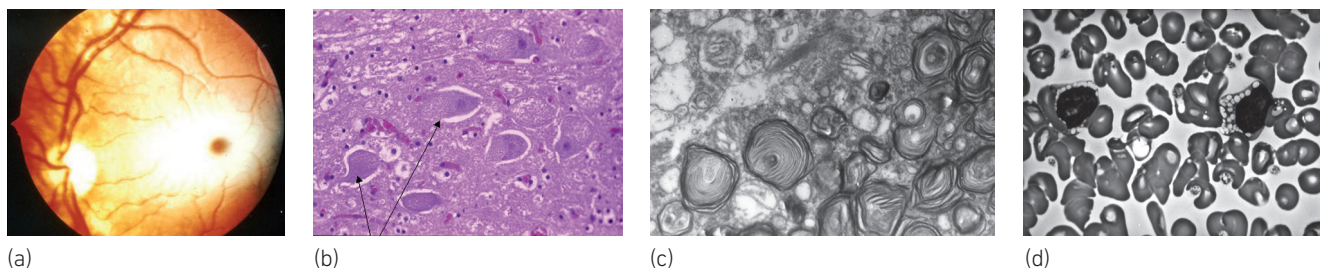


Figure 9.1 – Typical features of lysosomal storage disorders.

(a) The characteristic cherry red spot on the retina of a child with Tay–Sachs disease. (b) Ballooned neurons in the central nervous system (arrows). (c) Abnormal cell bodies seen under the electron microscope. (d) Vacuolated lymphocytes. Photos courtesy of Drs Ed Wraith and Guy Besley, Royal Manchester Children's Hospital.

9.2. Science toolkit

This chapter is about **allele frequencies**: what determines them, how they can change, and the way allele frequencies can be used to work out genetic risks in a variety of situations. Allele frequencies are often called gene frequencies – a common bad habit that we will try to avoid, because most autosomal genes have a fixed frequency of 1 copy per genome. Allele frequencies are numbers that lie between 0 and 1; traditionally they are symbolized as p and/or q .

Allele frequencies depend on the concept of a **gene pool**. This would consist of all the alleles at a particular locus in some defined population. The population might be anything from a small community to the whole of humanity. Within the gene pool the frequency of allele \underline{A} is the proportion of all the alleles at that locus that are \underline{A} . It is also the probability that an allele, picked at random from the pool, would be \underline{A} . Alleles can be defined in various ways for different purposes. Most discussion of allele frequencies in relation to disease classifies all alleles at a locus into just two categories, normal alleles and disease alleles. If you sequenced the DNA, you would probably find that each included quite a range of variants, and for some purposes you might wish to count each sequence variant as a separate allele. Whatever scheme you use to classify alleles, the frequencies of all the alleles in a population must add up to 1.

Clinicians are usually more interested in frequencies of genotypes than frequencies of alleles. They want to know the probability that a randomly selected person is homozygous or heterozygous for a disease allele. The Hardy–Weinberg distribution (*Box 9.1*) describes the relationship of genotype frequencies to allele frequencies. The Hardy–Weinberg distribution makes it possible to calculate the frequency of carriers of recessive conditions in a population without having to sample a large number of people and genotype them in the laboratory. Along with the basic mendelian rules, it is the central tool for assessing genetic risks.

The bean-bag model that we use in *Box 9.1* (below) to derive the Hardy–Weinberg distribution illustrates an important precondition of the distribution. The probability that the second bean will be \underline{A} or \underline{a} must be entirely independent of whether the first bean was \underline{A} or \underline{a} . In genetic terms, the genotype distribution will be Hardy–Weinberg only if there is **random mating**. Random mating does not mean free love; it just means that

The Hardy–Weinberg distribution

Suppose there are two alleles, \underline{A} and \underline{a} in a population (there may or may not be others as well; it doesn't matter). Say the frequency of allele \underline{A} is p and the frequency of \underline{a} is q . p and q will sum to 1 if \underline{A} and \underline{a} are the only alleles in the population; if there are other alleles $p+q$ will be less than 1. The Hardy–Weinberg distribution predicts that the frequencies of the three possible genotypes are:

$$\frac{AA}{p^2} \qquad \frac{Aa}{2pq} \qquad \frac{aa}{q^2}$$

Students often state this as $p^2 + 2pq + q^2 = 1$. This is wrong. The Hardy–Weinberg distribution is not an equation. It describes the relation between allele frequencies and genotype frequencies. p^2 , $2pq$ and q^2 only add up to 1 if those two alleles are the only alleles in the population, so that everybody must be either \underline{AA} or \underline{Aa} or \underline{aa} . If there are three alleles in the population (say, \underline{A} , \underline{a} and \underline{a}_1 , with frequencies p , q and r) the frequencies of \underline{AA} , \underline{Aa} and \underline{aa} are still p^2 , $2pq$ and q^2 , but there are also other genotypes in the population. The total distribution is then:

$$\frac{AA}{p^2} \qquad \frac{Aa}{2pq} \qquad \frac{aa}{q^2} \qquad \frac{Aa_1}{2pr} \qquad \frac{aa_1}{2qr} \qquad \frac{a_1a_1}{r^2}$$

The basis of the Hardy–Weinberg distribution can easily be seen by doing a little thought experiment. Imagine all the genes in a gene pool as a large collection of beans in a bag. A proportion p of the beans are of type \underline{A} and a proportion q are of type \underline{a} . Eyes shut, reach into the bag and pick a bean. The chance it is \underline{A} is p , and the chance it is \underline{a} is q . Replace the bean, shake up the bag and pick a second bean. Again the chance it is \underline{A} is p , the chance it is \underline{a} is q . The chance both beans were \underline{A} is p^2 ; the chance both were \underline{a} is q^2 . The chance the first bean was \underline{A} and the second \underline{a} is pq ; similarly, the chance you picked first \underline{a} then \underline{A} is qp . Overall, the chance you picked one \underline{A} and one \underline{a} is $2pq$.

you don't ask your beloved's genotype before you decide whether to pop the question. This is not as far-fetched as it sounds. There are many ways in which people tend to select mates based partly on genotype. There is **assortative mating** for ethnicity, height, intelligence, deafness and many other phenotypes that are at least partly genetic. The effect of assortative mating is to increase the proportion of homozygotes for each allele in the population, and decrease the proportion of heterozygotes, compared to the predictions of Hardy–Weinberg.

Assortative mating is mainly important in clinical genetics in the context of inbreeding. Relatives share genes, and so if you marry a relative rather than an unrelated person you increase the chance that your partner will have alleles in common with you. If you carry an allele for a recessive condition there is an increased chance that your partner will also carry it, and thus an increased risk of having a child with the condition, compared to an outbred couple. This is discussed further in *Section 9.3* in relation to the **Choudhary family (Case 18)**.

Using Hardy–Weinberg to calculate carrier risks

Suppose the sister of a boy with cystic fibrosis is getting married. She knows the disease is genetic, and she wants to know the risk that she might have an affected child. She is healthy, but of course it is quite likely that she is a carrier of cystic fibrosis (see below for the exact probability). This has no implications for her own health, but her children would be

at risk if her partner happened also to be a carrier. Assuming random mating, calculating the risk that her partner is a carrier is the same as calculating the probability that a person, picked at random from the population, is a carrier. Hardy–Weinberg enables us to do this.

If we ignore the extensive allelic heterogeneity and classify all alleles at the *CFTR* locus as \underline{A} (functional) and \underline{a} (non-functional), with frequencies p and q , the proportions of the genotypes are:

$$\begin{array}{ccc} \frac{AA}{p^2} & \frac{Aa}{2pq} & \frac{aa}{q^2} \end{array}$$

Cystic fibrosis affects one birth in 2000 in Americans of Northern European origin (it is less frequent among Hispanics, African–Americans and Asian Americans). Therefore q^2 is 1/2000. Hence q is the square root of 1/2000, which is nearly 1/45. One allele in 45 at the *CFTR* locus is \underline{a} ; the remaining 44/45 must be \underline{A} (we had decided to classify *all* alleles as \underline{A} or \underline{a} , so in this case $p+q=1$). We can now calculate the carrier frequency: $2pq = 2 \times 44/45 \times 1/45 = 1/23$ approximately. Assuming random mating, if the girl's partner is of white Northern European descent there is a 1 in 23 chance he is a carrier.

For an X-linked condition the calculation is even simpler. If hemophilia A affects one boy in 5000 in a certain population, what proportion of the women are carriers? Males can only be \underline{A} or \underline{a} , so the frequency of the disease allele is simply equal to the proportion of males who are affected:

Females			Males	
AA	Aa	aa	A	a
p^2	$2pq$	q^2	p	q

The carrier frequency among women is $2 \times 4999/5000 \times 1/5000$, or 1 in 2500.

For a rare condition the frequency of the normal allele can usually be approximated to 1, which makes the calculation easier. However, it is important to note that for rare autosomal recessive conditions, a significant proportion of cases are the offspring of consanguineous marriages. This is not a major factor with cystic fibrosis in the USA or UK, because cystic fibrosis is a relatively common disease. But the rarer a condition is, the higher is the proportion of cases where consanguinity is a factor. Ignoring this and simply assuming a Hardy–Weinberg distribution can lead to serious over-estimation of carrier frequencies for very rare diseases (see below).

Changing allele frequencies

Allele frequencies can change from generation to generation for a number of reasons:

- New mutations increase the number of disease alleles in the gene pool. There can also be back mutation, changing a disease allele into a normal allele. However, for loss of function mutations the process is largely unidirectional. Any one of a large number of possible sequence changes can convert a functioning gene into a non-functioning one, but only a very specific back mutation can restore the function of a non-functional allele. The ratio of forward to back mutation rates is likely to be of the order of 1000:1. Other mutations may well occur to non-functional alleles, but the allele remains non-functional.

- Natural selection will remove disease alleles from the gene pool if the disease is such that affected people are less likely to reproduce. Artificial selection might have a similar effect (see *Section 9.4*).
- A large influx of migrants from a population with substantially different allele frequencies could change the gene pool.
- Every population is made up of a finite number of individuals. In each generation, those who actually reproduce comprise only part of the total population. Reproducers are never exactly representative of their whole generation for purely statistical reasons, and this introduces chance fluctuations of allele frequencies between generations. These changes will accumulate over time, because the gametes that produce the next generation of children are a sample of the alleles in the parental generation only, and have no statistical memory of the allele frequencies in earlier generations. The smaller the size of the reproducing population, the greater the random fluctuations (termed **genetic drift**) that will take place at every generation, and the more rapidly drift will be observed over generations.

Whatever the initial genotype frequencies in a population, one generation of random mating is enough to establish a Hardy–Weinberg distribution. If, when a population is surveyed, the distribution is seen to differ significantly from Hardy–Weinberg, this might mean one of several things:

- *The survey methodology is flawed.* This is most often seen when a new DNA polymorphism is being checked. A non-Hardy–Weinberg distribution of genotypes most likely means that the genotyping method is producing errors and needs to be improved to make it reliable.
- *Selection has already occurred.* A genotype that causes substantial prenatal or infant mortality will be under-represented in a sample of living people.
- *There is significant assortative mating.* This might be simple inbreeding, or perhaps the population is not freely interbreeding, but consists of two or more groups that have different allele frequencies and that largely breed among themselves (**population stratification**).

Note that if you are using a chi-squared test to check whether genotype numbers fit Hardy–Weinberg, with two alleles there is only one degree of freedom, even though you have three observed and expected numbers. That is because once you have fixed q , everything else in the expected numbers follows.

Factors determining allele frequencies

People sometimes imagine that dominant characters should be common and recessive ones rare. Actually there is no connection between the frequency of an allele and whether it causes a dominant or recessive phenotype. Huntington disease is one example among many of a rare dominant; blood group O is a common recessive.

For a neutral character where selection plays no role, the allele frequencies reflect the frequencies in the founders of that population, modulated by genetic drift if, at any time, the breeding population was reduced to small numbers. Most common neutral DNA polymorphisms (SNPs or microsatellites, see *Box 8.1*) fit this category. Occasional mutations

will introduce some random variation by changing one allele into another. As explained in *Chapter 11*, deamination of methylated cytosines has tended, over evolutionary time, to convert CpG sequences systematically into TpG, but this effect on allele frequencies is only noticeable on an extremely large timescale.

For diseases, the allele frequencies reflect additional factors. Recurrent mutations tend to increase the pool of loss of function alleles, while selection acts to remove disease alleles from the population. For dominant and X-linked diseases selection is very effective, because everybody (dominant) or all males (X-linked) who carry the disease allele have the disease and so are exposed to selection. For autosomal recessive conditions selection acts much more slowly, because most of the disease alleles are in healthy heterozygotes, who are not subject to selection. For this reason, alleles for recessive diseases can persist for very many generations in a population, even if the disease is very severe. This leads to important **founder effects** in many populations. If a population, however numerous today, derives from a small number of founders, or passed through a bottleneck when only a few individuals contributed to the next generation, then any recessive allele that was present in one of the founders is likely to be present at high frequency in the modern population. Equally, if the allele responsible for a normally common recessive condition happened to be absent from the small pool of founders, it will be rare or absent in the modern population (*Figure 9.2*).

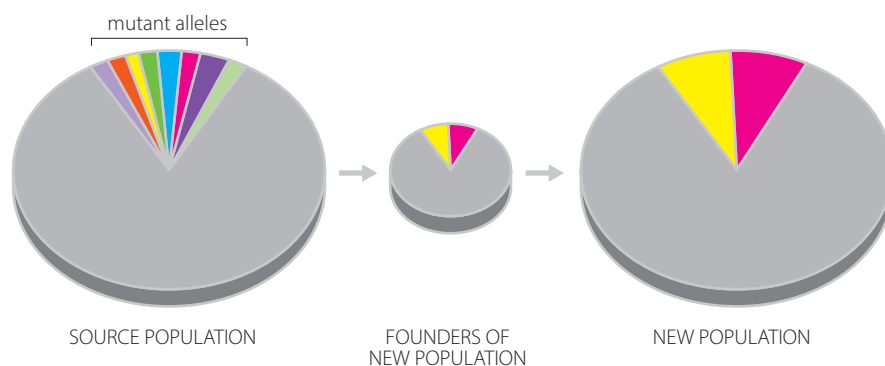


Figure 9.2 – Founder effects.

Founder effects are seen in many populations, mostly ones that are small and relatively isolated reproductively. Finns (*Table 9.1*) may seem to be an exception, but in fact the population of Finland is known to have passed historically through severe bottlenecks – see *Disease box 9*. For the reasons explained above, founder effects are seen mainly with recessive conditions but not usually with dominant or X-linked conditions. *Table 9.1* shows some examples, and *Disease box 9* describes in more detail some diseases characteristic of one such population, Ashkenazi Jews.

Table 9.1 – Diseases common in certain populations, probably because of a founder effect

Disease	OMIM	Mode of inheritance	Population	Comments
Diastrophic dysplasia	222600	AR	Finns	90% of Finnish cases have a splice donor mutation in intron 1
Aspartylglucosaminuria	208400	AR	Finns	Carrier frequency 1 : 30; 98% have p.Cys163Ser mutation
Neuronal ceroid lipofuscinosis	256730	AR	Finns	Single mutation, p.R122W, found in 97% of Finnish cases
Hermansky–Pudlak syndrome	203300	AR	Puerto Ricans	Carrier frequency 1 : 21; much rarer in most other populations
Bardet–Biedl syndrome	209900	AR	Bedouin	Two non-allelic forms, BBS2 and BBS3 both relatively frequent
Myotonic dystrophy	160900	AD	Quebec – Sanguenay	Prevalence 1 : 500, 30–60× prevalence in most other populations
Butyrylcholinesterase deficiency	177400	AR	Alaskan Eskimos	Deficiency allele frequency 0.1; three different alleles reported in this population
Usher syndrome type 1C	276904	AR	French–Acadians in Louisiana	43/44 cases homozygous for a c.216G>A mutation.
Charcot–Marie–Tooth disease type 4D	601455	AR	Bulgarian gypsies	The OMIM entry gives an interesting commentary

Note that all diseases are autosomal recessive (AR) except for myotonic dystrophy.

Heterozygote advantage

There is a second reason why a recessive condition may be common in a population. The classic example is sickle cell disease. This is common in many populations where falciparum malaria is or was recently endemic, but is absent from populations where malaria was not frequent. The driving force is that heterozygotes are relatively resistant to malaria. Thus in historical times normal homozygotes often died of malaria, and sickle cell homozygotes died of their disease, leaving the heterozygotes to contribute disproportionately to the next generation (*Figure 9.3*).

Even a very small degree of heterozygote advantage, continued over many generations, can greatly influence allele frequencies. If, relative to Aa heterozygotes, a proportion s_1 of aa and a proportion s_2 of AA homozygotes fail to reproduce, at equilibrium the ratio of allele frequencies, q/p , is s_2/s_1 . For cystic fibrosis s_1 is effectively 1 (we are talking about conditions during the past, not today), while q/p is 1/45 for people of Northern European origin. It follows that s_2 must also be about 1/45. In other words, in order to account for the high proportion of people of Northern European origin who are carriers of cystic fibrosis, carriers must have enjoyed a 2% reproductive advantage (average number of children surviving to reproductive age) over normal homozygotes during past centuries. Such an advantage is too small to be easily detectable in survey data, and its nature is controversial.

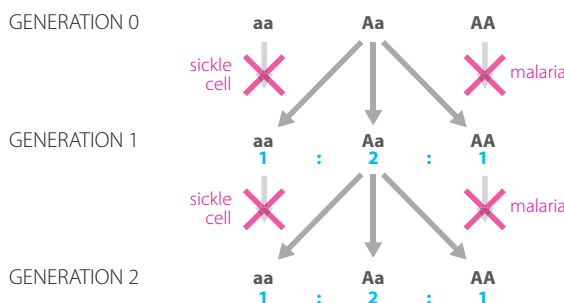


Figure 9.3 – Heterozygote advantage.

A hypothetical extreme example in which everybody with sickle cell disease (aa) fails to reproduce because of their disease, while all normal homozygotes (AA) fail to reproduce because of malaria. Only heterozygote \times heterozygote matings contribute to the next generation. Regardless how many generations the selection continues, at birth every generation always consists of 25% sickle cell homozygotes, 50% heterozygotes and 25% normal homozygotes. No real situation is this extreme.

Heterozygote advantage or founder effect?

If a disease is common in a certain population because of a founder effect, we would expect most affected people to have exactly the same ancestral variant. On the other hand, if heterozygote advantage has been the mechanism, a whole range of variants might be common. A good example is β -thalassemia in Jews from Kurdistan. The carrier frequency is as high as 20%, but 13 different β -globin mutations have been described in this small group. Clearly, heterozygote advantage (presumably resistance to malaria) has been the driving force here. On the other hand, among the Amish of Lancaster County, Pennsylvania, the rare recessive Ellis–van Creveld syndrome (OMIM 225500) is relatively common, with a disease allele frequency of 0.07. All affected people in nine families studied had the same sequence changes in each copy of the *EVC* gene on chromosome 4. All could trace their ancestry back to one couple, Samuel King and his wife, who immigrated in 1744 – one of whom must have been a carrier, though perfectly healthy. This is an example of a founder effect. In this particular case the founder is relatively recent, so the shared ancestry is demonstrable in the genealogy. In most cases it has to be inferred by identifying a shared pathogenic variant and, ideally, a shared haplotype of non-pathogenic genetic markers on the chromosomal segment carrying the disease allele. As it happens, the *EVC* variant in the Amish families is always accompanied by a second, non-pathogenic sequence change in the *EVC* gene. The change is not found on non-mutant chromosomes. This provides additional confirmation that we are looking at copies of a single ancestral chromosome.

There are several examples of isolated populations where more than one variant causing a certain recessive disease is relatively common. Heterozygote advantage is a likely explanation, but there may also be an observational bias. Mutations are normally identified by studying affected people, not by screening a whole population. If one disease allele is fairly common, there will be plenty of people heterozygous for that allele. Whenever such a person happens to carry another disease allele they will be affected, and so both their disease alleles will get identified. Thus a common recessive disease allele provides a mechanism for bringing the rarer alleles in a population to medical attention.

But as some of the Jewish diseases in *Disease box 9* illustrate, the effects of heterozygote advantage, founder effects and genetic drift can be hard to disentangle.

9.3. Investigations of patients

CASE 19 ULMER FAMILY

- Hannah, 6-month-old baby girl, Ashkenazi Jewish background
- Normal at birth but then increasing problems
- ? Tay–Sachs disease
- Enzyme test confirms diagnosis
- Test the sibs?

231

239

316

395

Tay–Sachs disease is one of the lysosomal storage diseases. A specific lysosomal degradative enzyme, in this case hexosaminidase A, is defective. Its high molecular weight substrate, G_{M2} ganglioside, continues to be imported into the lysosome but cannot be degraded and so accumulates. Lysosomes are not just cellular trash cans, they are part of a signaling network that sets the metabolic state of the cell, and the distended lysosomes, crammed with undegraded G_{M2} ganglioside gradually kill retinal ganglia and other neurons, leading to death of the child, usually at age 2–4 years. The disease is known from all ethnic groups, but is about 100 times more frequent in Ashkenazi Jews than in most other groups. About 1 in 30 North American Jews is a carrier, compared to perhaps 1 in 300 people in most other populations. This has led to carrier screening programs being set up specifically in Jewish communities in several countries, as described in *Chapter 11*. Because their families led entirely secular lives and were not part of any Jewish community, Rachel and Uzi had missed out on screening.

Having learned about Tay–Sachs disease, Rachel and Uzi were very concerned to know whether either of their healthy children was a carrier and pressed for them to be tested. The geneticist contended that it was unethical to test young children. Any test result would have no implications for the child's health or management. The time to offer a test was when the child was of an age to make an independent informed choice. They might prefer not to know, and it was wrong to close off that possibility for them when there was no compensating benefit in testing a child. The geneticist did point out that each child had a 2 in 3 risk of being a carrier (not 1 in 2 – see *Box 9.2*).

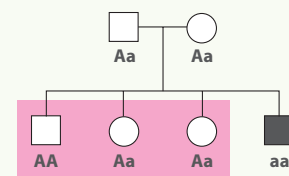
Rachel and Uzi thought they might have more children, and if so they would want prenatal diagnosis. To prepare for this possibility, mouthwash samples were taken from both of them and the *HEXA* gene checked. Three *HEXA* variants together make up over

The risk a healthy sib is a carrier

If two carriers of an autosomal recessive disease have a child, the chance is 1 in 4 that it will be affected, 1 in 2 that it will be a carrier, and 1 in 4 that it will be homozygous normal. However, the chance that a healthy sib of an affected child is a carrier is not 1 in 2, but 2 in 3. As the figure shows, we know the child is not affected, so it must be one of the three in the shaded box.

This is an interesting example of how hard it can be to spot an error when the error consists of giving the right answer to the wrong question.

- The right question is, what is the risk that a healthy sib is a carrier? Answer 2 in 3.
- The wrong question is, what is the risk that *any* child of a carrier couple will be a carrier? Answer 1 in 2 – but this is not the question the parents asked.



BOX 9.2

90% of all Tay–Sachs alleles in Ashkenazi Jews (*Table 9.2*). Rachel and Uzi each carried one of those variants. Rachel had a change affecting the intron 12 splice site and Uzi had a 4 bp frameshifting insertion in exon 11. Hannah was a compound heterozygote for these two loss of function variants.

Table 9.2 shows the results of a survey that looked just for the three major Ashkenazi variants in Tay–Sachs carriers. In non-Jewish populations these do not make up a particularly high proportion of all disease alleles, and other work has demonstrated the expected wide assortment of individually rare pathogenic alleles, typical of a loss of function condition. The situation in Ashkenazi Jews is interesting. As explained above, if the high frequency of Tay–Sachs disease were just a founder effect we would not expect more than one variant to be common. This is discussed further in *Disease box 9* at the end of this chapter.

Table 9.2 – Distribution of pathogenic variants in the hexosaminidase A gene

Variant	Percentage in Ashkenazi carriers (n = 156)	Percentage in non-Jewish carriers (n = 51)
4 bp insertion in exon 11	73	16
Intron 12 splice site mutation	15	0
p.Gly269Ser in exon 7	4	3
Unidentified	8	81

Data of Paw *et al.* (1990) cited in OMIM entry 272800.

CASE 18 CHOUDHARY FAMILY

- Baby girl Nasreen, healthy but deaf
- Multiply consanguineous family
- Autozygosity mapping
- Exome sequencing
- A second recessive condition?
- Calculate coefficient of inbreeding
- Possibilities for therapy

207

216

240

395

This family, with its multiple consanguineous marriages, is highly inbred. These complicated family structures are particularly seen in the Middle East and the Indian subcontinent. From a genetic, though arguably not a social, viewpoint, such inbreeding is undesirable – but people do tend to overestimate the degree to which inbreeding is deleterious. People from backward rural communities are often claimed to be mentally slow because of inbreeding – in reality, any apparent slowness probably has quite different causes. Researchers would like to estimate the true genetic costs of inbreeding by correlating the frequency of problems with the degree of inbreeding. This could be used as one factor – but only one – to be considered in individual counseling and general social policy. For this we need to calculate the coefficients of relationship and of inbreeding:

- The **coefficient of relationship of two people** is the proportion of their alleles that they share by virtue of having one or more definable common ancestors.
- The **coefficient of inbreeding of a person** is the proportion of loci at which the individual is expected to be homozygous because of the consanguinity of the parents. It is half the coefficient of relationship of the parents. Equally, it is the probability that at any given locus the individual receives two alleles that are identical by descent.

The easiest method for doing the calculation is Sewall Wright's path coefficient method (Box 9.3). Figure 9.4 shows how Nasreen's coefficient of inbreeding could be calculated. Using the path coefficient method, the coefficient of relationship of Aadnan and his wife Mumtaz is calculated as $10/64$ – a bit closer than the $1/8$ of first cousins. The coefficient of inbreeding of Nasreen is therefore $5/64$. Note that even quite highly inbred pedigrees produce relatively modest coefficients of inbreeding. Without resorting to incest, it is difficult to invent a pedigree where a child has a coefficient of inbreeding as high as $1/4$. Mouse geneticists use repeated brother–sister matings over many generations to produce highly homozygous lines. Only the pharaohs of ancient Egypt are recorded as trying the same thing in humans.

Inbreeding is recognizable in a person's genome sequence by the presence of stretches where genetic markers are homozygous. The total length of such runs of homozygosity gives a measure of the degree of inbreeding, independent of any checking of the pedigree.

Calculating the effects of inbreeding

What proportion of their genes do relatives share?

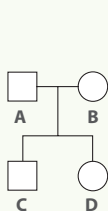
For close relatives the answer may be intuitively obvious:

- parent and child share half their genes (always)
- full sibs (same two parents) share half their genes (on average)
- half-sibs (one parent in common) share one-quarter of their genes (on average)
- uncles / aunts and nephews / nieces share one-quarter of their genes (on average)
- first cousins share one-eighth of their genes (on average)

If these figures are not intuitively obvious, or where the relationship is more complicated, Sewall Wright's path coefficient method is easy to follow and reliable:

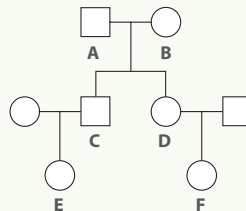
- (1) draw the pedigree showing only the common ancestor(s) and the links to them
- (2) choose one path between the two relatives, through a common ancestor and count the number of links
- (3) if that path has n links it contributes $(1/2)^n$ to the coefficient of relationship
- (4) if there is more than one possible path, do the same for each path
- (5) add together the contribution of each path

This is illustrated below for the simple cases of full sibs and first cousins, to show that it does indeed produce the right answer. A more complicated example is shown in Figure 9.4.



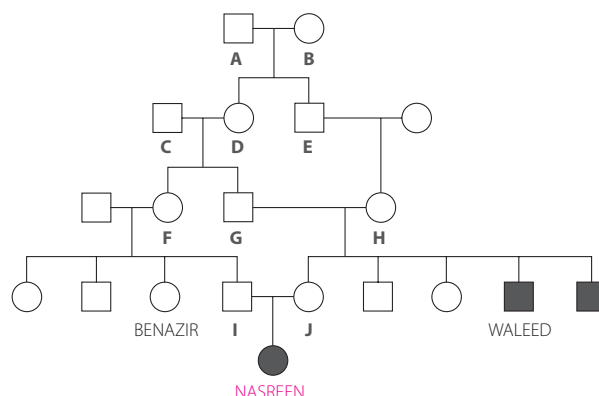
Full sibs:

Pathway	Contribution
C–A–D	$(1/2)^2 = 1/4$
C–B–D	$(1/2)^2 = 1/4$
Total:	$1/2$



First cousins:

Pathway	Contribution
E–C–A–D–F	$(1/2)^4 = 1/16$
E–C–B–D–F	$(1/2)^4 = 1/16$
	$1/8$



COEFFICIENT OF RELATIONSHIP OF I AND J:

PATH	STEPS	CONTRIBUTION
I - F - C - G - J	4	$(1/2)^4 = 1/16$
I - F - D - G - J	4	$(1/2)^4 = 1/16$
I - F - D - A - E - H - J	6	$(1/2)^6 = 1/64$
I - F - D - B - E - H - J	6	$(1/2)^6 = 1/64$

coefficient of relationship: **10/64**

coefficient of inbreeding of Nasreen

= $1/2$ coefficient of relationship of her parents = **5/64**

Figure 9.4 – Calculation of the coefficient of inbreeding of Nasreen Choudhary (Case 18).

Because there is more than one inbreeding loop, it is not easy to do this calculation intuitively. The path coefficient method is used to calculate the coefficient of relationship of Nasreen's parents.

9.4. Going deeper...

The frequencies of genetic diseases in a population depend on the opposing effects of mutation and selection, working on whatever frequencies were present in the initial founder population.

For dominant and X-linked diseases mutations normally have a short half-life.

- In the extreme case of a lethal dominant condition, every case must be a new mutation. The parents would not be parents if they were themselves affected. Thanatophoric dysplasia (see *Box 6.2*) is an example. Note that genetic lethality means inability to pass on one's genes, and not necessarily physical lethality.
- A mutation will usually persist for only one or a few generations if it causes a dominant condition that reduces a person's chance of passing on their genes, even if it does not wholly prevent reproduction. An unusual appearance or a mild disability can be quite enough to reduce a person's chances of success in the marriage market. Neurofibromatosis I (*Disease box 1*) and achondroplasia (**James Jenkins, Case 14**) illustrate this: a high proportion of cases of each condition have *de novo* mutations.
- For an X-linked recessive condition, selection is effective only against affected males. If a population has roughly equal numbers of XX females and XY males, one-third of all X chromosomes are in males (*Figure 9.5*). For a genetically lethal

condition such as Duchenne muscular dystrophy, this means that one-third of all DMD mutant alleles are wiped out each generation. Because DMD is still with us, this loss must be roughly balanced by fresh mutations. If the rates of mutation and selection balance out, one-third of cases of DMD would be fresh mutations. This has major implications for risk estimation. We cannot assume that the mother of a DMD boy is a carrier – the chance is only 2 in 3. Only if she also has an affected brother or other maternal relative is she an *obligate* carrier (as with **Judith Davies in Case 4**).

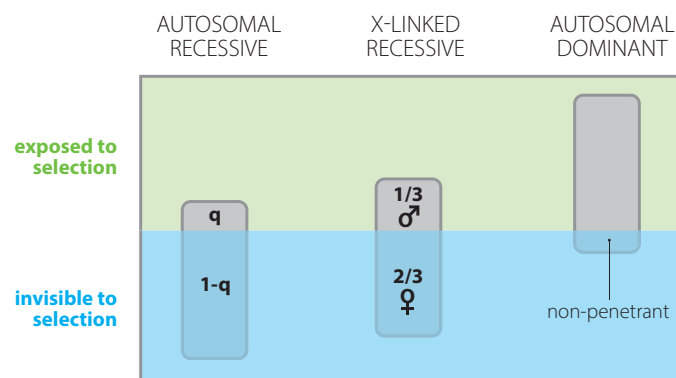
For recessive characters, selection works on mutant alleles only when they are in homozygous form. The ratio of genes in homozygotes to those in heterozygotes is $2q^2 : 2pq$ ($2q^2$ because there are two mutant alleles in each homozygote). Assuming p is very close to 1, the ratio simplifies to just q (Figure 9.5). In other words, for a typical recessive with $q = 0.01$, the mutant alleles are exposed to selection only 1% of the time. Over evolutionary time selection will be effective, but mutant alleles can easily persist over a few dozen generations. The result of these different dynamics of selection is that serious dominant and X-linked conditions are likely to show a high proportion of new mutations, while for most recessives, mutation can be discounted in considering families, and individual pathogenic alleles may be quite common in a population.

Clinical geneticists are naturally most concerned with conditions that are both common and serious, and all such conditions must have somehow avoided being wiped out by natural selection. Possible mechanisms include the following.

- A very high mutation rate. DMD is an example. Probably the remarkable structure of the dystrophin gene (*Chapter 3*) contributes to this.
- Heterozygote advantage for a recessive condition. Thalassaemia (**Case 13, Nicolaides family**) is a clear example. As mentioned above, cystic fibrosis (**Case 2, Brown family**) must also have some heterozygote advantage, otherwise it could not be so common. If a severe recessive condition is relatively frequent in a large population, this could indicate a slight degree of heterozygote advantage or, alternatively, chance fluctuations in the transmission of the disease allele that bring that particular condition to attention.
- Symptoms mostly appear after reproductive age. Huntington disease (**Case 1, Ashton family**) is a classic example, as are the familial breast and colon cancers described in *Chapter 7*.

Figure 9.5 – Conditions with different modes of inheritance have different degrees of exposure to natural selection.

See text for calculations.



- Selection during spermatogenesis. As mentioned in the discussion of achondroplasia (**James Jenkins, Case 14**) in *Section 6.3*, in 2009 Goriely and colleagues showed that the apparent high mutation rate in achondroplasia is due to spermatogonia that carry the *FGFR3* mutation enjoying a proliferative advantage in the male testes. These authors identified the same process in a small set of other genes (*HRAS*, *RET*, *PTPN11*), producing dominant conditions with an apparently unusual frequency of *de novo* mutations, almost all of which are of paternal origin, and where there is a strong paternal age effect.

What is the chance the offspring of a consanguineous marriage will have a recessive disease?

Consider Jack and Jill who are first cousins. Suppose the frequency of a disease allele is q . The chance that Jack is a carrier is $2pq$. For a rare disease p is virtually 1, so Jack's carrier risk is very nearly $2q$. Jill shares 1/8 of her genes with Jack by virtue of their relationship. That is to say, for any allele in Jack, there is a 1 in 8 chance that Jill has the same allele, inherited from a common ancestor. Therefore if Jack is a carrier, the chance that Jill is a carrier is 1/8. If they are both carriers, the chance of a child being affected is 1 in 4. Overall the risk is $2q \times 1/8 \times 1/4 = q/16$. *Table 9.3* shows how, the rarer a condition is, the greater is the relative risk for first cousins compared to unrelated people.

Table 9.3 – Inbreeding and the risk of a recessive disease

Disease allele frequency	Risk of affected child – unrelated parents	Risk of affected child – first cousin marriage	Relative risk for first cousins
q	q^2	$q/16$	$1/16q$
0.01	1 in 10 000	1 in 1600	6.25
0.005	1 in 40 000	1 in 3200	12.5
0.001	1 in 1 000 000	1 in 16 000	62.5

The calculation is only valid for rare diseases because it assumes the frequency of the normal allele is effectively 1, and it assumes that the only way first cousins would both carry a disease allele is if they both inherit it from a common ancestor.

Another way of looking at this is that, the rarer a recessive condition is, the greater the proportion of cases are the offspring of consanguineous marriages. As mentioned before, this means that Hardy–Weinberg calculations of carrier frequencies are likely to be misleading when applied to rare recessives.

Can we abolish genetic disease?

Clinical geneticists are very clear that their goal is not to wipe out genetic disease, and their output must not be measured by the number of abnormal pregnancies terminated. Their goal is to enable people with genetic diseases, or at risk of them, to lead as normal lives as possible, including having families. Nevertheless, health planners might find it attractive if a by-product of genetic services were a reduction in the frequency of genetic disorders.

The discussion above of mutations and selection shows that abolishing genetic disease is not a generally realistic aim. Most serious dominant conditions have a high proportion of new mutations, while for recessives, the great majority of pathogenic alleles are in healthy heterozygotes. If a country's dictator decided to sterilize all people with serious genetic diseases, this would not prevent fresh cases appearing in the next generation. Even continued over many generations this program could not succeed in its aim. A few dominant adult-onset conditions such as Huntington disease could be largely prevented, but the only way to prevent recessive diseases would be to sterilize all carriers. This would rebound badly on the dictator and his family. Although carrier frequencies for individual recessive conditions are typically of the order of 1 in 100, OMIM lists thousands of different conditions, and every one of us is a carrier for several lethal recessive conditions. A sterilization program could prevent genetic disease only by sterilizing the dictator, along with all the rest of the human race.

A slightly different question is whether medicine is storing up long-term problems by effectively combating natural selection (see *Box 9.4*).

Should treated people repay their debt to society by not having children?

Consider a treatable genetic condition such as phenylketonuria (PKU; see **Case 20, Vlasi family** in *Chapter 10*). Untreated phenylketonurics have severe intellectual disability and are likely to need care all their lives and be unable to live independently. When the treatment works, an affected person is enabled to live a normal life, including having children. Genetically the person is still homozygous for the pathogenic allele, and any child will inevitably inherit a copy. The treatment is expensive – should there be a bargain? Should we say that we agree to fund the treatment, but in return the treated person should agree not to pass on his or her genes?

Most clinicians would find such a proposition deeply distasteful – but they may have at the back of their mind an uneasy feeling that, as responsible physicians, perhaps they ought to face the question. The question should indeed be faced, because when it is examined, it goes away. We showed above that for a recessive condition the proportion of all mutant alleles that are present in affected people is q , where q is the allele frequency. In the UK, PKU affects about one person in 10 000. Thus q is 0.01. In other words, agonizing about whether or not a treated phenylketonuric has the right to pass on their genes is agonizing about 1% of the problem while ignoring the other 99%. Because we know we can have no influence on the 99%, it is pointless to worry about the 1%. Of course, each time a treated phenylketonuric passes on a mutant allele rather than failing to reproduce, that must marginally increase the frequency of those alleles in the next generation. Continued over 100 generations, a treatment program might double the frequency of the pathogenic allele. Most people would feel that humanity has rather more serious problems to face over the next 2500 years than a doubling of the frequency of PKU alleles.

BOX 9.4

Identifying relatives – and criminals

Clinical geneticists are not the only ones interested in looking at the extent to which individuals share DNA by virtue of a shared common ancestor. Police in many countries have long used DNA profiling to match a crime scene sample to the profile of an individual in their national DNA database. Where there is no perfect match, they may wish to see if the crime scene sample might have come from the brother or son of somebody in

their database. This practice (familial searching) is ethically contentious, and is permitted in some jurisdictions but prohibited in others. However, it has limited power. Forensic databases are based on genotypes of just a dozen or so microsatellites. These are ample to identify a full match, but this limited amount of genomic data is insufficient to identify more than first-degree, or possibly second-degree relatives. The wide availability of genome-wide SNP genotypes has opened new possibilities.

Millions of people send their DNA to companies that genotype it for a million or so SNPs and report back deductions about ancestry, eye color (wouldn't a mirror do that?) and possibly health. The companies are careful not to make individually identifiable information public. However, some people choose voluntarily to surrender their anonymity in order to try to find unknown relatives who may have done the same. Companies processing these requests look for shared segments in the genomes of pairs of individuals. On the basis of the number and length of shared segments they report a likely degree of relationship. The comprehensive SNP genotypes allow quite distant relationships to be identified.

The process is powerful because, although distant relatives share only small amounts of their DNA, the number of your distant relatives increases rapidly with distance. If every couple had two children you would have one sib, 7 first cousins, 15 second cousins and $(2^p - 1)$ relatives of degree p . A database with 1 million profiles drawn at random from a national population of 100 million is highly likely to include the profile of at least one of your 127 5th cousins, with whom you might share a detectable block of your genome. The real extent of sharing varies, of course, depending on the precise genealogy together with random events at every meiosis, but it is often sufficient to allow distant relatives to locate each other.

While individuals may relish the opportunity to identify unsuspected relatives, police forces may see a new tool for catching criminals. In a powerful extension of familial searching, the full genome-wide SNP genotype of a crime scene sample can be used to identify innocent distant relatives of the unknown criminal, and hence provide valuable leads for the investigation. The use of this method to identify a suspect in the case of the so-called Golden State Killer, an infamous serial murderer and rapist whose case has been open for over 40 years, brought widespread attention in 2018.

For an accessible discussion of these matters by well-informed population geneticists, including many caveats, see <https://gcbias.org/2018/05/07/how-lucky-was-the-genetic-investigation-in-the-golden-state-killer-case/>. Note that this cannot work with the limited genetic information in national forensic databases, and crime scene samples often contain too little DNA, or too mixed or degraded DNA, to allow 1 million SNPs to be genotyped. Despite these limitations, there are serious questions about the wisdom of making individual genotypes public, and this feeds into health-related discussions. The more extensively biobank data is shared, the more powerful it is for research – but the more it risks impinging on individual privacy or coming up against national data protection laws.

Jewish diseases and Finnish diseases

Ashkenazi Jews and Finns are the two best-studied examples of populations with strong founder effects. Both have high levels of education and well developed medical services that provide good data on the range and frequency of genetic diseases. *Table 9.1* listed some Finnish diseases, and *Box table 9.1* in this section shows some Ashkenazi Jewish diseases. Motulsky (1995) provides a good concise review of Jewish diseases.

Both populations have expanded greatly from a fairly small number of founders.

- Of the 13–14 million Jews in the world, about 80% are Ashkenazi, descendants of a population that migrated to the Rhineland in Germany in the ninth century, and later moved into Poland, Lithuania, Belarus and surrounding regions. For over 1000 years Ashkenazim have been distinct from Sephardic Jews, who are primarily from Spain, Portugal and North Africa. Until a few hundred years ago Sephardim constituted the majority of all Jews, and their communities in Spain and Portugal may have existed since Roman times. The Ashkenazi population probably comprised a few thousand individuals in the early modern period. Motulsky suggests that a small prosperous merchant class may have contributed very disproportionately to succeeding generations, producing very strong founder effects.
- In Finland there were two phases of expansion, either of which could have produced strong founder effects: one in prehistoric times soon after the south of the country was first settled, and another in the seventeenth century when pioneer groups populated the largely empty north (De la Chapelle, 1993). The Finns, incidentally, illustrate that language is not a reliable pointer to genetic origins. Finnish is completely unrelated to the Indo-European languages spoken by Finland's neighbors. It is a Uralic language, with deep affinities to various Siberian languages; it is closely related only to Estonian and two minority languages of North-Western Russia, Karelian and Veps. Nevertheless, Finns are genetically European, not Siberian. At some time in remote history the proto-Finns must have adopted a new language.

Box table 9.1 – Some diseases that are unusually prevalent among Ashkenazi Jews

Disease	OMIM number	Mode of inheritance	Carrier frequency	Comments
Tay–Sachs disease	272800	AR	1/30	Three fairly common Ashkenazi alleles – see Case 19 (Ulmer family) and <i>Chapter 11</i>
Familial dysautonomia	223900	AR	1/30	Very few non-Jewish cases reported; 99.5% of Jewish cases have the same splicing variant
Gaucher disease	230800	AR	1/15	Two common Ashkenazi alleles p.Asn380Ser (75%) and c.84insG (15%); p.Leu444Pro also frequent
Canavan disease	271900	AR	1/40 – 1/60	Two common Ashkenazi alleles, p.Glu285Ala and p.Tyr231STOP
Fanconi anemia complementation group C	227645	AR	1/90	Most Ashkenazi cases have a splice site change in intron 4
Niemann–Pick disease Type A	257200	AR	1/80	Three variants account for 65% of Ashkenazi cases
Bloom syndrome	210900	AR	1/200?	Most Ashkenazi cases have the same variant, deletion of ATCTGA and insertion of TACATTC at nucleotide 2281

Disease	OMIM number	Mode of inheritance	Carrier frequency	Comments
Familial breast cancer	113705 (<i>BRCA1</i>) 600185 (<i>BRCA2</i>)	AD	Total 2.2%	Three common Ashkenazi alleles: c.185delAG, c.5382insC in <i>BRCA1</i> ; c.6174delT in <i>BRCA2</i> .
Torsion dystonia 1	128100	AD	1/1000 – 1/300 affected	Genetics heterogeneous; many Ashkenazi and non-Jewish patients have the same 3 bp deletion

AD, autosomal dominant; AR, autosomal recessive.

There has been controversy over whether the distribution of alleles in Ashkenazi diseases can be explained just in terms of population history. With a simple founder effect, one might expect one pathogenic allele, on one specific marker haplotype, to explain all the excess cases of a disease (excess over the levels found in other populations). That is what we see in familial dysautonomia, Fanconi anemia and Bloom syndrome. But in Tay–Sachs, Gaucher, Niemann–Pick and Canavan diseases, two or more alleles are relatively common. This has been taken as evidence of heterozygote advantage. It is noteworthy that three of the four diseases listed are lysosomal storage diseases (Tay–Sachs, Gaucher and Niemann–Pick disease), and in fact another lysosomal storage disease, mucopolidosis IV, is also relatively common in Ashkenazim. Is this just coincidence, or could there be something about being a carrier of a lysosomal storage disease that was advantageous at some time in the history of this population? Lysosomes have often been pictured as simple intracellular garbage cans, the sole function of which is to degrade unwanted high molecular weight material – but it has become apparent that they have a much wider role as signaling systems that help to set the balance of cell metabolism between catabolism and anabolism. Thus it is not impossible to imagine that there might be circumstances in which a reduced dosage of certain lysosomal enzymes might be advantageous.

A natural observational bias could explain some of the data. For example, three different *BRCA1/2* mutations are fairly common in Ashkenazim and rare elsewhere (see *Chapter 12*). This seems remarkable – but it may be just an initial random fluctuation, amplified by the strong population expansion. Before demanding a special explanation, we should remember all the other genes where we don't find common Ashkenazi mutations. Maybe it is not so remarkable that one gene in a thousand should show a random fluctuation large enough to produce the present distribution.

9.5. References

- De la Chapelle A** (1993) Disease gene mapping in isolated human populations: the example of Finland. *J. Med. Genet.* **30**: 857–865.
- Goriely A, Hansen RMS, Taylor IB, et al.** (2009) Activating mutations in *FGFR3* and *HRAS* reveal a shared genetic origin for congenital disorders and testicular tumors. *Nature Genetics*, **41**: 1247–1252.
- Motulsky AG** (1995) Jewish diseases and origins. *Nature Genetics*, **9**: 99–101.
- Niskanen M** (2002) The origin of the Baltic–Finns from the physical anthropological point of view. *The Mankind Quarterly*, **43**: 121–153.

Paw BH, Tieu PT, Kaback MM, Lim J and Neufeld EF (1990) Frequency of three HexA mutant alleles among Jewish and non-Jewish carriers identified in a Tay–Sachs screening program. *Am. J. Hum. Genet.* **47**: 698–705.

9.6. Self-assessment questions

- (1) 200 unrelated people were typed for a single nucleotide polymorphism that has two alleles, C and T. 87 people were CT, 93 TT and 20 CC. What are the frequencies of the C and T alleles? Is the population in Hardy–Weinberg equilibrium?
- (2) Usher syndrome Type 1 is an autosomal recessive deaf–blindness syndrome that affects one person in 100 000 in a population. Although all cases are clinically indistinguishable, genetic analysis has shown that mutations in several different genes can cause Usher syndrome (locus heterogeneity). What is the total frequency of carriers, (a) if all cases are due to mutations at a single locus, and (b) if homozygosity at any one of 10 different loci contributes equally to the overall incidence?
- (3) Inability to taste low concentrations of bitter substances such as phenylthiourea and perhaps some types of cabbage is (usually) an autosomal recessive trait. People can be classified as tasters and non-tasters. 64% of people in a population dislike spring greens and won't eat them because they taste unpleasantly bitter. What is the frequency of the taster allele?
- (4) It is known that erythropoietic protoporphyria behaves as a dominant condition with reduced penetrance, but molecular analysis has shown that affected people are all compound heterozygotes, having a rare non-functioning allele and a common low-functioning allele ($q = 0.11$ in France). If the condition affects 1 person in 30 000 in France, what is the frequency of the non-functioning allele? What is the risk that an affected person, married to an unaffected person, will have an affected child?
- (5) A woman's only son has Duchenne muscular dystrophy. She has no brothers or sisters, and there is no history of the disease in the rest of the family. What is the chance her daughter is a carrier?
- (6) The healthy sister of a boy with cystic fibrosis marries an unrelated man whose family has no history of the disease. Both are of Danish stock. She is pregnant. What is the risk the child will have CF? If she had consulted you beforehand, would you have advised her not to have children?
- (7) An autosomal recessive disease affects one person in 40 000. A woman's marriage breaks up under the stress of caring for her affected child. She finds solace with a sympathetic cousin; eventually they marry and now she is pregnant. What is the risk their child will be affected?
- (8) Fred has an extremely rare autosomal recessive disease. His mother is Turkish, his father Nigerian. What is the chance that both of his grandfathers are carriers?
- (9) Waleed and Benazir (*Figure 9.4*) marry and have a deaf son, Aziz. Deaf people often prefer to marry a deaf partner, and Aziz marries Nasreen Choudhary. What is the chance their first child will be deaf? Calculate the coefficient of inbreeding of this child. Apart from deafness, is this child at high risk of other recessive conditions?

[Hints on questions 2, 3, 4, 5 and 6 are provided in the Guidance section at the back of book.]

10

How do our genes affect our metabolism, drug responses and immune system?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe the basic principles of inborn errors of metabolism and give examples of diseases caused by metabolic blocks
- Give examples of individual variations in response to drugs, explaining their importance
- Discuss critically the prospects for personalized medicine based on genetic testing
- Describe the general nature and function of the major histocompatibility complex and the role of HLA matching in transplantation
- Describe in outline the genetic mechanisms underpinning our ability to mount a specific immune response against virtually any foreign antigen

10.1. Case studies

CASE 20 VLASI FAMILY

- Valon, 6-year-old boy with serious learning problems
- Small, microcephalic, blue eyes, fair skin and hair, eczema; hyperactive
- ? Phenylketonuria



251 262 317 395 396

Valon, aged 6 years, was the only child of Adem and Flora Vlasi. They had had a very unsettled life. Valon was born in Kosovo. The family lived in a remote rural area where only basic medical care was available. Soon after he was born the political situation became unstable and the family moved several times before entering Australia as refugees. Over the years Adem had been worried about Valon's progress, but Flora said the problems were likely to be due to all the moves he had experienced and the fact that he had not been to school. When they settled and Valon was enrolled in school the teachers recognized at once that he had serious learning problems. They arranged for assessment by an educational psychologist and suggested the family doctor refer him to a paediatrician. In the clinic the paediatrician was surprised to see that Valon had a condition he had only read about in textbooks. Valon was small and microcephalic and had blue eyes, very fair skin and hair, as well as eczema. He was hyperactive and when restrained he rocked his body. The paediatrician also thought that Valon had a musty smell about him in spite of him being well cared for by his parents. The doctor strongly suspected that Valon had phenylketonuria and arranged for measurement of phenylpyruvic acid in a urine sample and phenylalanine levels in a blood sample.

Figure 10.1 – A patient with untreated PKU.

CASE 21 PORTILLO FAMILY

- Sickly boy, Pablo
- Family history of similar problems
- X-linked severe combined immunodeficiency

252

263

286

395

Pilar and Pedro Portillo come from very close families and three generations live in the same part of town. When Pablo was born in 1988 Pilar and Pedro were happy that they had three children, but Pablo was a much more sickly baby than his siblings. He always seemed to have a cough, an ear infection or diarrhea and he failed to gain weight. Pilar's maternal grandmother encouraged Pilar to get an appointment for Pedro at the specialist children's hospital because Pablo's problems were very similar to those of her own two sons who had both died before they were one year old. She hoped there might be treatment available to stop Pablo deteriorating further. At the hospital Pablo was admitted straight away for investigations.

The blood tests showed that Pablo had a very low lymphocyte count. T cells and NK (natural killer) cells were absent; B cells were present but non-functional. *Box 10.1* gives some basic details about these cells. These findings, together with the family history, suggested a diagnosis of X-linked severe combined immune deficiency (X-SCID). This was very bad news because without successful treatment the prognosis is very poor. The doctors suggested bone marrow transplantation was Pablo's best hope.



(a)



(b)

Figure 10.2 – Problems with immunodeficiency.

(a) Failure to thrive and skin problems. (b) Herpes simplex developing over an area with eczema (Koebner phenomenon). Photo (a) courtesy of Dept of Medical Illustration, Manchester Royal Infirmary and (b) courtesy of Dr Andrew Will, Royal Manchester Children's Hospital.

Types and functions of lymphocytes

All three types of lymphocyte are derived from bone marrow. B cells and NK cells mature in the marrow but T cells undergo a process of maturation in the thymus gland. B cells give rise to plasma cells that secrete immunoglobulins. NK cells are large granular lymphocytes with a characteristic morphology; they account for up to 15% of blood lymphocytes and provide a first line of defense against virally infected cells. **T lymphocytes** are involved in the regulation of the immune response and in cell-mediated immunity, and they help B cells to produce antibody. Mature T cells express antigen-specific T cell receptors plus the CD3 molecule. In addition, mature T cells express either CD4 or CD8 cell surface molecules that enable them to play a role in cell and antibody-mediated immunity (CD4+) or to become cytotoxic (CD8+).

10.2. Science toolkit

In this chapter we cover three areas of genetics, all of which concern causes of genetic differences between people:

- inborn errors of metabolism
- common variation in drug responses (**pharmacogenetics**)
- the systems that enable us mount an immune response against virtually any foreign antigen, but that also lead to transplant rejection (**immunogenetics**).

The present section and *Section 10.4* are each divided into three parts, covering these three areas.

Inborn errors of metabolism

The concept of inborn errors of metabolism was developed at the very dawn of clinical genetics (see *Box 10.2*). If a metabolic pathway requires the sequential action of several enzymes, a loss of function of any one of those enzymes will lead to a metabolic block (*Figure 10.3*). Substrate accumulates before the block, and there is a lack of product downstream of the block.

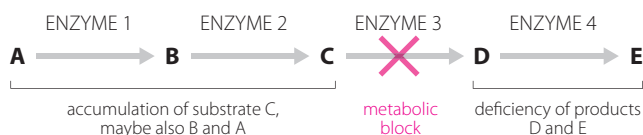


Figure 10.3 – Effects of a metabolic block in a simple pathway.

- In a biosynthetic pathway the most noticeable effect is likely to be the absence of the end product. For example, tyrosinase is the key enzyme for biosynthesis of melanin, and homozygous loss of function of tyrosinase causes albinism (*Figure 10.4*).
- In a degradative pathway, most often it is the accumulation of the blocked substrate that causes problems. The lysosomal storage diseases are typical examples. As briefly discussed in *Disease box 9*, lysosomes are vesicles that contain a collection of around 40 different hydrolytic enzymes that are required to degrade a variety of large molecules. Lysosomes import high molecular weight materials, but can only export their low molecular weight breakdown products. A deficiency of one or other of the lysosomal enzymes therefore causes undegraded or partially degraded high molecular weight material to accumulate within the lysosome. This can lead eventually to the death of the cell.

Patients with lysosomal storage diseases are normal at birth, but there is a progressive deterioration as undegraded material builds up in the lysosomes. We have already met one lysosomal storage disease in *Chapter 9*, Tay–Sachs disease in the **Ulmer family**, **Case 19**. *Figure 9.1* illustrated the consequences of accumulation of undegraded G_{M2} ganglioside in lysosomes. *Disease box 9* mentioned several other such diseases. Other frequent examples include mucopolysaccharidoses like Hunter syndrome (OMIM 309900) and Hurler syndrome (OMIM 607014). In these diseases lack of one or another enzyme required for the breakdown of glycosaminoglycans (mucopolysaccharides) leads

to accumulation of undegradable and unexportable high molecular weight material, with serious clinical consequences.

- The high concentration of substrate upstream of a metabolic block can also lead to the production of abnormal metabolites. Phenylketonuria (see **Vlasi family, Case 20** in Section 10.3) provides one example of this phenomenon. Production of an abnormal metabolite is the pathogenic mechanism in Type I tyrosinemia (OMIM 276700). Maleylacetoacetic acid (see Figure 10.4) is eliminated by first being converted into fumarylacetoacetic acid, which is then normally broken down by fumarylacetoacetate hydrolase (FAH) to fumarate and acetoacetate. Type I tyrosinemia is caused by lack of FAH. Fumarylacetoacetic acid accumulates and spills over into production of the toxic succinylacetone. This abnormal

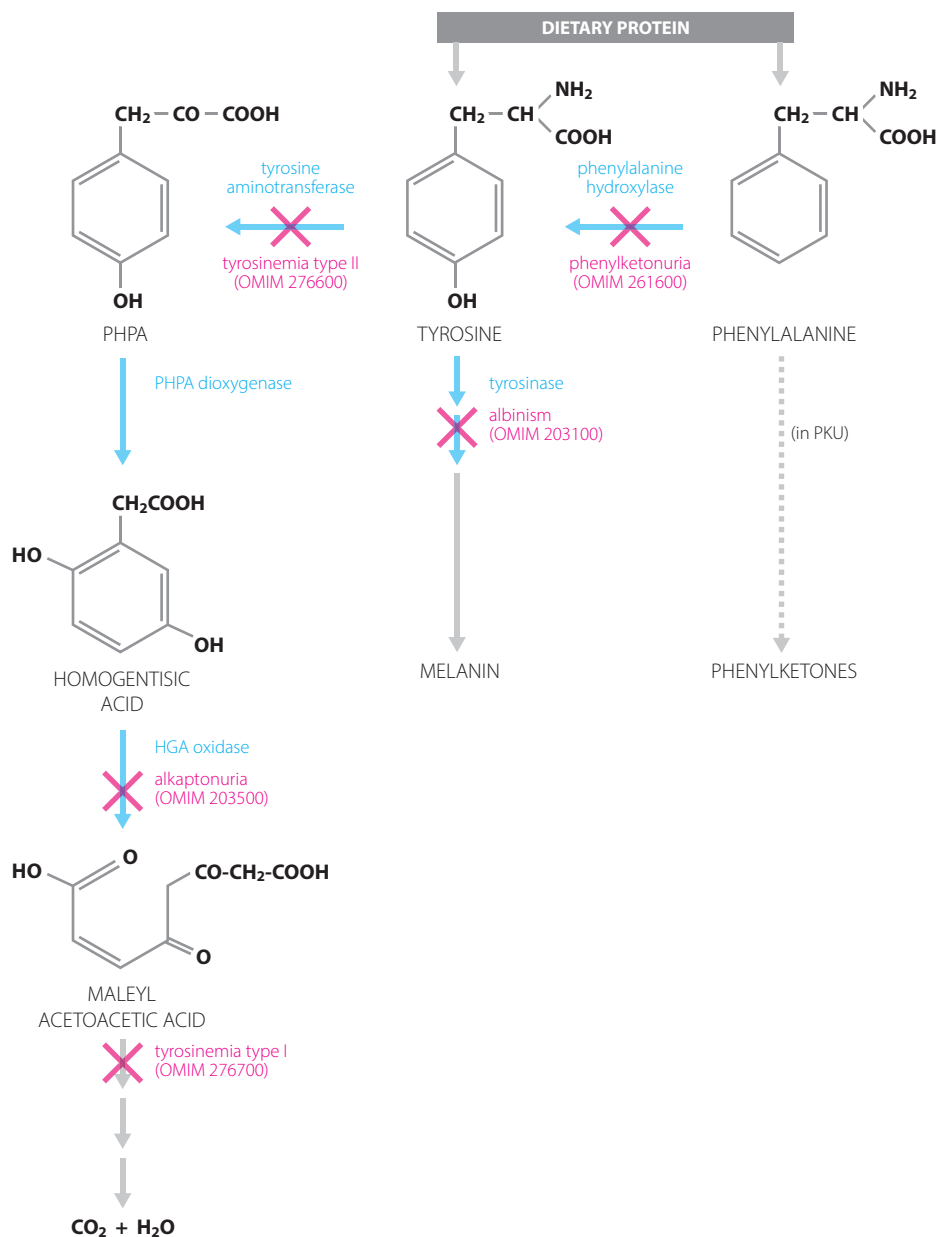


Figure 10.4 – Metabolism of the amino acids phenylalanine and tyrosine.

Consequences of a metabolic block can include absence of product (albinism), excretion of the material immediately upstream of the block (alkaptonuria), or excretion of alternative metabolites of a blocked substrate (phenylketonuria). HGA, homogentisate; PHPA, *p*-hydroxy phenylalanine.

metabolite causes the kidney and liver disease which are the hallmark of Type I tyrosinemia. The disease is treated with the drug Nitisinone, which is an inhibitor of PHPA dioxygenase (see *Figure 10.4*). The effect is to convert the lethal Type I tyrosinemia into the milder Type III disease, which is then treated with dietary restriction of tyrosine and phenylalanine.

Inborn errors of metabolism are among the most treatable genetic conditions because they often respond to dietary manipulation or drugs. For this reason, newborn screening programs often include many individually very rare inborn errors, as described in *Chapter 12*. Some of the treatments are summarized in *Chapter 14*.

Some history

The concept of *inborn errors of metabolism* goes back to the very early days of human genetics. In 1902 Archibald Garrod published a paper on 'The incidence of alkaptonuria: a study in chemical individuality'. Alkaptonuria (OMIM 203500 – see *Figure 10.4*) is a rare recessive condition in which affected individuals lack homogentisate 1,2-dioxygenase (also called homogentisic acid oxidase) and accordingly excrete in their urine large amounts of homogentisic acid, an intermediate in the catabolism of phenylalanine and tyrosine. This readily darkens on exposure to air, hence the name. Garrod noted that parents of patients were often first cousins, and that brothers and sisters were sometimes affected. He speculated that alkaptonuria might be a mendelian recessive condition – a remarkable insight, given that Mendel's work had only been rediscovered two years previously. In a series of lectures in 1908, Garrod coined the term 'inborn error of metabolism' and suggested cystinuria and pentosuria as further examples.

Rather like Mendel, Garrod was perhaps ahead of his time. Geneticists at the time were more concerned with understanding the basic mechanisms of heredity, and biochemists with understanding basic biological chemistry. Patients with exceedingly rare diseases were not amenable to experimental investigation. It was over 30 years later that Beadle and Tatum developed a suitable experimental system, X-ray mutagenesis and biochemical analysis of the fungus *Neurospora crassa*. Their 1941 paper 'Genetic control of biochemical reactions in *Neurospora*' does not contain the phrase associated with their names, 'one gene – one enzyme' but does say:

It should be possible, by finding a number of mutants unable to carry out a particular step in a given synthesis, to determine whether only one gene is ordinarily concerned with the immediate regulation of a given specific chemical reaction.

Within five years, Beadle had clearly enunciated the one gene – one enzyme hypothesis and placed it at the center of contemporary understanding of gene action. At that time neither the structure of proteins nor that of genes was known. A further seminal development was the demonstration by Ingram in 1956 of the difference between normal hemoglobin and hemoglobin S. By the early 1960s all the basic concepts of biochemical genetics were in place.

Biochemical genetics is defined more by its methods and practitioners than by anything else. During the period 1960–1990, before PCR became routine, possibilities for clinical genetic testing by DNA analysis were limited. Biochemists, however, had a sophisticated knowledge of metabolic pathways and enzymology, and they applied this to any suitable genetic condition. They used specialized tools such as gas chromatography coupled to mass spectrometry (GC–MS) to identify abnormal metabolites in blood or urine. They ran large-scale newborn screening programs, and were closely involved in managing the children in whom they diagnosed inborn errors. All this set them a little apart from mainstream clinical genetics. Nowadays the two have very much come together. The biochemists have not abandoned their GC–MS and other special technologies, but the use of DNA methods and of concepts from biochemistry and cell biology is now universal across all of clinical genetics.

Pharmacogenetics

Drugs often work in only a proportion of the people taking them, and some drugs can produce unwanted or dangerous effects in some people. Adverse drug reactions (ADRs) are a serious clinical problem. It has been estimated that they are responsible for around 100 000 deaths each year in the USA. In the UK a study of 18 820 consecutive admissions to two large general hospitals in 2001–2 concluded that 6.5% of admissions were related to ADRs, and 2% of those patients died. The projected annual cost to the NHS was up to £466 million (Pirmohamed *et al.*, 2004). These variable responses can have many different causes. Many are non-genetic, such as the age, sex, body mass or lifestyle (drinking, smoking, exercise, etc.) of the patient, concomitant illnesses and interactions with other drugs the patient may be taking – but often the cause is genetic (*Table 10.1*). Genetic factors affect both **pharmacokinetics** (the absorption, distribution, metabolism and excretion of the drug – in other words, what the body does to a drug) and **pharmacodynamics** (the actual effect of the drug on its target organ, or what a drug does to the body).

Adverse reactions can be classified into Type A and Type B. ADRs of Type A reflect individual quantitative differences in the normal metabolism of a drug. They account for 80–95% of all reported ADRs and are generally dose-dependent and predictable. A person who eliminates a drug unusually slowly is exposed to a higher effective dose of the drug than somebody who eliminates it rapidly. Thus a normal dose of the drug may produce symptoms of an overdose (see *Box 10.3*). The same is true of individuals who activate a prodrug unusually efficiently. By contrast, Type B ADRs are idiosyncratic reactions to a drug, unrelated to its normal mode of action. Type B ADRs are dose-independent, hard to predict and can be life-threatening. In *Table 10.1* only the reaction to carbamazepine is a Type B reaction. Whereas all the other examples are understandable in terms of the normal metabolism of each drug, it is not at all clear why a person's HLA type (see discussion of immunogenetics, below) should have any relevance to the effect of carbamazepine.

Many of the most striking individual differences in response, both in terms of efficacy and the risk of side effects, are due to large individual variations in the rate of metabolism of drugs. Many different enzymes are involved with different specific drugs (see the reviews by Evans and McLeod, 2003, and Weinshilboum, 2003) but the most frequent players are enzymes of the P450 family, reviewed by Guengerich (2008). This large family of iron-containing enzymes, known as cytochromes P450 after their spectral peak of light absorption, use molecular oxygen to insert hydroxyl or related groups into a wide variety of organic molecules. They are responsible for the initial metabolism of perhaps 75% of all drugs. Three family members, CYP2C9, CYP2C19 and CYP2D6 are especially important in drug reactions. For each of these enzymes, common polymorphisms affect their activity. People can be classified as poor, intermediate, extensive and (for CYP2D6) ultra-rapid metabolizers (*Figure 10.5*). For many drugs, a P450-catalyzed oxidation is the first step in elimination of the drug. The clinical effect of a given dose of drug is greater in a poor metabolizer, and much less in an ultra-rapid metabolizer because of the different rates of elimination. Ultra-rapid metabolizers may get no benefit from a standard dose of the drug, while poor metabolizers may suffer effects mimicking an overdose (see *Box 10.3* for an example). Some prescribed drugs are **prodrugs** that require enzymic conversion into their active form. Codeine is converted by CYP2D6 to its active form, morphine. Poor metabolizers get no pain relief from standard doses of codeine, while ultra-rapid metabolizers are at increased risk of adverse effects such as breathing problems and sedation.

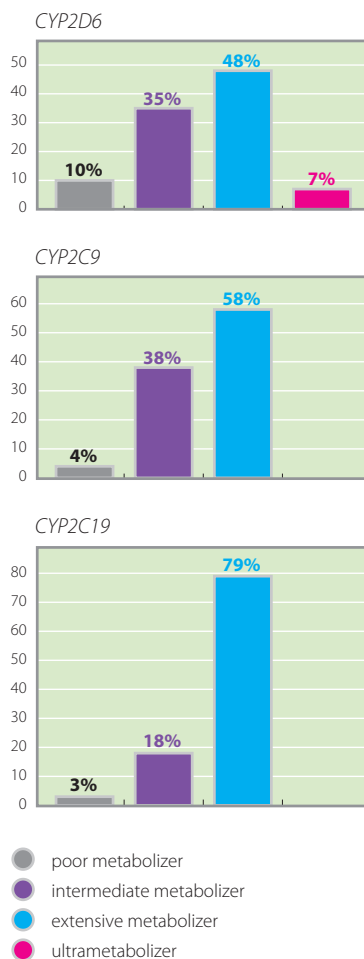


Figure 10.5 – Polymorphisms in the CYP2D6, CYP2C9 and CYP2C19 genes cause common variations in enzyme activity.

Variation is present in all populations but at different frequencies; these figures are for white people of Northern European origin and are redrawn from the data of Service (*Science*, 2005; **308**: 1858–1860).

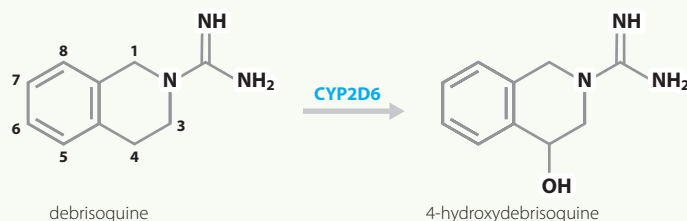
Table 10.1 – Examples of drugs that can produce serious side effects in people with certain genotypes

Drug	Effect
Azathioprine	Life-threatening bone marrow suppression from normal dose in people with low activity thiopurine methyl transferase
Carbamazepine	Life-threatening Stevens–Johnson syndrome in East Asians with HLA-B*1502 and Europeans with HLA-A*3101 genotypes
Fluorouracil	Potentially fatal toxicity in people with deficiency of dihydropyrimidine dehydrogenase
Irinotecan	Severe neutropenia and diarrhea in people homozygous for a low-activity variant of the <i>UGT1A1</i> gene
Isoniazid	Risk of polyneuropathy in slow acetylators
Succinylcholine	Prolonged apnea in people with butyrylcholinesterase deficiency
Warfarin	Excessive bleeding in people with low-activity <i>CYP2C9</i> or certain forms of <i>VKORC1</i>

Robert Smith's debrisoquine misadventure

Debrisoquine was a drug used to control high blood pressure. It is no longer prescribed, but it was the subject of an important historical episode.

In 1975, back in the heroic times when scientists experimented on themselves, Bob Smith, a laboratory director at St Mary's Hospital Medical School in London, ingested 32 mg of debrisoquine, as did some of his co-workers. He later described his adverse response to the drug: "Within two hours severe orthostatic hypotension [low blood pressure] set in with blood pressure dropping to 70/50 mmHg. Hypotensive symptoms persisted for up to two days after the dose...". His colleagues, who had taken a similar dose, had no significant effects.



Box figure 10.1 – The first step in elimination of debrisoquine is a reaction that is mediated by the CYP2D6 enzyme, producing 4-hydroxydebrisoquine.

Analysis of 4-hydroxydebrisoquine in the urine of the volunteers revealed that the extreme sensitivity was associated with a greatly decreased ability to carry out the hydroxylation reaction that removes debrisoquine. A later study of a larger number of participants led to the description of a genetic polymorphism, and eventually individuals were divided into four classes of 'ultra-rapid', 'extensive', 'intermediate' and 'poor' metabolizers, which reflect the variation in the activity of CYP2D6. Professor Smith was, of course, a 'poor' metabolizer.

Many other enzyme systems are involved in pharmacokinetic variations. A long-standing example is the surgical muscle relaxant suxamethonium (succinylcholine). Normally its effect is short-lived because it is rapidly broken down by butyrylcholinesterase. Individuals who are homozygous for a low-activity variant of the enzyme (see OMIM 177400) are unable to eliminate the drug in this way and suffer dangerously prolonged apnea. Other drugs are metabolized via acetylation, and people can be divided into fast and slow acetylators depending on the activity of *N*-acetyl transferase.

The examples above all concern pharmacokinetics. There are not so many examples of major pharmacodynamic effects, where common variations in a drug target have a major effect on the performance of a drug but, increasingly, drugs are being designed to act on one specific target genotype. Patients are grouped (stratified) by genotype and given genotype-specific treatments. **Stratified medicine** is well established in oncology, because different tumors have different acquired somatic mutations that drive their growth. Specific drugs target specific mutant proteins, and prescribing is governed by prior genotyping. *Disease box 7* showed one example, and this topic is considered more systematically in *Section 10.4*. Similar combinations of a drug and a companion diagnostic are increasingly seen as the future in other branches of medicine.

Immunogenetics

Immunogenetics involves two main aspects:

- understanding how we can produce an apparently infinite number of different specific antibodies, in a clear exception to the one gene – one polypeptide hypothesis
- understanding how we can distinguish self from non-self, and mount an immune response against almost any foreign cell or antigen.

Here we outline the genetics behind the recognition problem; the mechanisms for generating antibody diversity are outlined in *Section 10.4*. Immunogenetics is a large subject, and the treatment here is necessarily introductory. Any recent immunology textbook will go much more deeply into this fascinating area of genetics. For background and much further detail on the recognition problem as described below, a good source is the latest edition of *Janeway's Immunobiology* by Murphy and Weaver (2016); an earlier edition is available on the NCBI Bookshelf.

As is well known, transplanted organs are rejected unless they are tissue-matched. This will turn out to be a problem for **Pablo Portillo (Case 21)** who needs a bone marrow transplant (see *Section 10.3*). The major determinants of rejection are antigens encoded by genes in the major histocompatibility complex (MHC) on chromosome 6p21.3. Transplants that are fully matched at the MHC will not normally be rejected. Unfortunately, full MHC matching is seldom achievable, except between identical twins, because the MHC is the most polymorphic and variable region in the human genome.

Genes are densely packed within the MHC – the 'classical MHC' contains around 200 genes in a 4.1 Mb region; the 'extended MHC' (*Figure 10.6*) is 7.6 Mb long and contains over 400 genes – though nearly half of these are non-expressed pseudogenes. Unusually, many of the genes are functionally related. In higher organisms, functionally related genes are usually scattered apparently randomly over the genome, except for some clusters of recently duplicated and diverged genes. In the MHC, however, most of the

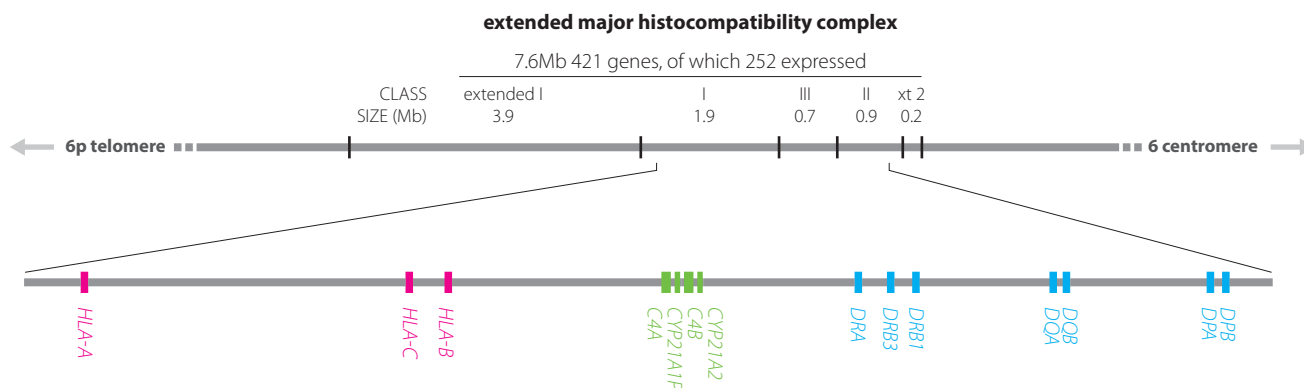


Figure 10.6 – The human MHC at 6p21.3 is conventionally divided into Class I, II and III regions, with extensions at either end.

The figure shows the most important genes for tissue matching, together with the C4/21-hydroxylase cluster. Data from Horton *et al.* (2004) *Nat. Rev. Genet.* **5**: 889–899.

genes play some part in the immune process – although there are exceptions, such as the 21-hydroxylase gene that is involved in steroid metabolism. Within the MHC, the key determinants of self versus non-self recognition are cell surface molecules encoded by a series of structurally related genes, the Human Leukocyte Antigen (HLA) genes.

HLA molecules are divided into Class I and Class II.

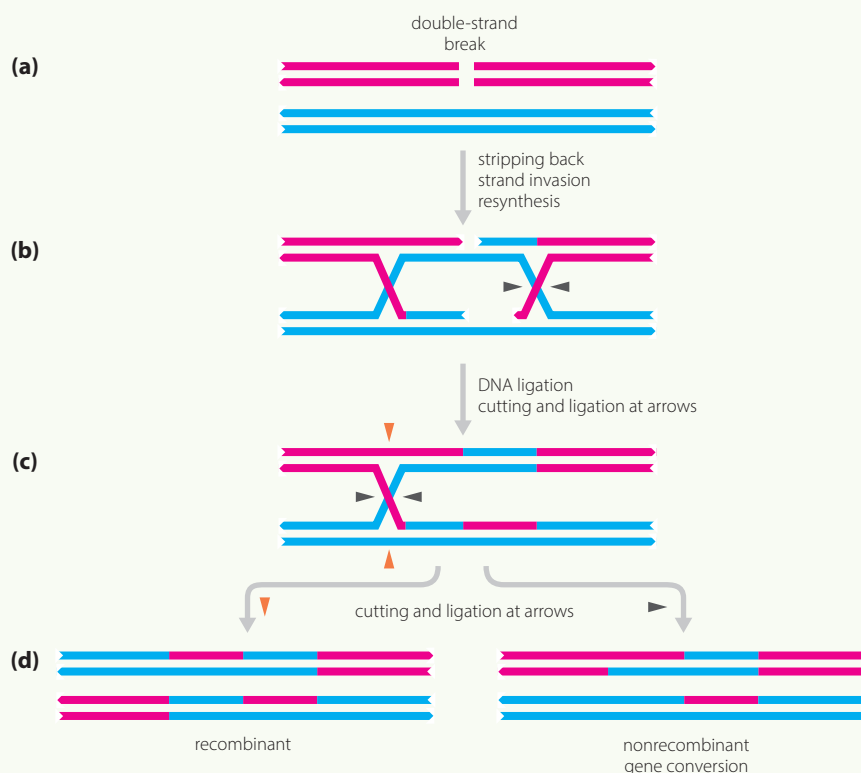
- Class I molecules are present on the surface of most nucleated cells. They consist of a heavy chain, encoded by an HLA gene, and a constant light chain, β 2-microglobulin, encoded on chromosome 15. There are 26 Class I genes in the MHC, but only nine are functional. The HLA-A and HLA-B molecules are the most important Class I antigens for tissue matching. Both loci are highly polymorphic: with 511 recorded alleles the HLA-B locus is the most polymorphic in the human genome.
- Class II antigens are found mainly on B lymphocytes and macrophages. They consist of alpha and beta chains, both encoded within the MHC. Of the 24 Class II loci, 15 are functional. The main class II molecules are DR, DP and DQ, and again these are highly polymorphic (in 2001 a World Health Organization committee listed 323 DR β alleles).

Clearly there has been selection in favor of this very extensive variability. Alleles frequently differ by substantial blocks of amino acid residues, suggesting that recombination and **gene conversion** (see Box 10.4) have both been important in generating the diversity.

HLA molecules present peptides derived from foreign proteins to T lymphocytes. Class I molecules present endogenous antigens to CD8⁺ T cells, while Class II molecules present exogenous antigens to CD4⁺ T cells. T cells initiate an immune response to either non-self peptides presented by self-HLA molecules, or cells carrying non-self HLA molecules. T cells that respond to self peptides presented by self-HLA molecules occur but are eliminated during early development ('clonal deletion'). The response to cells carrying non-self HLA molecules is the cause of transplant rejection. Ideally a transplant donor and recipient should be matched for both alleles at the HLA-A, -B and -DR loci (i.e. 6 matches). With modern immunosuppressive treatments, transplantation across mismatches is often successful, but immunosuppression brings problems of its own.

Recombination and gene conversion

As mentioned in *Section 2.2*, recombination (crossing over) in the first division of meiosis involves a lot more than a simple exchange of chromosome segments. It is initiated by a double-strand break in one of the chromosomes. The broken ends are then stripped back. As *Box figure 10.2* shows, subsequent events involve strand invasions, DNA synthesis and cutting and ligation of DNA strands.



Box figure 10.2 – Depending on which DNA strands are involved in the second round of cutting and ligation (steps c–d), the final result is two recombinant chromosomes (cuts at orange arrowheads) or non-recombinant chromosomes (cuts at black arrowheads). But with gene conversion, a small sequence from the blue DNA has been patched into the red chromosome. Both mechanisms contribute to the great genetic diversity at the major histocompatibility locus. Reproduced from Strachan & Read (2019) *Human Molecular Genetics* 5e, with permission from Garland Science/Taylor & Francis LLC.

Gene conversion is probably a frequent event in normal meiosis, comparable in frequency to recombination, but is difficult to detect. In humans it was first identified in individuals with loss of function changes in the *CYP21A2* gene (**Figure 10.6**) whose specific variant was the presence of portions of sequence of the nearby *CYP21A1P* pseudogene – evidently the product of non-allelic homologous pairing resolved by gene conversion rather than recombination (see OMIM 613815).

10.3. Investigations of patients

CASE 15 TIERNEY FAMILY

- 4-year-old boy, Jason
- Pale with extensive bruising and tachycardia
- ? Acute lymphocytic leukemia
- Diagnosis of ALL confirmed with *TEL-AML1* fusion gene
- TPMT test prior to chemotherapy
- Severe adverse reaction after false negative TPMT result
- Possibilities for therapy

175

190

261

395

To recapitulate, Jason was a 4 year old boy who was diagnosed with acute lymphocytic leukemia (ALL) on the basis of his clinical presentation and blood and bone marrow tests that revealed large numbers of immature lymphocytes (see *Figure 7.1*). Fluorescence *in situ* hybridization of the abnormal cells (*Figure 7.10*) showed that these had a reciprocal translocation, t(12;21)(p13;q22.3) that created a *TEL-AML1* fusion gene, a known driver of leukemagenesis.

Jason was admitted to hospital and treated with induction chemotherapy including prednisolone, vincristine, daunomycin and L-asparaginase. He responded well to this treatment and progressed to consolidation treatment with methotrexate. He was then started on maintenance treatment of 6-mercaptopurine and methotrexate, which was planned to continue for 3 years. However, after treatment started he suffered severe neutropenic sepsis, an infection associated with a very low neutrophil count. This serious adverse drug reaction was a consequence of Jason's genotype at the thiopurine methyltransferase (*TPMT*) locus.

6-mercaptopurine and azathioprine are prodrugs that are widely used for treating ALL, and also in transplantation, and for treating inflammatory bowel disease and inflammatory arthritis. In the circulation, azathioprine is converted by non-enzymic hydrolysis to 6-mercaptopurine. When taken into cells 6-mercaptopurine is converted into a series of thioguanine derivatives which are powerful inhibitors of DNA and RNA synthesis. These are the active molecules, but they are highly toxic and precise dosage is critical. Thiopurine methyltransferase (*TPMT*) catalyzes a side reaction in the circulation that converts 6-mercaptopurine into the inactive 6-methylmercaptopurine. People with low-activity forms of *TPMT* over-produce the active but toxic thioguanine species (*Figure 10.7*).

In 1980 cases were first described of *TPMT* deficiency in red blood cells, and subsequent studies established that reduced red blood cell *TPMT* activity was associated with adverse

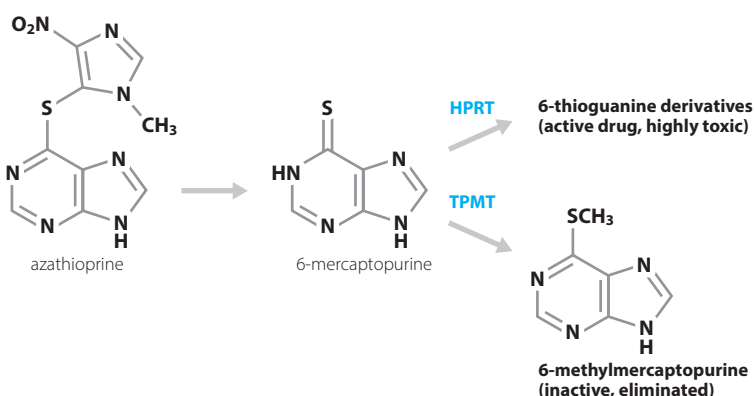


Figure 10.7 – Metabolism of azathioprine and 6-mercaptopurine.

HPRT, hypoxanthine phosphoribosyltransferase; TPMT, thiopurine methyltransferase. The conversion of azathioprine to 6-mercaptopurine is non-enzymic.

Table 10.2 – Common low-activity thiopurine S-methyltransferase alleles

Allele	Frequency in Caucasians
TPMT*2	0.5%
TPMT*3A	5%
TPMT*3C	0.5%

Note the nomenclature: in the pharmacogenetics literature specific alleles are often named by the locus, followed by an asterisk and a serial number. The variants are c.238G>C (TPMT*2); c.460G>A and c.719A>G (TPMT*3A) and c.719A>G (TPMT*3C). These are all mis-sense variants causing amino acid substitutions.

effects of thiopurine drugs, including azathioprine and 6-mercaptopurine. Three variant alleles account for 80–95% of intermediate or low activity cases (Table 10.2). About 10% of the UK population are heterozygous for a low activity allele, and 0.3% are homozygous. These people are much more sensitive than normal homozygotes to the effects of these powerful drugs. In low activity homozygotes, normal doses can cause life-threatening bone marrow toxicity and collapse of the hematopoietic system.

TPMT status can be tested either by measuring the enzyme activity or by DNA analysis of the gene. Prior to treating Jason with 6-mercaptopurine a blood sample had been taken for enzyme analysis. This had revealed a normal level of enzyme activity. With hindsight, the doctors noted that before he had the TPMT test he had been given a blood transfusion to correct his anemia. Because the TPMT enzyme test is performed on red blood cells, this may have provided an inaccurate measure of his TPMT status. A blood sample was therefore sent for TPMT genotyping. This revealed that he was homozygous for the TPMT*3A allele, predicting absent TPMT activity. He was given intravenous antibiotics and supportive treatment and recovered well. He was recommenced on the maintenance treatment with a reduced dose of 6-mercaptopurine and remained disease-free.

CASE 20 VLASI FAMILY

- Valon, 6-year-old boy with serious learning problems
- Small, microcephalic, blue eyes, fair skin and hair, eczema; hyperactive
- ? Phenylketonuria
- Testing for subsequent baby?

251 **262** 317 395 396

Testing Valon's blood and urine confirmed the diagnosis of phenylketonuria. This is normally due to inactivity of the enzyme phenylalanine hydroxylase (Figure 10.4). A small percentage of phenylketonuric babies have instead a genetic defect in the production or recycling of an essential cofactor, tetrahydrobiopterin (BH₄, see OMIM 261640). The laboratory work-up checks for these variant forms of PKU, and will often include DNA studies to define the mutations in the PAH gene.

In phenylketonuria, a block in the normal catabolic pathway for phenylalanine leads to an accumulation of phenylalanine in the blood and tissues (Figure 10.4). This is not cleared through the urine because amino acids are actively reabsorbed in the kidney tubules. Eventually the accumulation spills over into production of abnormal metabolites, phenylketones, which can pass into the urine, hence the name of the condition.

When the family came to Sydney, Flora was already 6 weeks pregnant. Having learned that Valon had a serious genetic disease, she and Adem were extremely worried about the risk of the new baby being affected. With all their other problems, they felt they could not cope with a second affected child, and they discussed terminating the pregnancy despite their general reservations about abortion. When their family doctor explained that even if the baby was affected (a 1 in 4 risk), it could be treated, they immediately asked for the treatment for Valon – but the doctor explained that the treatment was only effective if started very soon after birth (see Chapter 14). So they asked if the new baby could be tested as soon as it was born, and were assured that this would be done. Flora delivered a healthy girl, and mother and baby were discharged from hospital.

A few days later the midwife visited them at home. As well as checking on the health and progress of mother and baby, she pricked the baby's heel and collected a blood spot on to a special card (a Guthrie card). This is standard practice for every baby born in many countries, regardless of family history. The card was sent to a central laboratory and here the level of phenylalanine in the blood was measured. Although it would have

been easier to take the blood sample while the baby was still in the hospital, this was not done because, while it is *in utero*, a phenylketonuric baby's phenylalanine is cleared through the placenta by the mother (who is expected to be a phenotypically normal heterozygote; see Figure 10.8). Only after the placental connection is broken does the phenylalanine level in a phenylketonuric baby start to rise. It takes a few days for an elevated level to become readily apparent. The optimal time is day 5, though in countries without integrated healthcare systems, the best course may be to take blood at 24–48 hours, while the mother and baby are still in hospital. The test result showed that Flora's new baby had normal levels of phenylalanine in her blood. She might be a carrier (like her parents and 1 in 50 of the general population, see Chapter 9) or a normal homozygote; the important thing is that she did not have PKU.

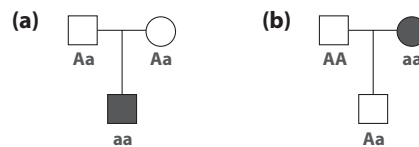


Figure 10.8 – While it is *in utero*, the level of phenylalanine in a fetus is determined by the mother's genotype and not its own.

(a) A phenylketonuric fetus develops normally *in utero* because the mother clears excess phenylalanine through the placenta. (b) Because the high level of phenylalanine in the maternal circulation crosses the placenta, the normal fetus of a phenylketonuric mother will be born severely brain-damaged and microcephalic unless the mother goes on a low phenylalanine diet throughout her pregnancy.

CASE 21 PORTILLO FAMILY

- Sickly boy, Pablo
- Family history of similar problems
- X-linked severe combined immunodeficiency
- Bone marrow transplantation
- Genetic cause defined
- Carrier tests for female relatives

252 263 286 395

Bone marrow transplantation is the treatment of choice for patients with severe immunodeficiencies such as Pablo. At first sight an immunodeficient patient would seem an ideal choice as a transplant recipient. Pablo lacked all T-cell function and therefore could not reject a graft. However, a problem with bone marrow transplants is graft-versus-host (GvH) disease. If the grafted bone marrow successfully reconstitutes an immune system, it will recognize the host tissue as foreign and initiate an immune response that could

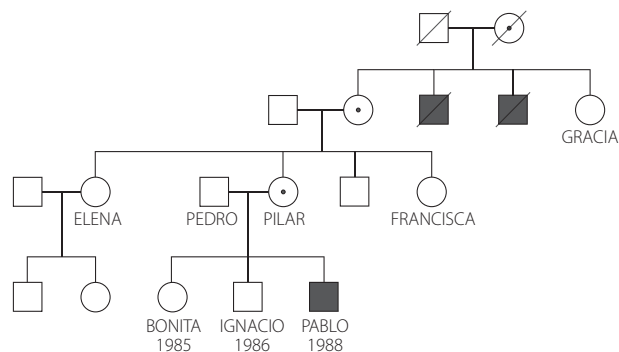


Figure 10.9 – Pedigree of severe combined immunodeficiency in the Portillo family.

The pattern shows that this is the X-linked form of this rare disease. Pablo's mother, grandmother and great-grandmother are obligate carriers. His sister, two aunts and cousin are at risk of being carriers.

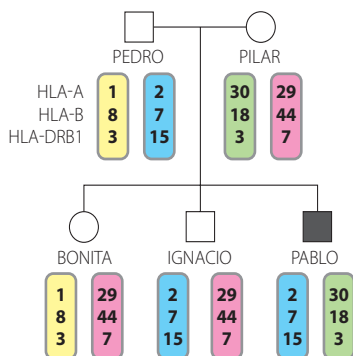


Figure 10.10 – Results of HLA typing in the Portillo family.

Because the *HLA-A*, *-B* and *-DR* loci are close together on chromosome 6, they are normally inherited as a block (haplotype).

be fatal. Good matching should reduce the risk of this. Pablo and other family members were typed for HLA-A, -B and -DR with the results shown in *Figure 10.10*.

Sibs have a 1 in 4 chance of sharing both MHC haplotypes. Unfortunately, neither of Pablo's sibs Ignacio or Bonita turned out to be a perfect match. His parents, of course, each shared one haplotype with him. Transplant donor registries were searched in the hope of finding a well-matched unrelated donor, but not surprisingly no perfect or even good match could be found. *Table 10.3* shows the frequencies in Spaniards from Murcia of each of the alleles and of the whole haplotypes carried by Portillo family members. Even though these are among the commonest haplotypes in that population, the chance a randomly selected member of the population would have the same two haplotypes as Pablo is only 1 in 1500.

Table 10.3 – Frequencies of Portillo family alleles and haplotypes in a Spanish population

Locus	Allele	Frequency
HLA-A	1	0.095
	2	0.230
	29	0.083
	30	0.107
HLA-B	7	0.056
	8	0.052
	18	0.071
	44	0.179
HLA-DRB1	3	0.123
	7	0.179
	15	0.052
Haplotype	Observed frequency	Calculated frequency
A1-B8-DR3	0.023	0.00061
A2-B7-DR15	0.019	0.00067
A29-B44-DR7	0.051	0.00266
A30-B18-DR3	0.035	0.00093

Note that HLA genes tend to occur in particular combinations. The observed frequency of each haplotype is much greater than the product of the individual allele frequencies (shown in the Calculated frequency column). This is an example of linkage disequilibrium. See *Chapter 13* for more discussion. Data from www.allelefrequencies.net; note that the nomenclature has been simplified.

Time passed, and eventually it was decided to act on his mother Pilar's offer to donate some of her own marrow. A new technique gave hope that GvH disease could be minimized. The problem is caused by T cells in the donor marrow, and in the 1980s techniques were developed to deplete human marrow of T cells. This made it, in principle, possible to restore immune function by bone marrow transplantation in patients with any form of SCID. Some T cells are inevitably present, so it is still preferable to match the transplant as well as possible, but mismatches are no longer a major problem. The

method had recently become available in her regional transplant center, and so bone marrow was taken from Pilar, depleted of T cells, and infused into Pablo. However, the results were disappointing. The child acquired only low levels of T-cell function. Evidently the marrow had engrafted poorly. Despite receiving a booster transplant, again from his mother, he succumbed to a cytomegalovirus infection and died at the age of 14 months. Experience suggests that the success rate is much lower in transplants performed after about 3.5 months of age. Buckley (2004) gives an authoritative and readable review of bone marrow transplantation in SCID.

Pablo died in 1989. At that time the gene causing the family disease had not been identified. The options open to his parents, if they wanted more children and wished for prenatal testing, were limited. They could opt for fetal sexing and terminate any pregnancy where the fetus was male. Given that in half those cases the fetus would have been unaffected, this was not acceptable to Pilar or Pedro. The geneticists could use genetic markers to try to identify whether a male fetus had inherited Pilar's normal or abnormal X chromosome, as described in *Chapter 8*. A DNA sample had been banked from Pablo to help with this, or with mutation analysis once the gene had been identified. The problem was that genetic mapping of X-SCID at that time had not provided an unambiguous localization. Probably the defective gene mapped to the proximal long arm, at Xq12–q13, but there was some question whether this was true of all cases. Thus there was a risk that an inappropriate genetic marker would be used, and a false negative result obtained. Pedro and Pilar were unwilling to take this risk, and in any case they already had two healthy children and decided not to have more; Pedro had a vasectomy.

In 1993 the faulty gene in X-SCID was identified as *IL2RG*. This encodes the gamma subunit that is part of the receptors for several different cytokines (interleukins 2, 4, 7, 9 and 15). Lack of cytokine signaling prevents development of T cells and NK cells; B cells are present but do not produce antibodies. The gene is quite small, consisting of 8 exons covering 4.2 kb at location Xq13. A variety of loss of function mutations have now been described in affected males. The DNA sample from Pablo Portillo was retrieved from the freezer and tested. A C>T substitution was found in exon 7 that converted the CGA codon for arginine 293 into a TGA stop codon (*Figure 10.11*).

Pilar's sisters Francisca and Elena were re-contacted and offered a definitive mutation test. Both accepted. PCR-amplifying and sequencing exon 7 of the gene showed that Elena was a carrier but Francisca was not. A note was made to contact Pilar's daughter Bonita and Elena's daughter when they were about 16 and could make an informed decision about carrier testing.

Years later another affected baby was born in a different branch of the family, and by that time there were more possibilities for treatment. This story will be taken up again in *Chapter 11*.

NORMAL SEQUENCE	CGGACGATGCCC	GAATTCCACCTGAAG
	R T M P	R I P T L K
MUTANT SEQUENCE	CGGACGATGCCC	TGAATTCCACCTGAAG
	R T M P	X

Figure 10.11 – The p.R293X mutation in the *IL2RG* gene that produced X-SCID in Pablo Portillo.

X designates a stop codon.

10.4. Going deeper...

Inborn errors of metabolism

Genotype–phenotype correlations

There is no neat one-to-one correspondence between enzymes and diseases. You cannot take a metabolic pathways chart and write in beside each reaction the disease that results from an inborn error in that step. To begin with, any disease that is the result of failure of a multi-step pathway could be produced by a block at any step in the pathway. Thus several different inborn errors might produce the same disease. This is similar to the way that defects in many different genes can produce inherited hearing loss or intellectual disability. Also, many enzymes are dispensable – perhaps lack of the reaction product has no obvious adverse effect, or maybe there are other ways of achieving the same function. Thus some enzyme defects would not produce a disease. *Figure 10.4* may give a contrary impression to all this, but this corner of metabolism was chosen precisely because it illustrated a particularly simple relation between biochemistry and disease.

Most inborn errors involve loss of function mutations, and as is usually the case with loss of function mutations, there is often extensive allelic heterogeneity (see *Chapter 6*). Because biochemists can measure enzyme activity quantitatively, inborn errors are a promising field for establishing genotype–phenotype correlations. We might expect a variant that causes a partial loss of activity of an enzyme to cause a milder disease than one causing total loss.

In many cases this expectation is fulfilled. However, the correlation between enzyme activity and phenotype is usually far from perfect. Variants of a sulfate transporter enzyme provide an example. High molecular weight sulfated polysaccharides are important components of connective tissue, and loss of function variants in the diastrophic dysplasia sulfate transporter (DTDST) enzyme cause skeletal dysplasias. Four different autosomal recessive skeletal dysplasia syndromes, which were distinguished on clinical grounds, turned out all to be caused by defects in DTDST. In ascending order of severity they are:

- multiple epiphyseal dysplasia 4 (MED4, OMIM 226900)
- diastrophic dysplasia (DTD, OMIM 222600)

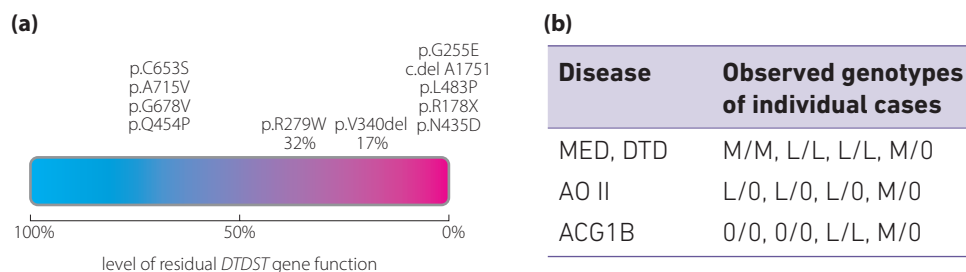


Figure 10.12 – Genotype–phenotype correlations in skeletal dysplasias caused by loss of function of the DTDST sulfate transporter.

(a) Level of enzymic activity of a series of reported variants. Variants were classified as having zero (O), low (L) or medium (M) activity. (b) Overall enzyme activity in a series of patients with clinical phenotypes of different severity. See text for key to the abbreviations. Data from Karniski (2001).

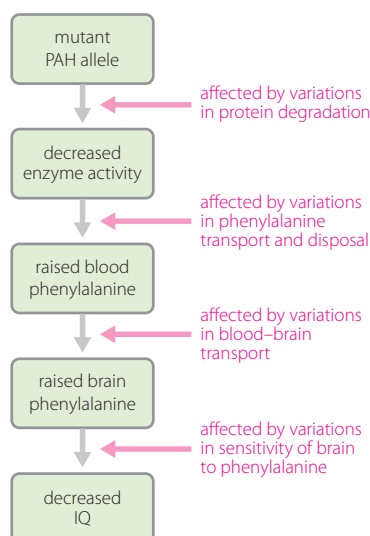


Figure 10.13 – In phenylketonuria many factors weaken the correlation between genotype and phenotype.

PAH, phenylalanine hydroxylase. This figure summarizes the arguments of Scriver and Waters (1999).

- atelosteogenesis type II (AOII, OMIM 256050)
- achondrogenesis type 1B (ACG1B, OMIM 600972).

Many different variants have been reported in the *DTDSD* gene. An interesting paper by Karniski (2001) explored the relationship between enzyme activity and clinical phenotype.

Enzyme activity was measured for a series of reported variant alleles (*Figure 10.12a*). Most patients are compound heterozygotes, and it is important to remember that what matters in physiology is the overall level of activity provided by the two alleles of the gene. Grouping the variants into zero (0), low (L) and medium (M) activity, Karniski reported the data shown in *Figure 10.12*.

The general conclusion is that genotype–phenotype correlations do exist, but they are usually rather loose. A paper by Scriver and Waters (1999) discusses why this might be so. The authors asked how far the IQ of an untreated phenylketonuric patient could be predicted from the enzymic activity of the phenylalanine hydroxylase variants they carried. The answer was, not very far, and *Figure 10.13* summarizes their arguments why this is so (their paper is recommended reading).

One gene – many enzymes?

The one gene – one enzyme hypothesis, though not true in all circumstances, is the essential guiding idea in biochemical genetics. How, then, can we explain diseases where a single gene defect results in defects in multiple enzymes? Two examples illustrate how defects in post-translational modification can cause loss of activity of multiple enzymes:

- **Mucopolidiosis II** (OMIM 252500) patients have disproportionate dwarfism, a coarse face and intellectual disability, similar to those with Hurler syndrome (OMIM 607014). Both diseases involve lysosomal enzyme defects, but whereas Hurler disease is caused by loss of function of one specific lysosomal enzyme, alpha-L-iduronidase, lysosomes in patients with ML II (also known as I-cell disease) have deficiencies of a whole range of enzymes. ML II is one of the family of carbohydrate-deficient glycoprotein diseases. Lysosomal enzymes synthesized in the cell cytoplasm require a specific attached carbohydrate as a signal that they should be transported to lysosomes. In ML II this signal is defective. There is almost complete absence of lysosomal targeting, and instead most lysosomal enzymes are secreted into the bloodstream. The signal consists of *N*-acetylglucosamine-1-phosphate attached to mannose sugars that are in turn attached to the polypeptide chains of many lysosomal enzymes. The underlying defect in ML II is in a single enzyme, the transferase needed to produce the signal molecule (*GNPTAB* gene).
- **Multiple sulfatase deficiency** (OMIM 272200) combines the features of six mendelian diseases that are each caused by deficiency of one specific sulfatase enzyme. mRNAs for each individual enzyme appear to be produced normally. This is again a defect in post-translational processing. All the affected enzymes have an unusual amino acid, formylglycine, as part of their active site. This is produced by modifying a cysteine residue in the protein after it has been synthesized (*Figure 10.14*). Multiple sulfatase deficiency is the result of loss of function mutations in the gene encoding sulfatase modifying factor 1 (SUMF1), the enzyme responsible for the modification.

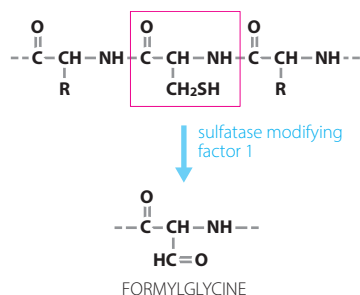


Figure 10.14 – The active site of several sulfatases requires post-translational modification of cysteine to formyl glycine by the SUMF enzyme.

Loss of SUMF function produces multiple sulfatase deficiency.

It is also worth noting that the whole concept of inborn errors of metabolism depends on a view of what constitutes 'normal' metabolism. *Boxes 10.5* and *10.6* describe conditions that would not generally be considered as inborn errors. Inability to synthesize vitamin C is a trait that all humans have, but that only manifests as disease in exceptional environments (*Box 10.5*). Lactose intolerance (*Box 10.6*) is a common mendelian phenotype that many Europeans might see as an inborn error of metabolism, widespread among non-European populations, but that is actually the normal ancestral state; it is the 'normal' ability of many European adults to tolerate fresh milk that is the derived (non-ancestral) state.

Inborn errors of metabolism are among the most potentially treatable genetic conditions. Many respond well to dietary manipulation or drug treatment (see *Table 14.2*). Because of this, in some countries newborn infants are routinely screened for a large number of different inborn errors (see *Table 12.4*).

Inability to make vitamin C – a universal inborn error in humans

If you lived on the diet that your dog or cat eats, you would develop scurvy because of lack of vitamin C. How do they stay healthy despite never eating lemons? It turns out that almost all animals have the enzymes necessary to synthesize ascorbic acid, and so are not dependent on an external supply. Exceptions to this include humans and other higher primates, guinea pigs, fruit-eating bats and the red-vented bulbul bird. All these species lack the enzyme L-gulonogamma-lactone oxidase (GULO), which catalyzes the last step of ascorbic acid biosynthesis. The human version of this gene, on chromosome 8p21, is a defective pseudogene with missing exons and other mutations, relative to the functional mouse *gulo* gene. Human cells transfected with the mouse *gulo* gene make their own ascorbic acid. Presumably the defective species all had such a fruit-rich diet that there was no selective pressure against *GULO* loss of function mutations.



Box figure 10.3 – Biosynthesis of ascorbic acid.

The final reaction is non-enzymic and occurs spontaneously. X indicates the metabolic block in humans.

BOX 10.5

Lactose intolerance – a common metabolic polymorphism

Most adults in Northern Europe have a dominant trait, hereditary persistence of intestinal lactase. This makes them able to tolerate a diet rich in milk. Conversely, most people in East Asia and in tropical and subtropical regions of the world shut off production of intestinal lactase in early childhood. It is quite common for people who lack intestinal lactase to suffer abdominal distension, pain, and diarrhea if they drink fresh milk. Dairy products such as cheese and yoghurt contain less lactose and cause fewer problems. Across the world the correlation with milk drinking is close – for example, certain pastoral tribes in Africa (e.g. the Bedouin, and the Beja people of Sudan) who drink fresh milk have high levels of persistence of intestinal lactase, whereas most African populations are non-persistent.

BOX 10.6

Persistence or non-persistence of lactase is a common polymorphism: both states are common among normal people. The ancestral state is undoubtedly non-persistence, as found in most mammals. Evidently there has been powerful selection for the persistence variant among populations who have taken up dairy farming. This must all have happened within the last 9000 years, making it one of the strongest selective changes in recent human history.

The causative DNA variant was hard to identify, but eventually turned out to be a C/T polymorphism in an enhancer 13 910 bp upstream of the start codon of the lactase gene on chromosome 2q21. In a survey of 236 individuals from 4 populations (Finnish, French, European American and African-American) every individual with one or more T alleles had persistent lactase, while every individual homozygous for the C allele had non-persistence. In Saudis and sub-Saharan Africans, however, different non-coding variants are responsible for lactose tolerance – an example of convergent evolution. See OMIM 223100 for more detail. These variants are typical of the sort of variants that are likely to surface as susceptibility factors for common diseases (*Chapter 13*). Mendelian diseases are usually the result of coding variants that cause major loss or gain of function of a gene, but here we have examples of changes that modify the timing of expression of a gene without affecting the integrity of the gene product.

Pharmacogenetics

Many genetically determined differences in pharmacokinetics and pharmacodynamics have been known for decades, but until recently personalized prescribing has been a topic for talk rather than action. There are several reasons for this slow uptake.

A cynic might argue that drug companies would not wish to develop a genetic test for drug efficacy if its only effect were to reduce the size of the market for their drug. They would be keener to identify the genotypes responsible for rare idiosyncratic adverse reactions – but that is a much more difficult task. Pharmaceutical companies now try to design drugs so that they are not metabolized by the P450 system or other highly variable enzymes. Thus the main targets for pharmacogenetic testing would be established drugs, most of which are now out of patent protection. The industry has little incentive to develop and market tests for drugs that are not bringing in much money.

One obstacle to personalized medicine concerns the logistics of incorporating genetic testing into routine clinical practice. It puts a delay between seeing the patient and prescribing a treatment. For some situations, for example with psychiatric drugs, a delay of a few days would be a price worth paying if the result were much more effective treatment. In many areas of oncology this is now standard practice (see below). In other areas personalized prescribing might really take off if there were a bedside dipstick test that allowed instant genotyping. An alternative in countries with integrated healthcare systems and electronic patient records might be to perform a single once-in-a-lifetime analysis of all the genes associated with major pharmacogenetic variation. Data from the analysis would become part of a person's standard medical record, available whenever there was a need to prescribe any drug.

The often poor genotype–phenotype correlations limit the potential value of genotype-driven prescribing. Even when there is reasonable understanding of the genetics, genotypes may not be strongly predictive of the optimum drug and dosage. The anti-coagulant warfarin is an important test case for the value of genotype-led prescribing.

Warfarin is very widely prescribed to people who have coronary artery disease or venous thrombosis, especially after surgery. The therapeutic window – the range of doses that are effective and not harmful – is narrow. Too little, and the patient gets no benefit; too much and there is a risk of serious, even fatal, bleeding. The safe but effective dose varies 20-fold between individuals, and adverse effects of inadequate or excessive doses are a major cause of emergency hospital admissions, second only to those caused by problems with insulin.

Warfarin targets vitamin K epoxide reductase (VKORC1), an enzyme that helps maintain the level of vitamin K, an essential clotting co-factor. Breakdown of the drug depends on several P450 enzymes, especially CYP2C9 (Figure 10.15). Individuals with low-activity variants of CYP2C9 and/or some variants of VKORC1 are at risk of serious bleeding episodes when given a standard dose of warfarin. However, many other factors also affect a person's response. These include the patient's age, the presence of other illnesses, the concurrent use of other drugs, and variation in other genes involved in handling the drug. Clinical trials have shown that genotype-led dosing can avoid some problems of the traditional trial and error procedure, but the overall benefit has not been sufficient to convince many physicians to adopt genotype-driven prescribing.

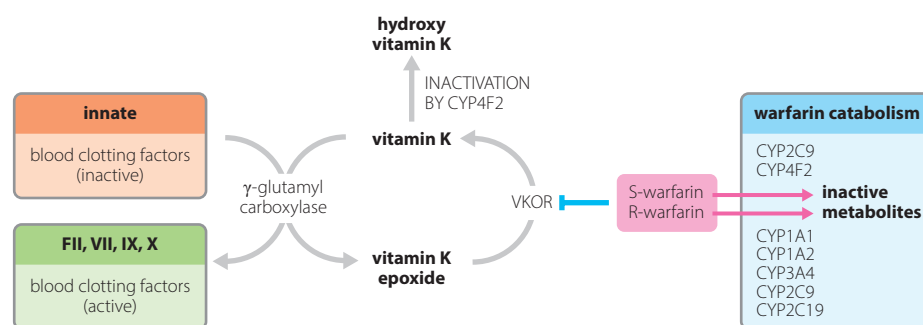


Figure 10.15 – Action and metabolism of warfarin.

Warfarin as usually prescribed is a mixture of two variants (stereoisomers), R-warfarin and S-warfarin. Both are active in reducing blood clotting, but the S-form is the more potent. They act by inhibiting VKOR, an enzyme necessary for recycling vitamin K, which is required to activate components of the blood clotting cascade. A number of different P450 enzymes (labeled CYP...) affect the efficacy of warfarin. Reproduced from *Human Molecular Genetics*, 5th edition (Strachan and Read, 2019) with permission from Garland Science/Taylor & Francis LLC.

Currently, oncology is the main area where stratified medicine has seen major successes. Traditional anticancer drugs simply target dividing cells. Not surprisingly, these drugs produce severe side effects. Newer drugs target specific signaling systems that drive the excessive cell proliferation of individual tumors. As we saw in *Chapter 7*, different tumors, even of the same histological type, show enormous heterogeneity of driver mutations, so genotyping prior to prescription is mandatory.

Recent years have seen a proliferation of targeted anticancer drugs. Some are small molecules, others are monoclonal antibodies (the other promising targeted agents, CAR-T (chimeric antigen receptor T-) cells are discussed in *Box 14.10*). The example of imatinib, a small molecule inhibitor of the BCR–ABL1 chimeric tyrosine kinase, was discussed in

Disease box 7. Trastuzumab (Herceptin) is an example of a monoclonal antibody. It targets breast cancers whose growth is driven by multiple copies of the *ERBB2* gene. A tumor biopsy must be checked for *ERBB2* over-expression before Herceptin is prescribed. *Table 10.4* shows examples.

These drugs are highly effective against tumors having the specific target genotype, and completely ineffective against similar tumors that lack the relevant variant. They are also extremely expensive. Thus in every case a biopsy must be genotyped to check for the variant. Unfortunately, tumors are rapidly evolving organs (see *Chapter 7*) and drug-resistant clones soon emerge. Combination therapy with several targeted drugs may be the way forward.

Cancer may be a special case, because drugs can target tumor-specific acquired variants that are not present in the normal cells of the patient. However, genetic studies raise the hope of splitting clinically defined conditions like schizophrenia into genetically defined subsets, and these may offer opportunities for a stratified medicine approach. An alternative promising approach to cancer therapy, immunotherapy, is discussed in *Box 14.10*.

Table 10.4 – Examples of oncogenic variants of genes that are targeted by specific drugs

Gene target	Oncogenic mechanism	Cancer type	Drug class
C-KIT	Activating tyrosine kinase domain mutation	Gastrointestinal stromal tumors	Small molecule tyrosine kinase inhibitor
B-RAF	Activating tyrosine kinase domain mutation	Melanoma	Small molecule tyrosine kinase inhibitor
HER-2	Amplification	Breast cancer	Monoclonal antibody to HER-2
EGFR	Activating tyrosine kinase domain mutation	Lung cancer	Small molecule tyrosine kinase inhibitor
EML4–ALK gene fusion	Gene fusion	Lung cancer	Small molecule tyrosine kinase inhibitor
ROS1 gene fusion	Gene fusion	Lung cancer	Small molecule tyrosine kinase inhibitor
RET gene fusion	Gene fusion	Thyroid cancer, lung cancer	Small molecule tyrosine kinase inhibitor
K-RAS	Activating tyrosine kinase mutation	Colorectal cancer	Negative predictor for cetuximab (monoclonal antibody treatment) targeting EGFR
BRCA1	Loss of function	Ovarian cancer	PARP inhibitor

Courtesy of Dr Fiona Blackhall, Christie Hospital, Manchester.

Immunogenetics

In *Section 10.2* we mentioned two challenging aspects of immunogenetics: how antigens are recognized as self or non-self, and how the effectively infinite diversity of antibody molecules is produced. The role of the MHC in recognition was described above; here we discuss the remarkable genetic mechanisms that underlie antibody diversity.

Antibodies are proteins (immunoglobulins, Ig) composed of heavy and light polypeptide chains. These can be assembled in various ways to produce the different classes of antibody – IgA, IgD, IgE, IgG and IgM. Each class has a specific heavy chain and a choice of κ or λ light chains. The five types of heavy chain are all encoded at the *IGH* locus on chromosome 14, and the light chains at the *IGK* or *IGL* loci on chromosomes 2 and 22, respectively. Each of these loci, however, contains a large pool of possible coding sequences (Figure 10.16). In the germ-line, and in all non-B cells, this arrangement is stable and the genes are not expressed. In B cells, however, the expressed immunoglobulin genes are the result of a series of DNA rearrangements. Mature immunoglobulin genes have a conventional multi-exon structure, with a single exon encoding the N-terminal variable region that carries the antigen binding site, and several exons encoding the constant region. However, the exon encoding the variable region is the product of highly variable DNA rearrangements during B-cell maturation.

As explained in the caption to Figure 10.16, the sequence encoding each variable region is made by joining a V (variable), J (joining) and, for heavy chains, D (diversity) segment, each chosen from a range of alternatives. Do not confuse this process with the splicing of exons and introns. Exons and introns are spliced in the RNA transcript by the spliceosome (see Chapter 3 and Disease box 10). Ig genes are spliced at the DNA level specifically in B lymphocytes by a special recombinase enzyme. The product is a multi-exon gene whose primary transcript is then spliced in the usual way.

The total diversity of antibodies depends several additional mechanisms.

- Unlike conventional genetic recombination, the specialized mechanism joining V, D and J segments adds or removes random small numbers of nucleotides at the

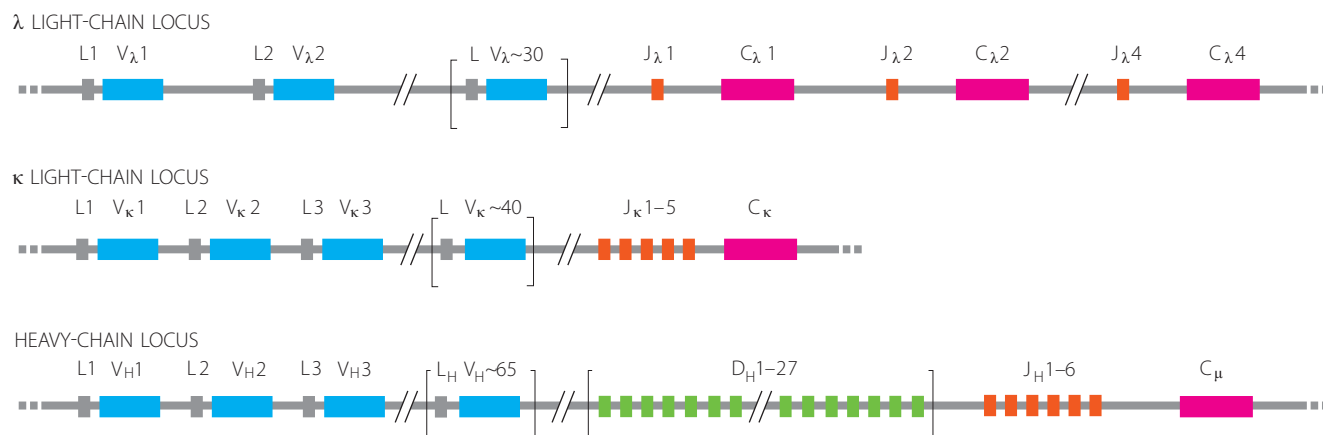


Figure 10.16 – Each type of immunoglobulin gene is assembled by DNA rearrangements that occur specifically in B lymphocytes.

In light chains one of 30–50 V regions is joined to one of 5–11 J regions, giving several hundred combinations. In heavy chains one of 27 D regions is joined to one of 6 J regions, and the new sequence is then joined to one of about 70 V regions, giving about 10 000 combinations. This combinatorial diversity is only one of the ways in which B cells are enabled to produce an almost infinite diversity of immunoglobulins. Reproduced from Murphy and Weaver (2016) with permission from Garland Science/ Taylor & Francis LLC.

junctions. Where the result produces a frameshift, the rearranged gene is non-functional, but in the one-third of cases where the reading frame is conserved, a whole extra layer of diversity is created.

- After gene rearrangement is complete, a process of somatic hypermutation introduces random point mutations at high frequency into the variable region of the genes in activated B cells.
- The different classes of heavy chain in IgA, IgD, IgE, IgG and IgM are made by yet another specialized recombination mechanism. Each rearranged IgH gene has at its 3' end sequences encoding each of the five types of constant region (only the C_μ is shown in *Figure 10.16*; the others are downstream of this). Only the C sequence closest to the VDJ exon is used. Class switching occurs by an intramolecular recombination event that physically excises one or more C sequences, so that a different one lies adjacent to the VDJ sequence.
- Finally, the combination of a heavy chain and a light chain in an antibody molecule provides yet another source of combinatorial diversity.

It remains only to add that a similar generator of diversity operates in T cells on the T cell receptor genes. For more detail of all of these processes, consult an immunology textbook, such as Murphy and Weaver (2016).

Disorders of the spliceosome

The spliceosome is the molecular machine that cuts the introns out of pre-mRNAs and splices together the exons (Matera and Wang, 2014). Like the ribosome, it is a large ribonucleoprotein complex, made up of non-coding RNAs and proteins that work together to perform its biochemical functions. Overall, five RNA species (snRNAs U1, U2, U4–U6) and around 170 different proteins are involved in its function. Interpreting cues for alternative splicing requires the spliceosome to perform a more flexible job than the ribosome, and to be more responsive to outside signals. Perhaps because of this, the spliceosome is a very dynamic structure compared to the ribosome, incorporating and discarding components and undergoing massive conformational shifts as the splicing reaction proceeds. Each of the snRNAs joins the spliceosome in the form of a preassembled ribonucleoprotein (snRNP) complex: first the U1 and U2 snRNPs, then the U4, U5 and U6 snRNPs together.

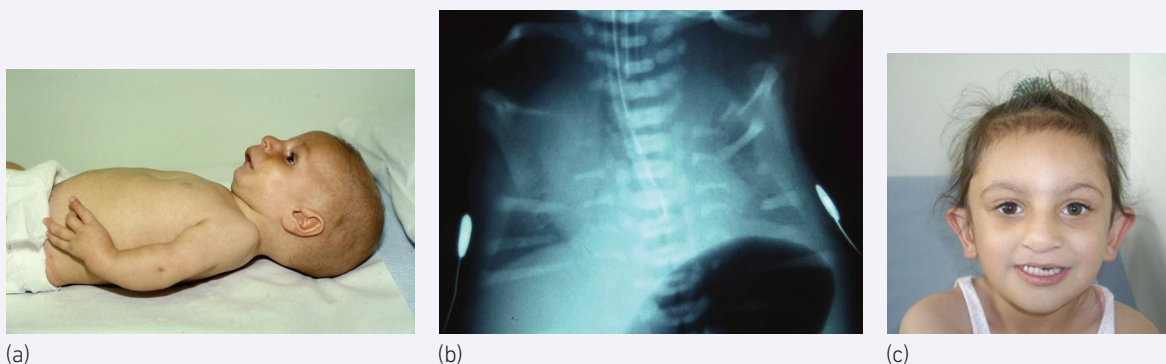
With so many working parts, it is not surprising that many things can go wrong. We are not concerned here with the very frequent variants in gene sequences that affect splicing of an individual pre-mRNA (see *Section 6.2*), but with defects in the splicing machinery itself. One might expect any defect in the core spliceosome to be an embryonic lethal, but in fact a surprising number of specific syndromes have turned out to be caused by mutations affecting the biogenesis or function of the spliceosome.

- **Retinitis pigmentosa** – variants in genes encoding several different spliceosomal proteins have been identified in families with autosomal dominant retinitis pigmentosa. PRPF3, PRPF8, PRPF31 and SNRNP200 (BRR2) proteins are all involved in the U4/U5/U6 tri-snRNP complex, one of the main components of the functional spliceosome. It is postulated that retinal neurons have an exceptionally high requirement for splicing factors, making them perhaps particularly sensitive to sub-optimal function of the spliceosome.
- **Facial dysostoses** – in a number of distinct syndromes there are defects in facial bone structures, likely caused by abnormal migration of neural crest cells to the pharyngeal arches and the face. In some cases there are also limb and other abnormalities. Some of these syndromes are caused by variants in spliceosomal proteins. About half of all patients with

Nager syndrome (OMIM 154400) have heterozygous mutations affecting the SF3B4 protein, a component of the U2 snRNP. The likely pathogenic mechanism is haploinsufficiency. A related but more severe condition, Guion–Almeida mandibulofacial dysostosis (OMIM 610536) is caused by haploinsufficiency for EFTUD2, a GTPase protein that is part of the U5 snRNP. The gene encoding another component of the U5 snRNP, TXNL4A, is mutated in Burn–McKeown syndrome (OMIM 608572). All patients described to date have been compound heterozygotes for a null mutation and a deletion in the promoter. The latter is a low-frequency population polymorphism that reduces, but does not completely abolish, expression of the gene. Thus patients retain a low level of TXNL4A function – maybe complete loss would be lethal. Finally, cerebrocostomandibular syndrome (OMIM 117650) is caused by variants in the *SNRNPB* gene that encodes a regulatory protein common to all the snRNP modules of the spliceosome. As with Burn–McKeown syndrome, the effect is to reduce but not abolish SNRNPB function. In this case the mechanism involves uprating production of a splice isoform that contains a premature termination codon, and so produces no protein.

- **An snRNA mutation** – mutations in the five major snRNAs have not been reported. However, around 700 introns in the human genome are spliced out by an alternative spliceosome in which snRNAs U11 and U12 replace U1 and U2, and U4atac and U6atac replace U4 and U6. U5 is common to both machines. Unlike the major snRNAs, these minor snRNAs are each transcribed from a single gene. Variants in the U4atac gene that interfere with assembly of the U4atac/U6atac/U5 complex are the cause of the recessive Taybi–Linder syndrome (microcephalic osteodysplastic primordial dwarfism Type 1, OMIM 210710).
- **Spinal muscular atrophy** – in Section 6.2 we saw how deficiency of the SMN protein causes spinal muscular atrophy (SMA, OMIM 253300) due to degeneration of the anterior horn cells of the spinal cord. This protein has a plethora of functions in the cell, but a major one is assisting the assembly of the five snRNP RNA–protein complexes that come together to form the spliceosome (reviewed by Matera and Wang, 2014). Complete absence of SMN protein is lethal; SMA patients rely on the residual levels of SMN protein provided by the poorly functional *SMN2* gene (see Section 6.2).

Spliceosome action is far from uniform – great flexibility is needed in order to respond to the many tissue-specific and context-specific signals governing alternative splicing. In addition, the various defects described here all reduce, rather than abolish, function of the spliceosome. This perhaps explains how defects in this basic cellular mechanism might lead to the tissue-restricted pathologies of the conditions described here.



Box figure 10.3 – Disorders due to defects in the spliceosome.

(a) Nager syndrome. Note micrognathia and radial aplasia with absent thumb. (b) Chest X-ray of an infant with severe cerebrocostomandibular syndrome. Note gaps and missing ribs. (c) A child with Burn–McKeown syndrome. Note hypertelorism, right unilateral cleft lip/palate, thin upper lip and prominent ears.

10.5. References

- Buckley RH** (2004) Molecular defects in human severe combined immunodeficiency disease and approaches to immune reconstitution. *Ann. Rev. Immunol.* **22**: 625–655.
- Evans WE and McLeod HL** (2003) Pharmacogenomics: drug disposition, drug targets and side effects. *New Engl. J. Med.* **348**: 538–549.
- Guengerich PF** (2008) Cytochrome p450 and chemical toxicology. *Chem. Res. Toxicol.* **21**: 70–83.
- Karniski LP** (2001) Mutations in the diastrophic dysplasia sulfate transporter (DTDST) gene: correlation between sulfate transporter activity and chondrodysplasia phenotype. *Hum. Molec. Genet.* **10**: 1485–1490.
- Matera AG and Wang Z** (2014) A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15**: 108–121.
- McLeod HL and Siva C** (2002) The thiopurine S-methyltransferase gene locus – implications for clinical pharmacogenomics. *Pharmacogenomics*. **3**: 89–98.
- Murphy K and Weaver C** (2016) *Janeway's Immunobiology*, 9th edn. Garland Science, New York. An early edition is available on the NCBI Bookshelf www.ncbi.nlm.nih.gov/books.
- Pirmohamed M, James S, Meakin S, et al.** (2004) Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Br. Med. J.* **329**: 15–19.
- Scriber C and Waters PJ** (1999) Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet.* **15**: 267–272.
- Weinshilboum R** (2003) Inheritance and drug response. *New Engl. J. Med.* **348**: 529–537.

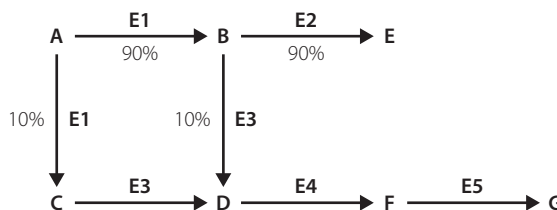
Useful websites

Pharmacogenomics Knowledge Base: www.pharmgkb.org

Personalised medicines: hopes and realities – a report by the Royal Society 2005: www.royalsoc.ac.uk/displaypagedoc.asp?id=17570

10.6. Self assessment questions

- (1) The diagram shows the biosynthetic pathways leading from precursor A to products E and G. Enzymes E1–E5 catalyze the reactions. E1 converts 90% of A to B and 10% to C. 90% of B is normally converted to E by E2, but 10% is converted to D by E3. Which enzyme(s) might be deficient in a condition marked by (a) deficiency of E and G, (b) deficiency of G only, and (c) deficiency of E together with a raised amount of G?



- (2) Describe ways in which a mendelian condition might be the result of loss of function of several enzymes if, (a) each of several unrelated patients shows a loss of function of just one of the enzymes, different in each patient and, (b) if each patient shows loss of function all of the enzymes.
- (3) Acute intermittent porphyria (OMIM 176000) is a dominant condition caused by mutations in the porphobilinogen deaminase (*PBGD*) gene, yet it is estimated that 80% of people heterozygous for a pathogenic *PBGD* mutation go through their lives completely unaware of the fact, and never suffer an episode of the disease. Discuss possible reasons for this low penetrance.
- (4) Discuss the case for regarding scurvy as a genetic disease. Can similar arguments be applied to other examples?
- (5) Arthur, Bridget and their three children Charles, Daniel and Eliza were typed for the HLA-A, -B and -DR loci. The results (using a simplified nomenclature) were:

Arthur	A3,23;	B7,27;	DR3,4
Bridget	A2,23;	B15,27;	DR3,4
Charles	A3,23;	B15,27;	DR3,4
Daniel	A2,23;	B7,27;	DR3,4
Eliza	A2,3;	B27;	DR4

Assuming there is no recombination, so that a person's A, B and DR alleles are passed on as an unbroken haplotype, work out the haplotypes in this family.

- (6) What features of the gene locus or mutational spectrum in a disease would lead you to suspect that gene conversion might be a major mutational mechanism?

[Hints on questions 2, 3 and 5 are provided in the *Guidance* section at the back of the book.]

11

How are genes regulated?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Define epigenetic, X-inactivation, imprinting, uniparental disomy, CpG island
- Give a general overview of the roles of DNA methylation, histone modification and chromatin conformation in regulating transcription
- Describe X-inactivation and its consequences for carriers of X-linked recessive conditions and balanced X;autosome translocations
- Describe what genetic imprinting is, how it happens and its possible evolutionary rationale
- Explain how DNA is methylated, how the methylation patterns can be heritable, and how they can be studied in the laboratory
- Describe how chromosomal events or point mutations involving imprinted genes can result in clinical conditions
- Give examples of pedigree patterns and sporadic syndromes dependent on imprinting

11.1. Case studies

CASE 22 QIAN FAMILY

- Girl, Kai, aged 2 years
- Developmental delay, seizures
- ? Angelman syndrome

277 287 395

Chu-Li and Chan are a hard-working couple who have an import business. When their first child Kai was born Chu-Li's mother moved from Hong Kong to help look after the baby. She had looked after several grandchildren before but even she found it hard to get Kai to feed well or to settle and sleep. Kai was slow to gain weight and achieve her developmental milestones. She appeared very jittery although she seemed happy and laughed a lot. By two years of age she wasn't showing any signs of talking and an appointment was made with a pediatrician. However, before this could happen she had a seizure and was admitted to the children's ward.

It was clear to the pediatrician that there were major problems with Kai's development. She noted that Kai, who had just learned to walk, did so with rather stiff legs held well apart. She had lots of jerky movements especially with her arms. She laughed a lot and tended to protrude her tongue and dribble. Her records showed that her



Figure 11.1 – A 10 year old girl with Angelman syndrome.

She has a mutation in the *UBE3A* gene. Photo. courtesy of Dr Jill Clayton-Smith, St Mary's Hospital, Manchester.

head circumference at birth was normal but it was now just below the 3rd centile. She arranged for Kai to have an EEG and the result confirmed her clinical suspicions. There were generalized EEG changes with runs of high-amplitude delta activity with intermittent spike and slow wave discharges. The pediatrician told the parents that she was sure Kai had a condition called Angelman syndrome. The family were shocked because they said no one else in the family had problems. They wanted to know what had caused the condition and if it might happen again. The pediatrician said that she needed to refer them to the genetics clinic because there were different ways that Angelman syndrome could occur and sometimes another child in a family could be affected.

CASE 23 ROGERS FAMILY

- Baby boy, Robert, born to older parents
- Normal 46,XY karyotype and pregnancy tests
- Severely hypotonic
- ? Prader–Willi syndrome



Figure 11.2 – A baby with Prader–Willi syndrome.

Note the marked hypotonia.

278

287

395

Ralph and Rowena Rogers had both been married before and each had a child by their first partners. Although Rowena was 38 years old when they married, they decided to try for a baby and were delighted when, after a few months, a pregnancy was confirmed. They wanted as many tests as possible to ensure a healthy baby. An amniocentesis test (see *Chapter 14*) showed a normal male karyotype and scans didn't show any problems. Rowena did mention that she didn't feel many fetal movements but put that down to the fact that she was very busy. Labor occurred on her due date and was a rather long affair, but a baby boy was born weighing 3.2 kg. The family decided to call him Robert. When he was put to the breast he didn't try to suck at all and the doctor noticed he was very floppy. The doctor mentioned his concerns to Ralph and Rowena and said he was going to ask for an urgent chromosome test to rule out Down syndrome. Rowena reminded him that the amniocentesis test had shown a normal chromosome pattern and so the doctor decided to wait and see how Robert progressed. He needed feeding by tube because he couldn't suck well and had marked truncal hypotonia. The pediatrician talked to the geneticist on the telephone and described Robert's problems. Suspecting Prader–Willi syndrome the geneticist said he would see the family urgently in his next clinic. In the genetics department the Rogers family were seen at the same clinic as the **Qian family (Case 22)**, which was an informative coincidence, especially for the medical student who attended because he was doing a special module in genetics.

11.2. Science toolkit

So far in this book we have concentrated on what genes are, how they work and how variant gene sequences can cause diseases or other variant phenotypes. Here we turn to the question of how gene expression is regulated. Understanding gene regulation is central to understanding what makes us human and how we function. Consider two observations:

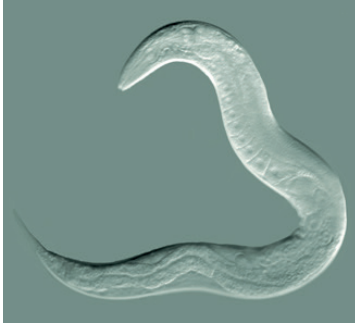


Figure 11.3 – *Caenorhabditis elegans* nematode worm.

Image by Bob Goldstein, UNC Chapel Hill, and reproduced under CC-BY-SA 3.0 Licence.

- Our body is made up of many hundreds of different types of cells. Brain cells, skin cells, liver cells and so on, all differ widely in their structure and function, yet they are all derived by repeated mitosis from the same original fertilized egg cell, and so they all have the same genomes. They differ because they use different readouts of the same genomes. All development, from egg to adult, depends on differential switching on and off of genes.
- The complex anatomy and functioning of the adult human body depends on an estimated 3.7×10^{13} cells of many different types. In *Section 3.4* we mentioned the 1 mm long nematode worm *Caenorhabditis elegans* (see *Figure 11.3*). This animal consists of just 959 or 1031 cells (depending on the sex). It is widely studied as one of the very simplest multicellular animals. Yet it has much the same number of protein-coding genes as we do (20 454 in *H. sapiens* versus 20 191 in *C. elegans*, according to ENSEMBL (as at 7 December 2019). Clearly our much greater complexity is not because we have more genes, but because we use the same genes in a more complex way. Thus sophisticated gene regulation is central to who we are.

Every step of gene expression is regulated: transcription, translation, post-translational modification, and localization of the gene product. A full description of all the mechanisms involved is far beyond the scope of this book, and so here we will focus on regulation of transcription: how cells decide which genes to transcribe and which remain silent. This is the major locus of on/off gene regulation. Although it is not an absolute distinction, in general and with many exceptions, controls acting later in the process tend to act more like dimmer switches, fine-tuning the level of expression.

Thus far we have seen how an RNA polymerase molecule attaches to the promoter, upstream of the coding sequence of a gene, and moves along the template strand synthesizing an RNA copy of the sense strand. It is easy to forget that our genome consists, not of naked DNA but of chromatin. As described in Chapter 2 (*Figure 2.17*), DNA is normally wrapped round an octamer of histone proteins to make nucleosomes, which then stack together to form the basic structure of chromatin. Nucleosomes contain four main types of histone, normally two molecules each of H2A, H2B, H3 and H4. This arrangement does not simply package the DNA in a static way; it dynamically governs what information in the DNA can reach the outside world. Chromatin-based mechanisms turn the invariant genome into a flexible, responsive and highly differentiated system. A subset of these mechanisms are collectively called **epigenetic** – literally, above genetics. Different authors tend to use the word ‘epigenetics’ in a narrower or broader sense (see box below).

Two definitions of epigenetics

- Mitotically and/or meiotically heritable changes in gene function that are not caused by changes in the DNA sequence (a frequently used narrow definition that insists on inheritance through mitosis, e.g. Holliday, 1994).
- The study of molecules and mechanisms that can perpetuate alternative gene activity states in the context of the same DNA sequence (a much broader definition, Cavalli and Heard 2019).

Epigenetic mechanisms are central to development. In the cascades leading to terminally differentiated cells, stem cells and their intermediate progeny acquire successively more differentiated patterns of gene expression that define the cell type, and these are conserved from mother cell to daughter cells by epigenetic mechanisms. However, many of the same mechanisms are involved in regulating gene expression without being transmitted through mitosis. Thus when in the final section of this chapter we describe the chromatin-based mechanisms that control transcription, we will follow Cavalli and Heard (2019) and not make a fundamental distinction between those cases where there is evidence that they are epigenetic in the narrow sense, and those that are just part of a cell's responses to its environment.

X-inactivation: a classic epigenetic process

The fact that people can be entirely normal while having either one (46,XY) or two (46,XX) X chromosomes requires explanation. Having an extra or missing copy of an autosome that contained that many genes would be lethal. The explanation lies in a mechanism of dosage compensation called **X-inactivation** or **lyonization** (named after its discoverer, Dr Mary Lyon).

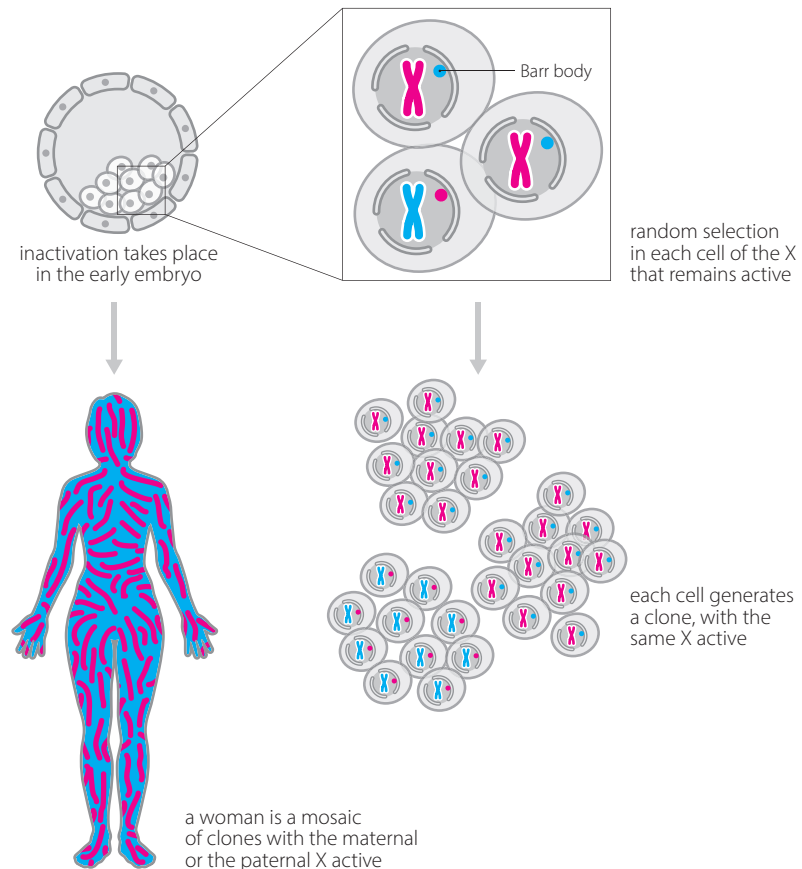
Early in the life of every human (and other mammalian) embryo, at the blastocyst stage, each cell somehow counts its number of X chromosomes. All Xs except one are then permanently inactivated by epigenetic mechanisms. All the DNA of the inactivated chromosome remains intact, but most of the genes are not expressed (some X-linked genes do escape the general inactivation; why and how they do so is the subject of current research). Which X remains active is a random choice in each cell. However, once the choice is made, it is remembered by all the daughters of that cell (*Figure 11.4*).

X-inactivation is an epigenetic process, depending on a change in chromatin conformation that is heritable through mitosis from cell to daughter cell. On a female karyotype the two X chromosomes are indistinguishable (see, for example, *Figures 2.10, 2.14 and 2.15*), but this is because the chromosomes are being seen during mitosis when all chromosomes are highly condensed and largely inactive. On completion of mitosis, while the other chromosomes of a female cell decondense, the inactive X remains condensed. It can sometimes be seen as a spot of densely staining chromatin at the edge of the nucleus, known as the Barr body. Counting Barr bodies was used in the past as a way of counting the number of X chromosomes (for sex tests on athletes, for example). Normal females and XXY males have one Barr body per cell; normal males and 45,X females have none, and 47,XXX females have two. However, X-inactivation is not heritable through a pedigree. In the female germ-line the inactive X is reactivated, and meiosis picks one X chromosome at random to go into the egg. Both X chromosomes in a 46,XX fertilized egg are active, and it is random which one will later be inactivated in any particular cell of the conceptus.

The 2.6 Mb of sequence immediately adjacent to the tip of the X chromosome short arm has special properties. A homologous sequence is present at the short arm tip of the Y chromosome; the two pair end-to-end in meiosis (see *Figure 2.7*) and have an obligatory crossover in this region. Genes in this region escape X-inactivation. Men and women each have two active copies of these genes, and the pattern of inheritance of variants appears autosomal. For this reason the region is called the pseudoautosomal region. There is another small pseudoautosomal region, 300 kb long, at the tip of the long arm,

Figure 11.4 – X-inactivation.

A normal woman is a mosaic of clones in which either the X chromosome from her mother or from her father is active; the other X is epigenetically inactivated.



but this does not usually pair or cross over with its counterpart on the Y chromosome in meiosis.

Normal females have roughly half their cells with the maternal, and half with the paternal X active, subject to the normal statistical variation. This has implications for a female carrier of any X-linked condition. Assuming random X-inactivation, she will be a mosaic of clones in which either the normal X or the mutation-bearing X chromosome is active. The consequences depend on what the gene product does, and where the relevant cells are.

- Where the gene product is diffusible, there is an averaging effect. Carriers of X-linked hemophilia have half the normal level of the affected clotting factor (subject to the usual individual variation). The clotting time is noticeably increased above normal, but the blood still clots sufficiently well to avoid disease.
- Where the gene product is fixed, there are patches of normal and affected tissue. The size of the patches depends on whether the tissue consists of many small clones or a few large ones, and on how much cell mixing occurred during development of that particular tissue. The patches may be demonstrable, for example, in anhidrotic ectodermal dysplasia (OMIM 305100), where the affected skin lacks sweat glands. A female carrier of this condition has patches of skin lacking sweat glands interspersed with patches having normal glands. If her skin

is painted with iodine and she then exercises enough to get a bit hot and sticky, starch powder will stick to the skin patches with sweat glands and form the dark starch-iodine color, but not to the patches lacking sweat glands. The pattern of skin clones will be revealed. They follow Blaschko's lines, named after the German physician who first described the pattern (see *Disease box 6*).

- If a woman is heterozygous for an X-linked variant that prevents the production of a particular type of cell, all her cells of that type must have her normal X active and the variant X inactive. Thus she would show completely skewed X-inactivation in those cells, but the normal 50:50 pattern of X-inactivation in all her other cells. The **Portillo family (Case 21)** provides an example, and this is discussed in the following section.

A second cause of skewed X-inactivation is a balanced X;autosome translocation (*Figure 11.5*). Usually, like most other balanced translocations, this will have no phenotypic effect (see the case of **Ellen Elliot, Case 5**) but in some cases there is an effect. Her translocated X chromosome is the only active X chromosome in every cell of her body, and so any mutated gene on it will affect her phenotype just as it would affect a male. Equally, if the translocation breakpoint happens to disrupt a gene, she will show the full phenotype associated with loss of function of that gene in males. For example, about two dozen unrelated women are known worldwide who have severe Duchenne muscular dystrophy despite having no family history of the disease. Each of these women has an X;autosome translocation. Each translocation involves a different autosomal breakpoint, but in each case the break on the X chromosome is at Xp21 and disrupts the dystrophin gene.

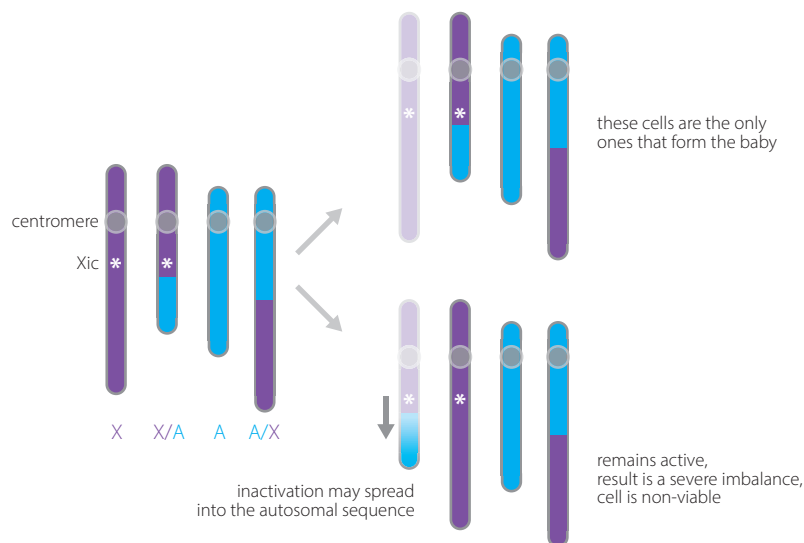


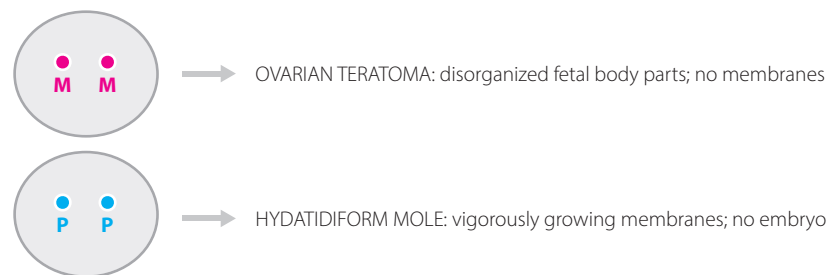
Figure 11.5 – Consequences of X-inactivation in a carrier of a balanced X;autosome translocation.

The inactivation process starts at the X-inactivation center (Xic, asterisk) on the proximal long arm and spreads across the chromosome. Cells that inactivate the translocated X suffer a fatal imbalance because of inability to inactivate the detached portion of the translocated X chromosome. Only cells that inactivated the intact X can contribute to forming the person.

Imprinting – why you need a mother and a father

For a heterozygous person, the parental origin of each allele is not normally relevant when thinking about their phenotype. However, some observations suggest that there are functional differences between the maternal and paternal components of somebody's genome. Occasional accidents start embryonic development in 46,XX cells that have either two maternal or two paternal genomes. Despite being ostensibly chromosomally normal, such cells always develop very abnormally, and quite differently from each other (Figure 11.6). Experiments in mice demonstrate the same effects. Evidently there is some difference between maternal and paternal genomes, and normal development requires one of each.

Figure 11.6 – Conceptuses that have two maternal or two paternal genomes do not develop normally.



More refined experiments in mice allowed the role of each individual chromosome to be studied. Ingenious manipulations allow mice to be generated that have correct chromosome numbers, but both homologs of one particular pair are derived from just one of the parents. This is called **uniparental disomy** (UPD). For some chromosomes the resulting mice are normal, but for others they are abnormal, and the particular abnormalities seen depend on whether the mice have two maternal or two paternal copies of the chromosome in question. Rare human cases of UPD were also discovered by chance. Later it became apparent that certain human syndromes could be caused by UPD, as described below.

These and other observations suggest that there are human (and mouse) genes that behave differently depending on their parental origin. They must carry some sort of *imprint* that marks their origin. It is important to remember that genes are not intrinsically paternal or maternal. If a man passes on to his child an imprinted gene that he inherited from his mother, he received it with a maternal imprint but passes it on with a paternal imprint. Thus imprinting must be reversible, so that it can be erased and re-imposed with each generation (Figure 11.7). Imprinting is an epigenetic phenomenon: the expression of imprinted genes is modified, and the modification persists through all the cell divisions that lead from a fertilized egg to an adult person, but the DNA sequence is not changed.

These observations fit into our general understanding of the crucial role of epigenetics in development. The whole of human development can be seen as an epigenetic process, because differentiating cells acquire new epigenetic identities that they pass on to their progeny. Sperm and egg cells have their own very different specific epigenetic identities. In the fertilized egg these identifying epigenetic marks must be stripped away, leaving an epigenetic clean slate for all subsequent development. In Section 11.4 we will discuss what these epigenetic marks are, how they can be added or removed and how they work to regulate gene expression. Against this general background, a few genes retain some of

Figure 11.7 – Addition and removal of tissue-specific epigenetic marks through human development. Most marks from sperm and egg are removed in the early zygote, leaving an epigenetic clean slate to allow subsequent development, but a few are not removed. These are the parent-specific imprints. Imprinting is epigenetic and reversible.

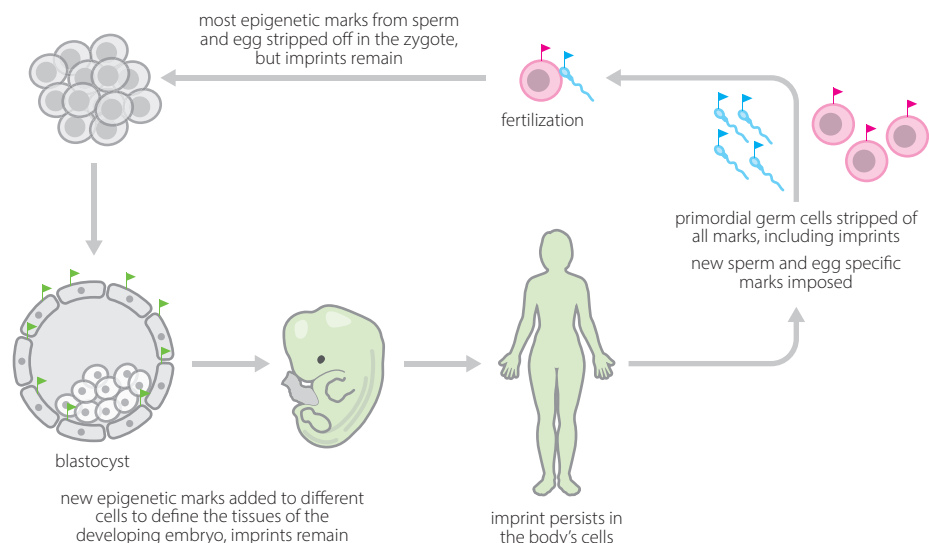
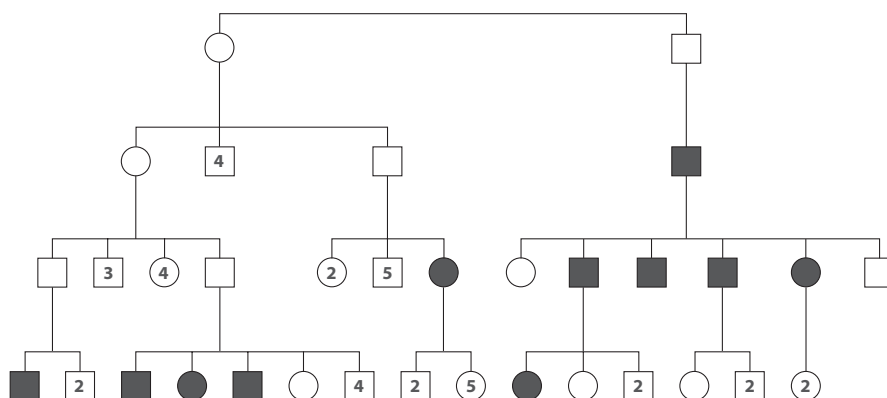


Figure 11.8 – Paragangliomas (hereditary glomus body tumors, OMIM 168000) are caused by mutations in the *SDHD* gene. The gene is imprinted and only expressed when inherited from the father. Thus offspring of a mutation-carrying mother are unaffected, whether or not they inherit the variant – but they can still pass it on. Family reported by Heutink *et al.* (1992), *Hum. Molec. Genet.* **1**: 7–10. To save space, where somebody has had several unaffected children a single symbol is used, with the number of children marked in the symbol.



11.3. Investigations of patients

CASE 4 DAVIES FAMILY

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy
- X-linked recessive inheritance
- Problems of testing dystrophin gene
- Exon 44–48 deletion identified by MLPA
- Molecular pathology
- Implications of X-inactivation

4

11

68

98

156

285

315

395

MLPA was used to define the pathogenic deletion in Martin, and to show that of his two sisters; Lisa was a carrier of the deletion while Jessica was not (see *Figure 4.11*). Martin's mother and maternal grandmother are obligate carriers of this X-linked disease. Because of X-inactivation they will be a mosaic of clones, some of which have the paternal X active, carrying a functional copy of the dystrophin gene, while others have the deletion-bearing maternal X active. The picture is complicated by the fact that muscle cells are multinucleate, being formed by fusion of myoblasts. Staining a muscle biopsy with a dystrophin antibody shows a patchy distribution of dystrophin in individual muscle cells of heterozygous females, reflecting the random nature of X-inactivation (*Figure 11.9*). Most carriers show biochemical evidence of subclinical muscle damage in the form of elevated levels of the muscle enzyme creatine kinase (CK) in their serum. As mentioned in *Chapter 4*, the CK level can be used to give an estimate of a woman's risk of being a carrier, but the estimate is rarely sufficiently close to 0% or 100% to be a useful guide for reproductive decisions. Occasional carriers have significant muscle weakness, presumably because by bad luck they inactivated mainly the normal X in their muscle cells. They are known as **manifesting heterozygotes**.

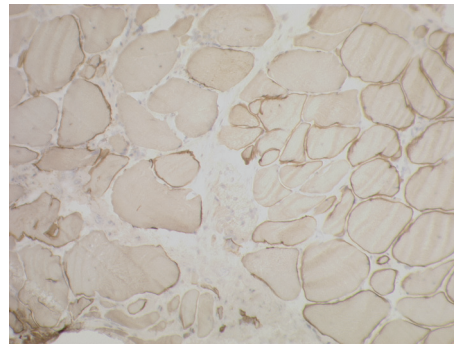


Figure 11.9 – A muscle biopsy from a female carrier of Duchenne muscular dystrophy stained with an antibody against dystrophin (brown).

Note the patchy distribution of staining around the outer membranes of cells (compare with the sections from an affected boy and a normal control in *Figure 1.4*). Photo. courtesy of Dr Richard Charlton, Newcastle upon Tyne.

CASE 9 INGRAM FAMILY

- Isabel, 10 years old with small stature and possibly delayed puberty
- ? Turner syndrome
- 45,X karyotype
- Risk of Y-chromosome DNA
- PCR test for Y sequences negative
- Questions around X-inactivation
- Possibilities for therapy

26

42

70

103

285

395

Isabel Ingram has a single X chromosome, 45,X and has Turner syndrome. Given that the X-inactivation mechanism exists in order to allow normal development in people with differing numbers of X chromosomes, one might ask why there is anything wrong with her. The answer is probably that not all genes on the X chromosome are subject to X-inactivation. An investigation of transcription levels (Carrel and Willard, 2005) showed surprising deviations from the conventional picture of blanket X-inactivation. Even outside the pseudoautosomal regions, about 15% of X-linked genes escaped inactivation partially or totally, and a further 10% showed different degrees of inactivation among the inactive X chromosomes in different cells.

Some X-linked genes that escape inactivation have counterparts on the Y chromosome, and these would be expected to have lower expression levels in Turner women than in normal men or women. Genes without functional Y-linked counterparts will have levels of expression in Turner females similar to those in normal males – but maybe the higher expression of these non-inactivated genes in XX females is responsible for some of the differences between normal males and females. Intriguingly, it has been claimed that Turner women whose X is of maternal origin may have behavioral problems of the sort that are mainly seen in boys (whose X chromosome, of course, is of maternal origin), while Turner women with a paternal X are free of such problems. If true, this would be evidence of imprinting.

CASE 21 PORTILLO FAMILY

- Sickly boy, Pablo
- Family history of similar problems
- X-linked severe combined immunodeficiency
- Bone marrow transplantation
- Genetic cause defined
- Carrier tests for female relatives
- Implications of X-inactivation
- Possibilities for therapy

252

263

286

395

Severe combined immunodeficiency can be either an autosomal recessive or an X-linked condition. At first it may seem surprising that an immunodeficiency should be X-linked – one might expect immunodeficiencies to be caused by mutations in the immunoglobulin genes, which are located on chromosomes 2, 14 and 22. But production of antibodies requires not simply intact structural genes for those proteins, but also properly functioning B cells, whose successful development must require numerous other genetically controlled steps. In *Chapter 10* we saw some of the complicated processing that is needed to enable these cells to produce an effectively infinite repertoire of antibody molecules. In combined immunodeficiency there is an absence of T cells as well as a functional defect in the B cells that produce antibodies. Its cause must be a failure of a much earlier stage of cell differentiation.

When the pedigree (*Figure 10.9*) was taken it seemed clear that baby Pablo had the X-linked form of SCID. If the disease had been the autosomal form, any relatives who were carriers would not be at risk of having affected children unless their partner was also a carrier. Because autosomal SCID is very rare, this risk is low provided they marry outside the family. However, for the Portillo family, now we know the condition is X-linked we can see that Pilar's sisters, aunt and daughter are at substantial risk, although for her brother and other son the risk is negligibly low.

At the time the family came to attention, it was not known what gene defect was responsible for X-SCID. However, it was possible to do carrier testing by looking at the pattern of X-inactivation in at-risk females. Female carriers of X-SCID like Pablo's mother Pilar have clones of cells in which the active X carries the pathogenic variant (subsequently identified as in the *IL2RG* gene, see *Figure 10.10*). As described above, any of those cells that were destined to differentiate into lymphocytes would be unable to do so. All the lymphocytes she does make would be derived from cells in which the normal X was active. Thus the lymphocytes of such a woman would show completely skewed X-inactivation, while all her other cells would show a normal roughly 50:50 pattern. Before the causative gene was identified this was sometimes used as a carrier test.

CASE 22 QIAN FAMILY277 **287** 395**CASE 23 ROGERS FAMILY**278 **287** 395**Qian family**

- Girl, Kai, aged 2 years
- Developmental delay, seizures
- ? Angelman syndrome
- Causes and genetic tests
- Possibilities for therapy

Rogers family

- Baby boy, Robert, born to older parents
- Normal 46,XY karyotype and pregnancy tests
- Severely hypotonic
- ? Prader–Willi syndrome
- Causes and genetic tests
- Possibilities for therapy

These two cases are considered together because although the phenotypes of PWS and AS are completely different, the causes of both conditions have a great deal in common. Three-quarters of cases of each condition are caused by a microdeletion of chromosome 15q11–q13. It was natural to suppose that since the two syndromes are so different, the deletions causing them must also be different at the molecular level – but they are not. The deletions are caused by recombination between misaligned repeats – the same mechanism that we saw in Williams syndrome (*Disease box 2*). The same 6 Mb stretch of DNA is normally deleted in each condition.

The deletion is sometimes just visible under the microscope but it can readily be checked using FISH or MLPA (see *Section 4.2*). In the present two cases a FISH probe for 15q11 identified a deletion in Kai Qian, confirming her diagnosis of Angelman syndrome. However, no deletion was seen in Robert Rogers, and so the provisional diagnosis of PWS still needed to be confirmed.

The breakthrough in understanding these conditions came when it was realized that the difference between them was due, not to different deletions, but to different parental origins of the deleted chromosome. In PWS it is always the paternal chromosome 15 that is deleted, while in AS it is always the maternal copy. Thus imprinted genes at 15q11–q13 lie at the heart of the pathology. PWS is always caused by lack of a paternally imprinted copy of the 15q11–q13 region, but this can arise in various ways other than by deletion (*Table 11.1*). In AS, but not PWS, some cases are due to a point mutation in an imprinted gene.

Table 11.1 – Causes of Prader–Willi syndrome and Angelman syndrome

Cause	PWS	AS
Del15(q11–q13)	75–80% (paternal)	70–75% (maternal)
Uniparental disomy	20–25% (maternal)	1% (paternal)
Point mutation	–	10% (<i>UBE3A</i>)
Imprinting error	1%	4%

In about 10% of cases with clinically diagnosed Angelman syndrome, none of the listed causes are found; in many an alternative diagnosis is discovered by whole exome sequencing.

The next step was to check Robert's DNA for uniparental disomy (UPD). Between 20 and 25% of PWS patients have two intact copies of chromosome 15, but both inherited from the mother. The effect is the same as the more frequent deletion: the patient lacks a paternal copy of the region. UPD can be detected by looking at the inheritance patterns of polymorphisms in the DNA of chromosome 15. A series of microsatellite markers (see *Box 8.1*) from chromosome 15 were genotyped in Robert and his parents. With the first marker (*Figure 11.10a*) it happened that the particular alleles in the three samples were such that we can't work out which of Robert's two alleles came from which parent. Remember that these are non-pathogenic polymorphisms that have no role in causing PWS or any other disease, so it is purely down to chance as to which alleles a person happens to have. The result does show that Robert has two copies of this particular sequence, and since this

is a sequence from within the PWS critical region, it confirms the FISH test showing that he does not have a deletion. The second marker (*Figure 11.10b*) is more informative. We see that Robert has no paternal alleles (correct paternity was confirmed using unlinked markers). By itself this result is compatible with either a deletion or UPD. Because we already know there is no deletion, it demonstrates UPD. This is isodisomy, in which he has two copies of the same maternal chromosome. Microsatellite alleles occasionally mutate, so it is prudent to confirm the finding using a second independent microsatellite.

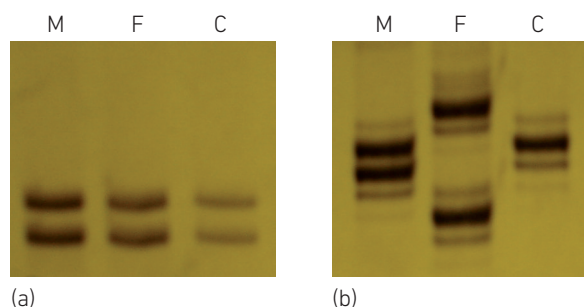


Figure 11.10 – Typing the Roberts family for two microsatellite markers from the PWS critical region.

(a) The genotypes show that Robert has two copies of the sequence, but do not identify the parental origin. (b) Robert lacks paternal alleles of this marker which, together with the result from the first marker, shows that he has maternal UPD. M, mother; F, father; C, child.

The phenotype of PWS is identical in patients with UPD or those with the more common paternal deletion, showing that PWS is caused by lack of a paternal copy of the 15q11–q13 sequence, and that a double dose of the maternal copy or lack of paternal copies of genes elsewhere on chromosome 15 have no additional effect. A few cases of non-deletion Angelman syndrome also are caused by UPD – in that case, having two paternal and no maternal copy of chromosome 15.

The origins of uniparental disomy

The UPD explains why Robert has PWS – but how does UPD arise? When the first example was reported (a child with cystic fibrosis in 1988) it was supposed that an egg that happened through non-disjunction to contain two copies of the relevant chromosome had, by extraordinary good luck, been fertilized by a sperm that happened by non-disjunction to lack any copy of that chromosome. If such a lucky coincidence were its sole cause, UPD would be vanishingly rare. In fact, though uncommon, it is seen far too frequently to be explained by such ultra-rare coincidences. A much more likely origin is through **trisomy rescue** (*Figure 11.11*).

We know that every possible trisomy occurs at conception, but nearly all are incompatible with survival to term, and miscarry spontaneously. But as *Figure 11.11* shows, a chance non-disjunction in an early mitotic division of a trisomic conceptus might generate one cell with a normal chromosome count. If that happened early enough in embryogenesis, that one cell might be able to develop into a complete baby. Assuming the mitotic non-disjunction is random, one time in three the result would be UPD. Supporting this view, UPD is much more frequent among PWS than Angelman patients (*Table 11.1*). We know

that the non-disjunctions that produce trisomies usually occur in the maternal meiosis, so we would expect most trisomic conceptuses to have two maternal and one paternal contribution. Therefore trisomy rescue is far more likely to generate maternal UPD than paternal UPD. It may be significant that Robert's mother Rowena was 38 years old when Robert was conceived.

Trisomy rescue produces **heterodisomy**: the two affected chromosomes are copies of the two different homologs in the parent. A possible alternative way of developing UPD is if a conceptus that, through nondisjunction, is monosomic for chromosome 15 was rescued by duplication of the monosomic chromosome. That would produce **isodisomy**, in which the two copies of the relevant chromosome are identical. Evidently that was what happened to produce Robert Rogers (*Figure 11.10b*).

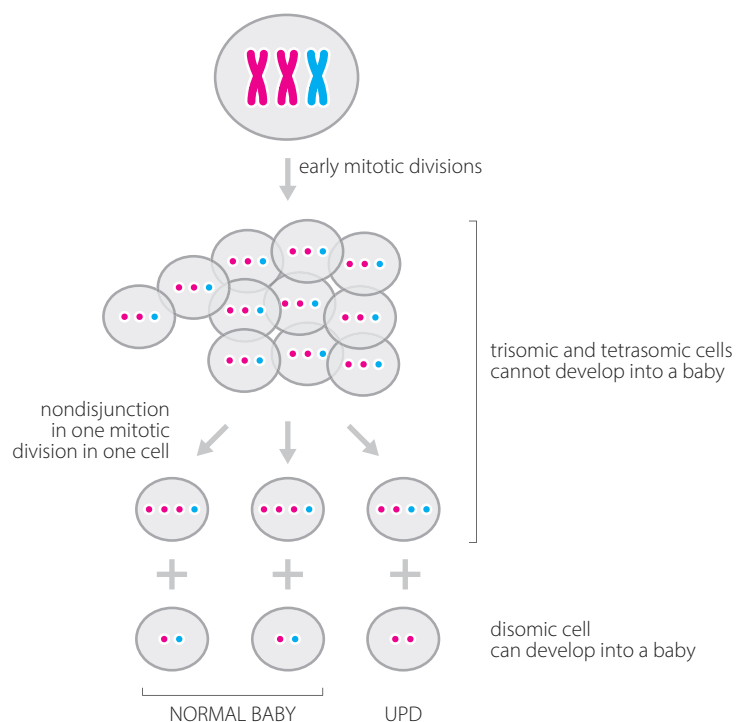


Figure 11.11 – Origin of uniparental disomy through trisomy rescue.

The initial conceptus is trisomic. A non-disjunction event in one mitotic division generates one disomic cell. If this happens very early in development that one cell is able to form the whole baby. By chance, one time in three the two copies will both be from the same parent.

Molecular pathology of PWS and Angelman syndrome

It turned out that the PWS/AS critical region contains a number of imprinted genes, some of which are expressed only from the paternal chromosome while others have the opposite pattern (*Figure 11.12*). It appears that Angelman syndrome is entirely caused by lack of a functional maternal copy of the *UBE3A* gene. As *Table 11.1* showed, a few Angelman syndrome patients have a loss of function variant of this gene as their only molecular abnormality. Interestingly, this gene shows imprinted expression in brain but

not in other tissues – thus imprinting can be tissue-specific. Identifying such cases is important because if this is not a *de novo* mutation there is a significant risk of another child inheriting the variant.

No point mutations of a single gene have been found in PWS, but studies of patients with rare small deletions have implicated the *SNHG14* transcript. This is a huge paternal-specific transcript, at least 460 kb long, that contains at least 140 exons. Only the first ten exons encode protein; the remaining exons are all non-coding. However, many small nucleolar RNAs (snoRNAs) are made from the RNA of *introns* from this region. These snoRNAs are required for modifying bases in ribosomal and other functional non-coding RNAs. It appears that PWS is caused by a deficiency of snoRNAs that, in turn, affects the biogenesis of ribosomes and maybe other functional RNAs, and so affects expression of other, unrelated, genes.

In a few otherwise unexplained cases of either syndrome, there seems to be a fault in the imprinting mechanism. Imprinted gene clusters always contain differentially methylated regions (DMRs), where the pattern of DNA methylation (see Section 11.4) is different on the maternal and paternal chromosomes. These are thought to be the actual effectors of the differentially imprinted gene expression. In these rare cases marker studies (as in Figure 11.10) show that chromosomes from both parents are present and complete, but the methylation pattern at the DMR shows that both carry the same parental imprint. Evidently something has gone wrong with the imprinting mechanism, so that either the paternal chromosome carries a maternal imprint, causing PWS, or vice versa, causing Angelman syndrome. These rare cases are examples of **epimutations** – mutations that change the epigenetics but not the DNA sequence. They provide valuable research material for scientists investigating the imprinting process.

This rather daunting complexity is also seen at a number of other imprinted loci (Table 11.2). For example, on chromosome 11p15 there are two imprinted domains, a growth-promoting domain (DMR1) normally expressed from the paternal chromosome, and a growth-suppressing domain (DMR2) normally expressed from the maternal chromosome. Beckwith–Wiedemann syndrome, with fetal overgrowth, is caused by various changes that enhance expression of DMR1 genes or repress expression of DMR2 genes. Silver–Russell syndrome, with intrauterine growth retardation, results from opposite changes.

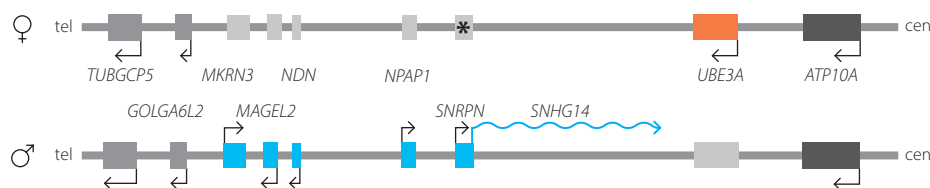


Figure 11.12 – Genes and transcripts in the Prader–Willi / Angelman syndrome critical region.

Genes in blue are expressed only from the paternal chromosome, those in orange only from the maternal chromosome, while those in dark gray are expressed from both. Genes in light gray are not expressed on the relevant chromosome. The wavy blue line shows the large paternal-specific SNHG14 transcript. The asterisk marks the differentially methylated region that controls the imprinting. Reproduced from Strachan and Read (2019; *Human Molecular Genetics*, 5th edition) with permission from Garland Science/Taylor & Francis LLC.

Table 11.2 – Some diseases that involve imprinting-related mechanisms

Syndrome	OMIM number	Chromosomal location	Affected gene(s)
Prader–Willi	176270	15q11–q13	<i>SNRPN</i>
Angelman	105830	15q11–q13	<i>UBE3A</i>
Beckwith–Wiedemann	130650	11p15.5	<i>IGF2, KCNQ10T1</i>
Silver–Russell	180860	7p11.2 11p15.5	<i>GRB10?</i> <i>H19, CDKN1C</i>
Temple	616222	14q32	<i>DLK1 / MEG3</i>
Kagami–Ogata	608149	14q32	<i>DLK1 / MEG3</i>
Pseudohypoparathyroidism 1A	103580	20q13.2	<i>GNAS1</i>
Transient neonatal diabetes mellitus	601410	6q24	<i>PLAGL1</i>

See OMIM and Soellner *et al.* (2017) for more detail. The University of Otago maintains a useful website giving extensive detail (<http://igc.otago.ac.nz/home.html>).

In addition to phenotypes caused by misregulation at a single imprinted locus, some patients have multilocus imprinting disturbance, where several loci are affected. A number of different causative genes have been identified; the non-functional variants in some cases are in the affected patient, but in others are in the mother. They seem to be involved either in writing epigenetic marks in egg cells or protecting them from erasure in the zygote.

What is the purpose of imprinting?

Several theories have been proposed to explain why imprinting should have evolved (see Wilkins and Haig, 2003). It is specific to placental mammals. The leading theory suggests that, seen in ‘selfish gene’ terms, parents have conflicting biological interests. A father’s genes are best propagated by his having plenty of children. If his partner dies from exhaustion, he can always hope to find another woman from the next cave. A woman’s genes are best propagated if she takes care of herself and doesn’t devote too many of her resources to any one child. Thus paternal genes promote fetal growth, even at the expense of the mother. This fits with the observations on hydatidiform moles: the paternal genes promote proliferation of the placenta and membranes, which extract nutrients from the mother.

11.4. Going deeper...

Every stage of gene expression – transcription, translation, post-translational modification of the gene product and sending it to its final destination – is subject to sophisticated and complex regulation. As described at the start of this chapter, we focus here on control of transcription. This fundamental layer of regulation governs the selective readout of the genome that defines cell identity and programs development. Disordered control of transcription is the basis of many clinical conditions, including the whole of cancer.

Controls on transcription operate at every level from the single gene to the overall genome. The best understood, and arguably most clinically relevant, parts operate over

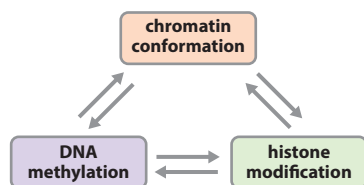


Figure 11.13 – Three pillars of local transcription regulation.

There is not a linear hierarchy of control; the three mechanisms all interact.

scales of a few kilobases – a small scale compared to the size of the whole genome. Broadly speaking, these controls involve the three interlinked mechanisms shown in *Figure 11.13*. Higher-level controls involve large-scale chromatin structures within the nucleus. These are currently poorly understood and are the subject of intensive research; they are briefly described in *Box 11.2*.

- **DNA methylation** – DNA methyltransferase enzymes attach methyl ($-\text{CH}_3$) groups to certain cytosine residues.
- **Histone modification** – a multitude of enzymes attach various groups including methyl, acetyl ($-\text{COCH}_3$) and phosphate groups to specific amino acid residues of histones in nucleosomes.
- **Chromatin conformation** – to interact with regulatory molecules the DNA must be accessible and not locked away in tightly packed chromatin. A number of different mechanisms control local and larger-scale chromatin conformation. Large ATP-driven multiprotein machines (chromatin remodeling complexes) move nucleosomes along the DNA to expose or occlude regulatory elements such as promoters and enhancers. **Small RNA molecules** can also play a role with their ability to target specific sequences in the genome.

Regulation often works in a combinatorial way: a series of relatively weak effects combine and interact to produce a strong effect. That allows more flexible and sensitive regulation than if everything depended on a single strong interaction. Thus, although individual interactions correlate with an overall effect, it is often not possible to draw a linear path from a single prime cause to the final effect. In so far as any prime causes can be identified, they are most likely to be binding of transcription factors – but these usually require chromatin changes to make their target sequences accessible.

Of the three mechanisms, only DNA methylation is currently the subject of clinical diagnostic tests. We also clearly understand how patterns of DNA methylation can be transmitted from mother cell to daughter cells, making it an epigenetic mechanism in the strict sense. However, it is worth noting that some of the best studied model organisms such as *Drosophila* flies and *Caenorhabditis* worms scarcely methylate their DNA at all, despite clearly being able to run epigenetic developmental programs. Thus DNA methylation cannot be the sole epigenetic mechanism – it is just the one we understand the best.

DNA methylation

In humans and other mammals methylation normally takes the form of adding methyl groups to the 5-position of cytosines to form 5-methyl cytosine (5MeC). DNA methylation does not affect base-pairing, so it does not alter the basic genetic message (*Figure 11.14*), nor does it in itself alter gene expression. But the methyl groups, sitting on the outside of the double helix, act as binding sites for methyl DNA-binding proteins, and these are part of regulatory systems that effect changes in gene expression. Thus the pattern of methylation along the genome can be ‘written’ by DNA methyltransferases and ‘read’ by methyl DNA-binding proteins, without affecting the primary gene sequence.

Methylation is almost entirely restricted to cytosines that lie immediately 5’ of guanines in so-called CpG dinucleotides (the ‘p’ represents the phosphate group that links the two together). Each 5’-CpG-3’ in one strand of the double helix is partnered by a 5’-CpG-3’ in

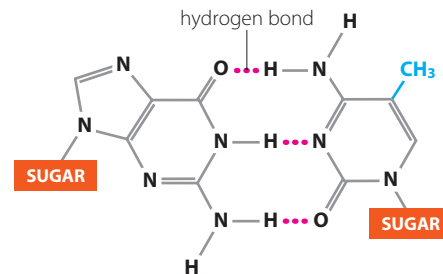


Figure 11.14 – 5-methyl cytosine base-pairs with guanine in exactly the same way as unmodified cytosine.

Thus DNA methylation does not alter the primary genetic message.

the opposite strand (remember, the strands are antiparallel). Not every CpG is methylated, but if a CpG in one strand is methylated, so is its partner CpG on the opposite strand, so that both strands carry the same pattern of methylated and unmethylated CpGs. This is the result of the action of a maintenance methyltransferase enzyme, DNMT1. When DNA is replicated, all CpG sequences on the newly synthesized strand are initially unmethylated. However, DNMT1 subsequently methylates any CpG on the newly synthesized strand that lies opposite a methylated CpG on the parental strand (*Figure 11.15*). By this process the pattern of methylation is inherited from mother cell to daughter cell.

Though heritable, the pattern of methylation is not fixed during the life of a cell. It differs according to the type of cell and its current metabolic state. Methyl groups can be added by *de novo* DNA methyltransferase enzymes, and removed through the action of the three Tet ('ten-eleven translocation') enzymes. Environmental factors affect the patterns and degrees of DNA methylation, as does the passage of time. Cancer and normal aging affect methylation. The degree of methylation of a selected set of 513 CpG sequences is the best current measure of biological (as distinct from chronological) age – and, rather alarmingly, is claimed to predict a person's remaining lifespan, at least in a statistical sense (Levine *et al.*, 2018).

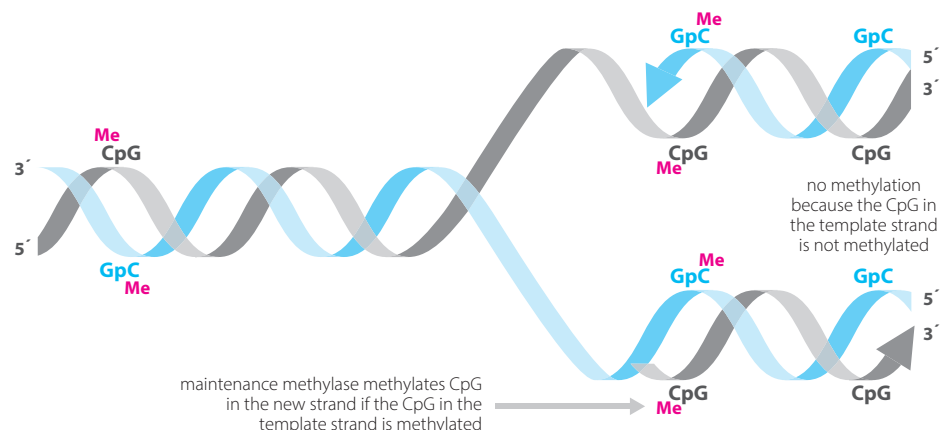


Figure 11.15 – The role of 5-methyl cytosine in epigenetics.

When the DNA is replicated a maintenance methyltransferase specifically methylates CpG sequences in the newly synthesized strand that lie opposite a methylated CpG in the parental strand. As a result, patterns of CpG methylation are heritable.

Studying DNA methylation

Methylation patterns in DNA are more difficult to study than sequence changes. As shown in *Figure 11.14*, 5MeC base-pairs with G just like unmethylated cytosine does. The hybridization properties of a DNA strand are unaffected by its methylation status, so hybridization-based tests cannot be used to investigate methylation. PCR and sequencing both depend on making copies of the sequence under investigation. Copies made *in vitro* (as distinct from copies made in cells that have the maintenance methylation machinery) will be unmethylated. Thus none of the methods described in *Chapters 4* and *5* can be used to check the methylation pattern of a DNA sequence. The main general method for studying DNA methylation is **bisulfite sequencing** (*Figure 11.16*).

When DNA is treated with sodium bisulfite under carefully controlled conditions, cytosine is converted by deamination into uracil, but 5MeC is resistant to the reagent. Uracil is not a natural base in DNA, but it base-pairs with adenine in just the same way as thymine does. If a bisulfite-treated template is copied by PCR or used in a sequencing reaction, each U in the original DNA is represented by T in the copies. Thus every C that was unmethylated in the original sequence appears as a T, while methylated Cs remain as C. Comparing the untreated and bisulfite-treated sequence reveals the pattern of methylation.

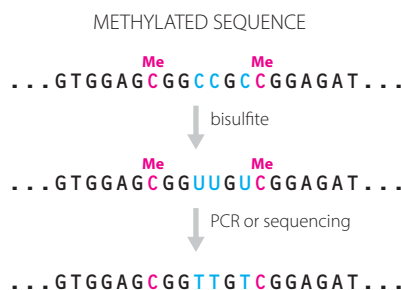


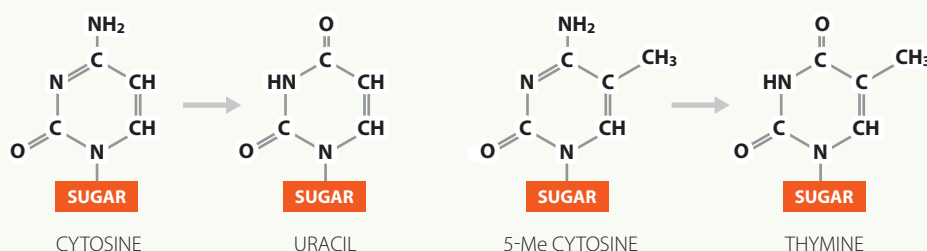
Figure 11.16 – Using sodium bisulfite treatment to identify methylated cytosines.

Unmethylated cytosines are converted to uracil, which appears as thymine when the product is PCR-amplified or sequenced. Methylated cytosines are unchanged. Comparing the sequence before and after bisulfite treatment identifies methylated Cs.

Other ways of checking the pattern of methylation of a DNA sample include using specific anti-MeC antibodies to immunoprecipitate methylated DNA fragments, or digesting the DNA with certain restriction enzymes whose recognition site includes a CpG sequence but that will cut the site only if it is unmethylated. For example, the restriction enzyme *HpaII* cuts CCGG sequences, but will not cut C^{Me} CGG. Other enzymes are unaffected by methylation – for example, *MspI* cuts CCGG regardless of methylation. This difference can be exploited in various ways to reveal whether or not a particular CpG is methylated. Most simply, PCR primers are designed to flank the CpG in question, and the DNA is digested with *HpaII* before amplification. If the site is unmethylated the template will have been cleaved and no PCR product will be obtained.

DNA methylation and CpG islands

We have seen how sodium bisulfite converts cytosine to uracil by deamination (removal of the amino group). But cytosine in DNA also has a tendency to deaminate spontaneously. It is estimated that in every cell 100 cytosine bases lose their amino group every day. Cells have an enzyme that recognizes uracil in DNA and repairs the damage by replacing uracil with cytosine. 5MeC also deaminates spontaneously. Deamination of 5MeC produces thymine (*Box figure 11.1*). This is a natural component of DNA, so the change is not obvious and is not always repaired. Thus methylated CpG dinucleotides have a natural tendency to mutate to TpG. A review of the mutation databases that have been established for many diseases clearly shows that CpG sequences are hotspots for mutation. The *IL2RG* mutation in **Pablo Portillo (Case 21)**, see *Figure 10.11* is one example among many.



Box figure 11.1 – Deamination of cytosine produces uracil, an unnatural base in DNA, but deamination of 5-methyl cytosine produces thymine.

The mutability of CpG sequences has had evolutionary consequences. Of the bases in the human genome 41% are C or G, so we might expect 4.2% (0.205×0.205) of all dinucleotides to be CpG. The observed frequency is one-fifth of this. Bulk human DNA is highly depleted of CpG sequences – they have been methylated and over evolutionary time converted to TpG by deamination. However, scattered around the genome are about 27 000 so-called **CpG islands** (the exact number depends on how an island is defined). These are stretches of DNA, normally 1 kb or less in length, where there has been no loss of CpG sequences. These sequences are somehow protected from being methylated.

About 60% of human genes have a CpG island in or near the promoter region. The regulation of transcription may be different in genes that have CpG islands from those that do not. CpG islands normally remain unmethylated, regardless of whether the associated gene is active or not. They become abnormally methylated in some cancer cells, which silences the gene (see *Chapter 7*) but not in normal cells. Promoters that do not have CpG islands nevertheless do contain individual CpG dinucleotides. Reversible methylation of these, and probably also of CpG dinucleotides at the boundaries of CpG islands, are important in gene regulation.

Histone modifications

Covalent modification of the protruding N-terminal tails of histone molecules in nucleosomes are the second major contributor to local control of transcription. Histone modifications (*Figure 11.17*) can be seen in terms of writers, readers and editors.

- **Writers** add methyl, acetyl, phosphate or other groups to histones. They include large families of histone acetyltransferases (HATs), methyltransferases, kinases, etc. Each enzyme is specific for a particular amino acid residue of a specific histone.

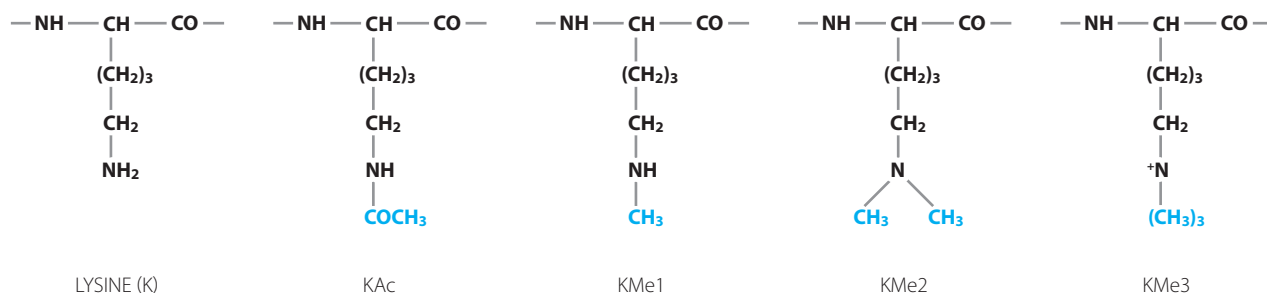


Figure 11.17 – Formulae of acetylated and mono-, di- and trimethylated lysine residues.

Other frequent modifications include phosphorylation of serines and methylation of arginines.

So, for example, EZH2 methylates lysine 27 of histone H3 (H3K27), while SETD2 trimethylates H3K36.

- **Readers** dock on to modified histones and mediate biochemical responses. For example, TAF3 (part of the transcription initiation complex) recognizes trimethylated H3K4 and helps position the initiation complex at promoters. Families of proteins containing chromodomains and bromodomains ‘read’ histones carrying specific methyl and acetyl marks, respectively.
- **Editors** remove histone marks. They include histone deacetylases (HDACs), demethylases and phosphatases. For example, the lysine demethylase KDM1A demethylates H3K4me, while HDAC1 is a histone deacetylase that functions in several large multiprotein repressor complexes.

An increasing list of genetic diseases turn out to be due to mutations affecting some of these players. The literature can be confusing because many of the enzymes have multiple names reflecting their independent discovery in different species. The nomenclature was rationalized by Allis *et al.* (2007). *Disease box 11* illustrates some of these ‘chromatin diseases’.

There is not a simple one-to-one ‘histone code’, but different histone marks are characteristic of different functional regions:

- Trimethylation of H3K9 and H3K27 marks repressed, non-transcribed chromatin.
- **Promoters** tend to have di- or trimethylated H3K4, acetylated H3K27 and a variant histone, H2A.Z.
- **Enhancers** have mono- or dimethylated H3K4 together with acetylated H3K27.
- **Gene bodies** are enriched for trimethylated H3K36 and dimethylated H3K79.

Chromatin conformation

The nucleus of a human cell might be 10 µm in diameter, but in a normal diploid cell it contains an astonishing 2 m of DNA. All this DNA is not just squashed in, it is intricately packaged in ways that are crucial for gene regulation. Only recently has much of this organization become amenable to investigation. For many years it was known that at the most basic level the DNA is organized round nucleosomes, and that the nucleosomes could be either compacted together, forming heterochromatin that is largely transcriptionally silent, or in a more open conformation, euchromatin, where genes may

or may not be transcribed. Within euchromatin, regulatory sequences could be identified by looking for DNA that was particularly accessible to molecules in its environment, rather than packed up in nucleosomes. These sequences generally turned out to be promoters or enhancers (as described in *Section 3.4*) and, as mentioned above, they are marked by particular histone modifications.

The distribution of accessible DNA is not fixed, but changes according to the needs of the cell. DNA methylation and histone modification play a role in this, but chromatin remodeling complexes are also crucial. As mentioned above, these are large multiprotein machines that use the energy of ATP to physically shuffle nucleosomes along DNA, swap variant histones into or out of nucleosomes, or disassemble them. They are involved whenever chromatin changes its activity – not only in activation or silencing of genes but also in DNA replication, DNA repair and segregation of chromosomes.

DNA methylation, histone modifications and nucleosome positioning interact and reinforce one another. For example, methylated histones can attract DNA methyltransferases so as to direct DNA methylation to specific genomic targets, while

Higher-level chromatin organization

Above this gene-level chromatin organization, until recently little was known about higher levels of organization except once we reach the level of whole chromosomes. These were known to occupy distinct territories within the interphase nucleus, as revealed by specialized FISH methods. A new technique, chromosome conformation capture (CCC; for details see Dekker *et al.*, 2013) has shed considerable light on the higher levels of organization. CCC identifies DNA sequences that lie physically close to one another in the interphase nucleus, even though they may be far apart in the linear genome. Application of the technique has shown that chromatin, above the nucleosome level, is organized into a series of loops called topologically associating domains (TADs). These are typically of the order of 100 kb in size, but with wide variations. Crucially, regulatory elements such as enhancers can only affect the expression of genes within the same TAD. The recognition of TAD boundaries has helped explain some otherwise baffling effects of small deletions, inversions or rearrangements in non-coding DNA. Sometimes these move TAD boundaries, and that can lead to abnormal phenotypes as an enhancer is prevented from driving expression of its normal target gene, but perhaps starts to affect expression of a different gene. The paper by Lupiáñez and colleagues (2016) gives examples. TADs seem to be relatively fixed features of chromosomes, conserved across cell types. However, as they are defined in experiments that use thousands of cells they may only represent an average of shifting and fluid interactions.

Above the level of TADs are larger-scale domains of two types, A and B. Type A domains contain transcriptionally active chromatin, type B domains are silent. These domains are dynamic structures that may each contain multiple TADs and that form or dissolve according to the state of activity of a cell. The relation between form and function of these chromatin structures is uncertain – do they determine the function of the chromatin they contain, or does the activity of the chromatin, controlled in some other way, cause these structures to form? The 4D Nucleome Project is one among several focusing on these questions (see www.4dnucleome.org). Research tools include sophisticated forms of CCC together with super-resolution microscopy at the single-cell level – but at present all we can say for sure is that gene regulation is exceedingly complex. That is hardly surprising when it allows a whole human to be constructed using only the same number of protein-coding genes as a small worm, and for individual cells and systems to respond appropriately to all manner of external factors.

proteins such as MeCP2 (see *Disease box 11*) that bind methylated DNA can recruit proteins that modify histones. Chromatin remodeling complexes that shunt nucleosomes along the DNA include components that recognize various histone modifications. It is seldom possible to point to a single initiating event and simple chain of causation, although ultimately binding of transcription factors plays a major role. Readers wishing more detail on the bewildering complexity could consult the reviews by Cavalli and Heard (2019) and Feinberg (2018).

At a descriptive level, we now have genome-wide maps of accessible DNA, of DNA methylation, of nucleosome positions and of histone modifications at single nucleotide resolution. Several consortia of research groups (for example, the International Human Epigenome Consortium and Roadmap Epigenomics Consortium, see *Useful websites*) are developing these resources, building on the work of the ENCODE Project Consortium (2012). Their task is very great: unlike DNA sequence, these features are specific to each individual type of cell. We may each of us have just two different genomes, but we each have many hundreds of different epigenomes in different cell types and at different times. And on top of that, we all vary from one another.

How far do epigenetic effects determine normal individual differences?

For scientists and philosophers the miracles of human development provide endless material for speculation and wonder. Evidence (disputed) of transgenerational epigenetic effects in humans (*Box 11.3*) raises profound questions about the full role of heredity in making us who we are, and suggests examples of the inheritance of acquired characters. For clinicians, especially those involved in public health, an important question is how far a person's general health is determined by environmentally acquired epigenetic marks imposed before or around the time of birth. The 'Barker hypothesis' (see Barker, 1990, 1995) holds that the general balance of a person's metabolism (their tendency to obesity, hypertension, etc.) is largely determined by their nutritional status in intrauterine and early neonatal life. Such long-term adaptation would surely require epigenetic mechanisms. Although the hypothesis has been controversial, much evidence supports some version of it. This leads on to some interesting questions. Epigenetic marks change in response to signals from the environment. Could a fuller understanding of these responses lead to effective prevention? More speculatively, could the current epidemics of obesity and type 2 diabetes owe something to the lifestyles of the parents, and not just those of the currently affected generations? None of that would alter the urgent public health need for interventions here and now, but it might help explain the remarkable intensity of those epidemics in the current generation.

Can epigenetic effects operate across generations?

While epigenetic changes can be transmitted from cell to daughter cell, they are not normally transmitted from parent to child. In X-inactivation, for example, the inactive X in a female is reactivated in the germ-line and the two X chromosomes become indistinguishable. However, there are many well-studied examples in plants of epigenetic marks being transmitted across generations, and a few clear-cut examples of transgenerational epigenetic effects in mammals (summarized by Daxinger and Whitelaw, 2012). It is controversial whether such effects occur in humans and, if so, how far any such cases are isolated oddities, and how far they are the tip

BOX 11.3

of an iceberg of wholesale transgenerational effects. This is a difficult area to study. Possible transgenerational effects down the female line are hard to disentangle from transmission of metabolic signals across the placenta. However, there are several plausible examples of effects transmitted down the male line, where epigenetic transmission is a strong candidate. Examples of possible male-line transgenerational effects (see Pembrey *et al.*, 2006) include:

- Historical studies from northern Sweden that show an association between the childhood food supply of the father and/or the paternal grandparents and a proband's longevity or risk of diabetic / cardiovascular mortality.
- A contemporary UK cohort study that shows an association between paternal *onset* of smoking in mid-childhood and increased body mass index in their future sons at 9 years.

Primordial germ cells in embryos undergo extensive demethylation of their DNA, which would seem to limit their ability to transmit epigenetic marks to offspring (see *Figure 11.7*) – but the demethylation is unlikely to be complete, and anyhow the signal might be transmitted in some other way, by RNA molecules in sperm, for example. One has to ask why, given that it is desirable to be adapted to one's environment, the proposed transgenerational epigenetic effects appear to adapt us to our grandparents' environment rather than our own. Supporters of transgenerational epigenetic programming might reply that these effects are only a very small part of the general epigenetic programming, but because they are counter-intuitive they force us to take epigenetic explanations seriously. The whole subject of inheritance that does not depend on DNA sequence is full of fascinating speculations and observations whose wider significance is much debated. Interested readers could consult the reviews by Blake and Watson (2016) and Miska and Ferguson-Smith (2016).

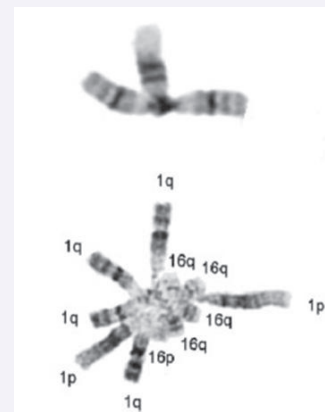
Chromatin diseases

The list of diseases caused by malfunction of epigenetic mechanisms is long and growing. The clinical features are not readily predictable from knowledge of the molecular defect, but it is striking how often the result is a clinically recognizable pattern of dysmorphic features.

Here we illustrate a few examples and then give a broader list in *Box table 11.1*.

ICF syndrome: a problem with methylating DNA. ICF syndrome (immunodeficiency, centromeric instability, facial anomalies; OMIM 242860) results from a defect in the *DNMT3B* DNA methyltransferase gene. Patients with this autosomal recessive disease show variable degrees of immunodeficiency and intellectual disability, and there is a characteristic facial appearance. When lymphocytes of ICF patients are cultured, a proportion of cells show abnormalities of the heterochromatin surrounding the centromeres of chromosomes 1, 9 and 16, including sometimes forming bizarre multiradial structures (*Box figure 11.2*).

Rett syndrome: a problem recognizing DNA methylation. Rett syndrome (OMIM 312750) is a childhood onset disorder, affecting almost exclusively girls. Usually there is a period of normal early development followed, between 6 and 18 months, by apparent regression with loss of purposeful use of the hands and the onset of characteristic hand wringing or other stereotypic movements. Head growth slows and up to 90% develop seizures. Breathing irregularities including hyperventilation and apnea are common and eye contact is difficult to achieve. The course, severity and age of onset of the condition vary from child to child. Some girls never



Box figure 11.2 – Unusual chromosome configurations in lymphocytes of a patient with ICF syndrome. Reproduced from *Am. J. Med. Genet.* 2007; **143A**:2052–2057 with permission from John Wiley & Sons.



Box figure 11.3 – A girl with Rett syndrome, showing the characteristic hand-wringing.

learn to walk or talk whilst others retain these skills to a certain extent. The condition can remain stationary for many years, although complications like scoliosis occur. In later years further deterioration may lead to loss of muscle bulk, limitation of mobility and liability to chest infections.

Most cases of Rett syndrome are caused by mutations in the gene encoding the methyl CpG-binding protein MeCP2. Mutations in males are largely lethal, but have been found in some males with a severe neonatal encephalopathy. The protein is a major 'reader' of DNA methylation. With different target genes it can be either a repressor or an activator. The genes regulated by MeCP2 activity are not yet all identified. It is also not understood why a period of normal development should occur.

Rubinstein–Taybi syndrome: a problem with a histone acetyltransferase. Patients with Rubinstein–Taybi syndrome (RSTS; OMIM 180849) have intellectual disability, a characteristic facial appearance, and broad thumbs and big toes (*Box figure 11.4*). 50–70% of patients have mutations in the *CREBBP* gene, encoding a histone acetyltransferase; a few have mutations in another HAT gene, *EP300*. In the remaining cases the cause is unknown.



(a)



(b)



(c)

Box figure 11.4 – Clinical features of Rubinstein–Taybi syndrome.

(a) The typical face; arched eyebrows and long nose with low columella, (b) broad laterally deviated thumb, and (c) broad great toe.

Kabuki syndrome 1: a problem with histone lysine methyltransferase.

Kabuki syndrome 1 (OMIM 147920) was named because of the supposed similarity of the facial appearance of patients to the make-up used in traditional Japanese Kabuki theatre. This largely sporadic disease usually results from *de novo* mutations in the *KMT2D* gene, encoding an enzyme that methylates lysine 4 of histone H3 in nucleosomes. Affected individuals have intellectual disability, short stature and can have a range of physical and developmental problems including structural cardiac and renal abnormalities.

Alpha-thalassemia / mental retardation syndrome: a defect in a chromatin remodeling protein. Affected males with this X-linked condition (OMIM 301040) have intellectual disability, a characteristic facial appearance, sometimes show male-to-female sex reversal and, unexpectedly, have a mild form of α -thalassemia (*Box figure 11.6*). Alpha-thalassemia is

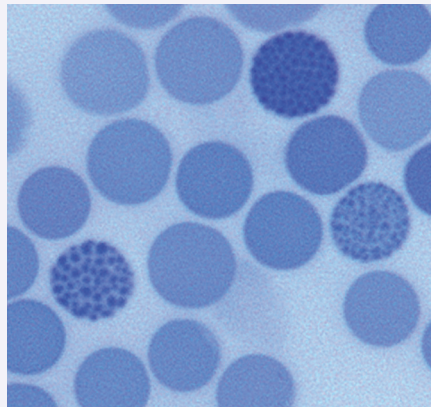


Box figure 11.5 – Clinical features of Kabuki syndrome.

The typical face with a gap in the eyebrows, long palpebral fissures and downturned corners of the mouth



(a)



(b)

Box figure 11.6 – ATRX syndrome.

(a) Facial appearance and (b) lymphocytes showing characteristic inclusions. Photo. (a) reproduced from *J. Med. Genet.* 1991; **28**: 742–745 with permission from the BMJ Publishing Group. Photo. (b) Courtesy of Dr Richard Gibbons, Oxford.

normally caused by deletion or inactivation of the alpha globin genes on chromosome 16. Lack of the *ATRX* gene product must affect the chromatin structure, and hence the expression, of several genes, including the alpha globin genes

Box table 11.1 – Examples of chromatin diseases

Function	Syndrome	OMIM number	Gene	Target
Writers:				
DNA methyltransferase	Tatton–Brown–Rahman	615879	<i>DNMT3A</i>	CpG
	ICF	242860	<i>DNMT3B</i>	CpG
Histone lysine methyltransferase	Sotos 1	117550	<i>NSD1</i>	H3K36 (me2)
	Luscan–Lumish	616831	<i>SETD2</i>	H3K36 (me3)
	Weaver	277590	<i>EZH2</i>	H3K27
	Kabuki 1	147920	<i>KMT2D</i>	H3K4
	Kleefstra	610253	<i>EHMT1</i>	H3K9
	Schinzel–Giedion	269150	<i>SETBP1</i>	
	Rubinstein–Taybi 1	180849	<i>CREBBP</i>	
Histone acetyltransferase	Rubinstein–Taybi 2	613684	<i>EP300</i>	
	Genitopatellar	606170	<i>KAT6B</i>	
	SBBYS	603736	<i>KAT6B</i>	
	Coffin–Lowry	303600	<i>RPS6KA3</i>	H3S10
Editors				
Lysine demethylases	Kabuki 2	300867	<i>KDM6A</i>	H3K27me2/3
	Claes–Jensen XLMR	300534	<i>KDM5C</i>	H3K4me2/3
	Siderius XLMR	300263	<i>PHF8</i>	H4K20me1
Histone deacetylase	Brachydactyly – ID	600430	<i>HDAC4</i>	
	Cornelia de Lange 5	300882	<i>HDAC8</i>	SMC3

Box table 11.1 *continued*

Function	Syndrome	OMIM number	Gene	Target
Chromatin remodeling	ATRX	301040	ATRX	
	Cockayne B	133540	ERCC6	
	CHARGE	214800	CHD7	
	Helsmoortel–Van der Aa	615873	ADNP	
	Coffin–Siris*	135900	SWI/SNF	
	Nicolaides–Baraitser*	601358	SMARCA2	
	Floating Harbor*	136140	SRCAP	
	Schimke immuno-osseous dysplasia*	242900	SMARCA1	

A wide variety of clinical syndromes are caused by defects in epigenetic processes. Histone methyltransferases and demethylases usually target a specific residue; acetyltransferases and deacetylases have wider specificities. *Syndromes caused by mutations in components of the BAF (SWI/SNF) chromatin remodeling complex; see Kosho *et al.* (2014). XLMR, X-linked mental retardation; ASD, autism spectrum disorder; ID, intellectual disability.

11.5. References

- Allis CD, Berger SL, Cote J, et al.** (2007) New nomenclature for chromatin-modifying enzymes. *Cell*, **131**: 633–636.
- Barker DJ** (1990) The fetal and infant origins of adult disease. *Br. Med. J.* **301**: 1111.
- Barker DJ** (1995) Fetal origins of coronary heart disease. *Br. Med. J.* **311**: 171–174.
- Blake GET and Watson ED** (2016) Unravelling the complex mechanisms of transgenerational epigenetic inheritance. *Curr. Opin. Chem. Biol.* **33**: 101–107.
- Carrel L and Willard HF** (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**: 400–404.
- Cavalli G and Heard E** (2019) Advances in epigenetics link genetics to the environment and disease. *Nature*, **571**: 491–499.
- Daxinger L and Whitelaw E** (2012) Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat. Rev. Genetics* **13**: 153–162.
- Dekker J, Marti-Renom MA and Mirny LA** (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Rev. Genet.* **14**: 390–403.
- ENCODE Project Consortium** (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**: 57–74.
- Feinberg AP** (2018) The key role of epigenetics in human disease prevention and mitigation. *New Engl. J. Med.* **378**: 1323–1334.

Holliday R (1994) Epigenetics: an overview. *Dev. Genet.* **15**: 453–457.

Kosho T, Miyake N and Carey JC (2014) Coffin–Siris syndrome and related disorders involving components of the BAF (mSWI/SNF) complex: Historical review and recent advances using next generation sequencing. *Am. J. Med. Genet.* **166C**: 241–251.

Levine ME, Lu AT, Quach A, et al. (2018) An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, **10**: 573–590.

Lupiáñez DG, Spielmann M and Mundlos S (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.* **32**: 225–237.

Miska EA and Ferguson-Smith AC (2016) Transgenerational inheritance: models and mechanisms of non-DNA sequence-based inheritance. *Science*, **354**: 59–63

Pembrey ME, Bygren LO, Kaati G, et al. (2006) Sex-specific male-line transgenerational responses in humans. *Eur. J. Hum. Genet.* **14**: 159–166.

Soellner L, Begemann M, Mackay DJ, et al. (2017) Recent advances in imprinting disorders. *Clin. Genet.* **91**: 3–13.

Wilkins JF and Haig D (2003) What good is genomic imprinting?: the function of parent-specific gene expression. *Nat. Rev. Genet.* **4**: 359–368.

Useful web sites

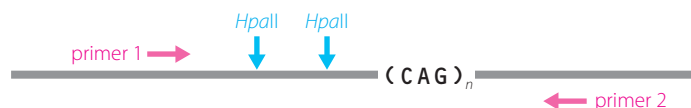
International Human Epigenome Consortium: www.ihec-epigenomes.org

Roadmap Epigenomics Consortium: www.roadmapepigenomics.org

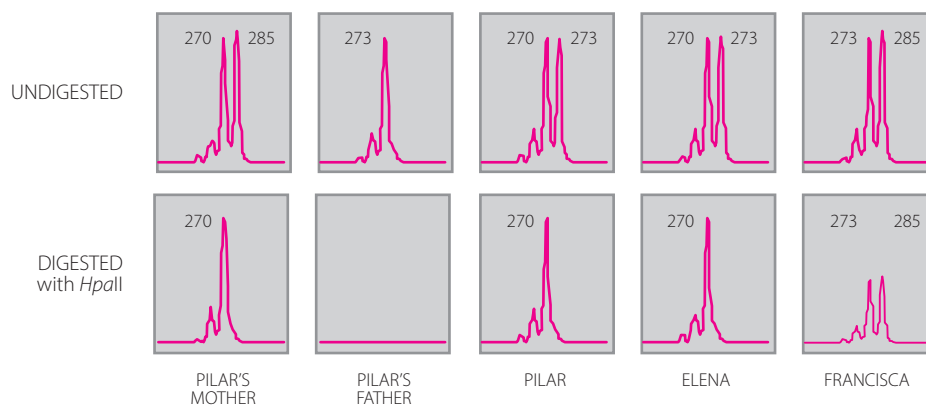
<http://igc.otago.ac.nz/home.html> is a source of information on imprinting and lists of imprinted genes.

11.6. Self-assessment questions

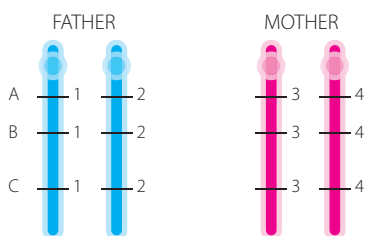
- (1) Suppose a gene on chromosome 6 is imprinted so that it is expressed only when it is inherited from the father. Complete absence of gene expression causes an unusual facial appearance. Draw a possible pedigree that you might see if a loss of function mutation in the gene is segregating in a large family.
- (2) Repeat the exercise in SAQ1 assuming that imprinting is such that the gene is expressed only from the maternal chromosome.
- (3) To check for skewed X-inactivation in women in the **Portillo family (Case 21)**, see pedigree in *Figure 10.9*, DNA was extracted from blood samples and an X-chromosome sequence was PCR-amplified that contained a variable trinucleotide repeat and two CCGG sites that are cut by the *HpaII* restriction enzyme, but only if they are unmethylated (see *Figure* below). For each woman the PCR was run twice, once with and once without prior digestion of the DNA with *HpaII*.



Results were as follows:



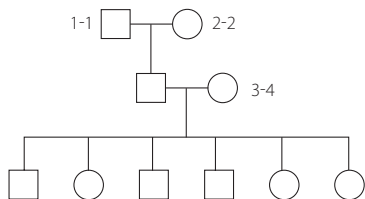
Explain these results.



- (4) The diagram shows types for three DNA polymorphisms on chromosome 15 in the parents of a child. Marker A maps within the PWS / Angelman syndrome critical region; the other two markers map distal to this region. Write possible marker genotypes for their child if he has:
- (a) PWS due to a deletion
 - (b) Angelman syndrome due to a deletion
 - (c) PWS due to uniparental disomy
 - (d) Angelman syndrome due to uniparental disomy
 - (e) PWS due to an imprinting error
 - (f) Angelman syndrome due to a mutation in the *UBE3A* gene.
- (5) Repeat SAQ4 assuming that in the paternal meiosis there was a crossover between the positions of markers A and B, and in the maternal meiosis between the positions of markers B and C. (NB this is not unrealistic – there is normally at least one crossover in each chromosome arm in each meiotic division).
- (6) Part of a DNA sequence is as follows:

5' CACTG^mCGGCCAAACAAGCA^mCGCCTG^mCGGCG^mCGCAGAGGCAG 3'

The cytosines indicated are either all methylated or all unmethylated, depending on the parental origin. The DNA is treated with sodium bisulfite and then a PCR primer is used to make a complementary strand. The primer used is off the diagram to the right. Design 10-nucleotide primers to specifically amplify the methylated and unmethylated versions of this sequence in conjunction with the downstream primer, and write out the sequence of this part of the PCR product. (Real primers would be 20–40 nucleotides long – they might need to be longer than normal PCR primers because it may be impossible to avoid some mismatches depending on which cytosines were methylated).



- (7) The pedigree shows genotypes for a polymorphic DNA marker that is located in the Xp–Yp pseudoautosomal region. Mark in possible genotypes for all the people in the pedigree.

[Hints on questions 1, 3 and 7 are provided in the *Guidance* section at the back of the book.]

12

When is screening useful?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Distinguish between screening and diagnosis
- Describe the parameters commonly used to define the performance of a screening program
- Describe the technical, social and ethical requirements that a screening program should fulfill
- Give examples of prenatal, neonatal and postnatal screening programs, and discuss their advantages and disadvantages to individuals and to society

12.1. Case studies

CASE 24 SMIT FAMILY

- Sam Smit, familial hypercholesterolemia
- Identified through cascade screening
- LDLR mutation detected, treatment started
- Affected relatives, including a homozygote

305

318

395

Back in 2001 at the age of 30, Sam Smit was contacted through the Dutch cascade screening program and found to have an elevated LDL-C (low density lipoprotein cholesterol) level (240 mg/dl; normal upper limit 190 mg/dl) and eventually a heterozygous mutation (c.551G>A, p.Cys184Tyr) in the *LDLR* (low density lipoprotein receptor) gene was found. Sam was started on regular statin medication which normalized his cholesterol level and he remained well. This finding was not a surprise to him since his father had died of a heart attack at the age of 48, but he wasn't aware of any further family history – his grandparents died in World War II and he had lost touch with his father's extended family.

Recently, however, he had been contacted by an older cousin, Pieter, the son of his father's sister, with information that worried him. Pieter's son Jan had married his second cousin Lotte and their young daughter Ana, aged 8 years, had just been found to have a very high level of LDL-C (700 mg/dl) after a doctor treating her for a fractured wrist noticed she had fatty deposits (xanthomata) between her fingers. Genetic testing showed Pieter, Jan and Lotte had a heterozygous mutation of the *LDLR* gene (the same variant as Sam) and they were prescribed statins and monitored regularly. However, Ana was homozygous for the mutation which leads to complete loss of LDL receptor function. Ana was now undergoing treatments with several medications on a trial basis with LDL apheresis twice weekly in a specialist center to try to lower her cholesterol levels.

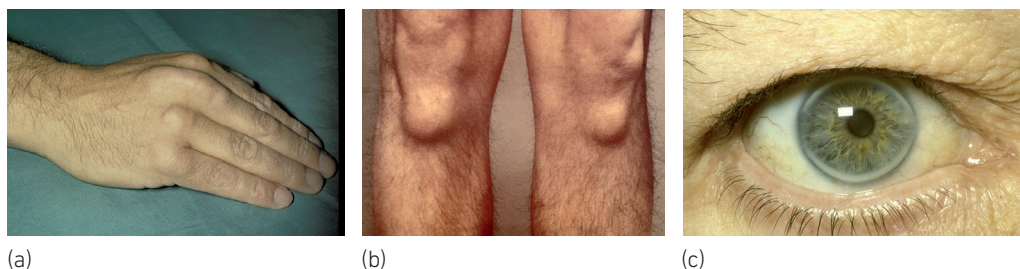


Figure 12.1 – Cholesterol deposition in patients heterozygous for familial hypercholesterolemia.

(a, b) tendon xanthomata, and (c) corneal arcus. Photos courtesy of Dr Paul Durrington, Manchester Royal Infirmary.

Sam, who had expected to be contacted about testing his son and daughter, found out that active cascade contact and testing posed more of a problem in the Dutch health system for relatives of people not followed up in a specialist clinic, because of new privacy regulations, and so he requested referral to the lipid clinic. In time his children were tested and neither was found to carry the family variant in *LDLR*.

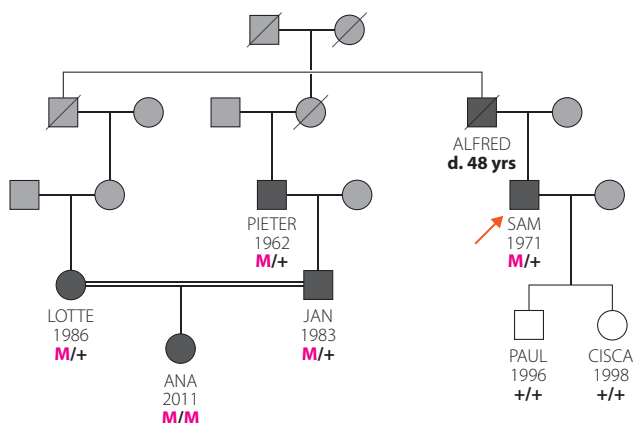


Figure 12.2 – Pedigree of the Smit family.

LDLR genotypes:

M mutation

+ wild-type

Symbols in pale gray show individuals not tested.

12.2. Science toolkit

Screening versus diagnostic tests

The word 'screening' is often used loosely as a synonym for testing, but the essentials of a true screening program are that it is applied to a whole population. In contrast to diagnostic tests, where a person has a problem and approaches the clinician for a test, screening is usually a top-down process, offered by some organization to a large cohort of people (probably defined by age, reproductive state or ethnicity). In any genetic screening program, ethical considerations are at least as important as technical questions. In this section we will consider some of the straightforward technical matters, and return to more general questions in the final section of this chapter.

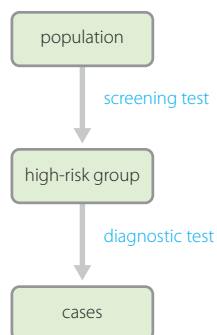


Figure 12.3 – Screening versus diagnostic tests.

Most screening tests do not result in a definite diagnosis. Because they are applied to large numbers of people, screening tests need to be cheap and simple. They can have rates of false positive and false negative results that would be unacceptable in a diagnostic test. The aim is not to make a diagnosis but to define a high-risk group of people, who are then offered a definitive diagnostic test (*Figure 12.3*). Because the numbers are then much smaller the diagnostic test can be more elaborate and expensive. Sometimes the distinction is blurred – for example, pregnant women are offered a detailed ultrasound malformation screen at 18–20 weeks, which for some malformations such as neural tube defects can also make a definitive diagnosis.

The distributions of screening test results in affected and unaffected people usually overlap (*Figure 12.4*) so that it is necessary to set an arbitrary cut-off point. This is always a compromise. Setting the cut-off too high means missing an unacceptably large proportion of the target group, while setting it too low means exposing too many unaffected people to the worry, inconvenience and expense of further investigation. *Box 12.1* shows some of the measures used to define the performance of a screening test, and these will inform the decision about where to place the cut-off. A website (<http://archive.wolfson.qmul.ac.uk/rscv2/>) allows interactive exploration of the trade-off between detection rate, false positive rate and the predictive strength of the risk factor that the screening test uses.

Even screening programs for mendelian diseases are subject to these compromises. DNA tests might give an immediate diagnostic answer, but they would rarely be used for population screening. We discuss the examples of phenylketonuria and cystic fibrosis below. Quick and cheap DNA tests always check for just one or a few specific variants,

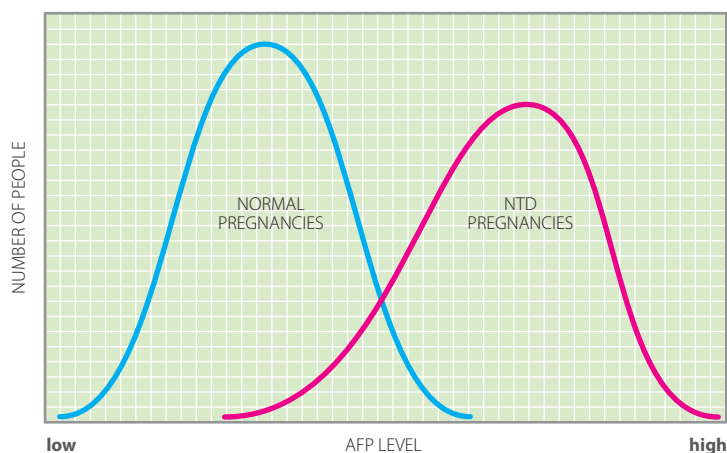


Figure 12.4 – A typical screening test.

The level of α -fetoprotein (AFP) in the serum of a pregnant woman is an indicator of the risk that her fetus has a neural tube defect (NTD: spina bifida or anencephaly). A raised level of AFP in the mother's blood indicates a higher risk, but the distributions in normal and affected pregnancies overlap. An arbitrary cut-off is used to select about 5% of women for further investigation. This involves detailed ultrasound examination and/or amniocentesis (see *Chapter 14*). Most women with raised AFP eventually deliver normal babies. Though largely superseded as a screening test by detailed ultrasound examination, the maternal serum AFP test illustrates typical features of a screening test.

but with loss of function conditions like these, many different variants in a gene can cause the loss of function, as described in *Section 6.4*. The time may come when DNA sequencing has become so cheap and convenient that it would be the natural tool to use for screening or, more probably, when most people's full genome sequence would already be stored in their medical record. In that case the technical (but not the ethical) distinction between screening and diagnosis would disappear – but we are not there yet.

We do not yet have routine whole-population genome sequencing, but we do already have DNA-based 'opportunistic screening'. This happens when a patient's exome or genome is sequenced as a diagnostic investigation for some clinical problem. This may or may not identify the cause of the problem for which the investigation was performed, but it may reveal a variant elsewhere in the exome or genome that is clinically significant, but for a condition totally unrelated to the original problem. For example, exome sequencing of a sick newborn might incidentally reveal a variant conferring a high risk of adult-onset cancer. Effectively, the variant was found as a result of a screening process. There has been fierce controversy about what to do with such 'incidental findings' – see *Section 12.4*.

Parameters of a screening test

Considering, for example, the AFP test shown in *Figure 12.4*, women can be positive or negative on the test, and can have a fetus with or without a neural tube defect (NTD). In the table, a, b, c and d are actual numbers of women in each category.

	Fetus has NTD	Fetus does not have NTD
Positive on test	a	c
Negative on test	b	d

Sensitivity of test = proportion of affected picked up = $a / (a+b)$

Specificity of test = proportion of all unaffected that are true negatives = $d / (c+d)$

False positive rate = proportion of all tests that give a false positive result = $c / (a+b+c+d)$

False negative rate = proportion of all tests that give a false negative result = $b / (a+b+c+d)$

Positive predictive value = proportion of all positives that are affected = $a / (a+c)$

Odds ratio = odds of being a case after a positive test ($a:c$) compared to odds of being a case after a negative test ($b:d$) = $(a/c)/(b/d) = ad/bc$

BOX 12.1

When might screening be done?

Genetic screening tests (*Figure 12.5*) fall into three groups:

- **Prenatal screening** – pregnant women in the UK, for example, are offered screening of their fetus for conditions including trisomies 13, 18 and 21 and various structural abnormalities including NTDs. Screening for Down syndrome is discussed below. Many individual mendelian and chromosomal conditions are also detectable prenatally, but these are not population screening tests; they are offered to individual couples who are known to be at high risk because of their family history.

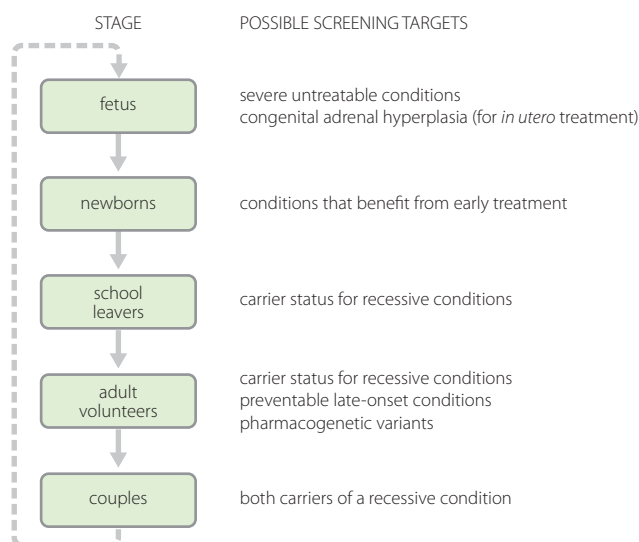


Figure 12.5 – When might screening be done?

- **Newborn screening** – as described in the next section, in many countries all newborn babies are screened for phenylketonuria. A variety of other tests may also be routinely offered (see *Section 12.4*). The list differs in different countries and often from institution to institution within a country. In the UK, national newborn screening programs are in place for phenylketonuria and five other inborn errors of metabolism, plus cystic fibrosis, sickle cell disease, hypothyroidism, hearing, and various problems detectable by clinical examination (www.gov.uk/guidance/newborn-blood-spot-screening-programme-overview). In many other countries the list is much longer (see *Section 12.4*). Additionally, it is becoming common for sick newborns to have their exomes sequenced as the quickest way to reach an urgently needed diagnosis, and this can amount to opportunistic screening, as mentioned earlier.
- **Adult screening** – this is most likely to be for carrier status for recessive diseases, for example, Tay–Sachs carrier screening in Ashkenazi Jews (see *Section 12.3, Case 19, Ulmer family*). Other ways of defining a high-risk group for genetic investigation might formally be described as screening: high-risk groups of adults defined by being relatives of a person with a certain condition (see the description of cascade screening in familial hypercholesterolemia, below) or defined by taking a certain drug (as described in *Chapter 10*, testing for pharmacogenetic variants affecting drug metabolism is mandated for a few drugs, and may perhaps become more common over the coming years).

For some conditions there are several different ways in which screening might be offered. Carrier screening for a recessive condition might be performed on newborn infants, on school leavers, on couples of childbearing age selected by family doctors, on pregnant women and their partners, or on anybody who cares to visit a drop-in center. Each approach has its advantages and drawbacks. Points to consider include:

- can the subjects give properly informed consent?
- how easy is it to access the group to be screened?

- how relevant will the information be at that time to the person screened?
- what are the practical and ethical implications of a positive result?
- what will the program cost, and do the benefits justify the cost?

These considerations are discussed a little further in *Section 12.3* in relation to Tay–Sachs carrier screening.

Who should be screened?

Screening can be offered to the whole population or to specific groups. Singling out particular groups for screening may become politically fraught, no matter how sound the epidemiological case for doing so. When Down syndrome screening was restricted to older women because of their higher risk (see *Figure 2.12*), some younger women who subsequently had a baby with Down syndrome complained that they had been unfairly discriminated against. Plans to screen specific ethnic groups have often caused trouble. In the UK, all babies are screened for sickle cell disease, and carrier screening is offered to all women early in pregnancy, even though the risk for white native British people is extremely low; however, the methodology used for carrier screening depends on the risk level in the local population.

Cascade screening is a halfway house between family-based clinical genetics and population screening. This is particularly used for familial hypercholesterolemia, as described in the next section, but could be used for any condition where risk extends beyond the nuclear family – mostly dominant adult-onset conditions. Like other screening programs, it is a top-down approach but family based, targeting distant relatives of an affected person.

How should screening be done?

A DNA test is seldom the best screening test (non-invasive prenatal testing is an exception, see *Disease box 12*). As mentioned above, the cheap and rapid DNA tests that would be suitable for large-scale screening always check for specific variants, but usually we simply want to know whether there is a loss of function, regardless of how it is caused. It is usually more efficient to use a functional test for large-scale screening. Examples discussed below include:

- Tay–Sachs screening normally measures the ability of the affected enzyme to break down an artificial substrate
- PKU screening is based on measuring the level of phenylalanine in a blood spot
- cholesterol level is used to screen for familial hypercholesterolemia
- cystic fibrosis screening primarily relies on measuring the level of immunoreactive trypsin, although carrier screening would necessarily use DNA tests.

These laboratory considerations are only part of the story. Questions of how a screening program is to be delivered, to whom, by whom and within what administrative framework are at least as important as what test the laboratory uses.

Antenatal screening for Down syndrome and other trisomies

Antenatal screening for Down syndrome and other chromosomal anomalies has been available for many years. Until the 1980s the screening test consisted of asking the mother's age. As *Figure 2.12* shows, the risk of having a baby with Down syndrome rises

sharply for older women. Over a certain cut-off (normally in the range 35–38 years, depending on resources available) women were offered a definitive diagnostic test. The diagnostic test is analysis of fetal material obtained by amniocentesis or chorion villus biopsy (Box 14.5). These procedures are invasive, unpleasant for the woman, and carry a 0.5–2% risk of causing a miscarriage. They are only appropriate for women at high risk.

Although the individual risk is higher for a woman over 38, there are so many more pregnancies among younger women that the majority of babies with Down syndrome are in fact born to younger women. Thus screening by maternal age alone has low sensitivity. A combination of maternal age, ultrasound findings and maternal serum biochemical markers can achieve a detection rate above 75% with a false positive rate of less than 3%.

- The ultrasound test measures the fluid under the skin of the neck of the fetus (nuchal translucency).
- Possible maternal serum biochemical markers include alpha-fetoprotein (AFP), beta-human chorionic gonadotrophin (hCG), unconjugated estriols (uE3), pregnancy-associated plasma protein A (PAPP-A) and inhibin A. For each, the distribution of levels in Down syndrome pregnancies (after adjustment for gestational age and maternal weight) is rather different from that in normal pregnancies. In the large US FASTER study (Malone *et al.*, 2005) median second trimester values for Down syndrome pregnancies, as multiples of the median for normal pregnancies, were 0.74 (AFP), 1.79 (hCG), 0.72 (uE3) and 1.98 (inhibin A). All the distributions for normal and trisomic pregnancies overlap strongly, so that none of the levels is diagnostic by itself, but in combination, and including the nuchal translucency and the mother's age, they give a composite risk figure that is far more predictive than age alone.

Women whose composite risk is greater than some predetermined cut-off are offered the diagnostic test. Deciding what cut-off to use involves a trade-off between sensitivity and false positive rate. Any increase in sensitivity will be accompanied by a decrease in specificity. Table 12.1 shows illustrative data from the FASTER study. First trimester screening allows women with positive results to have chorion villus biopsy and a less traumatic procedure for any termination of the pregnancy. Thus in the UK all women are offered a 'combined test' (ultrasound plus assay of PAPP-A and hCG) at 11–14 weeks of

Table 12.1 – Performance of different protocols for antenatal screening for Down syndrome

Protocol	Test sensitivity (%)		
	75	85	95
	False positive rate (%)		
NT, PAPP-A and hCG at 12 weeks	1.4	4.8	21
Quadruple test at 15–20 weeks	3.1	7.3	22
Fully integrated test	0.2	0.8	5.0

Figures show the percentage of false positive test results with different screening protocols and cut-off thresholds. The test sensitivity is determined by the threshold value of the composite risk that is chosen for declaring the screening result positive.

NT: nuchal translucency assessed by ultrasound.

Quadruple test: maternal serum AFP, hCG, uE3 and inhibin A.

Fully integrated test: NT + PAPP-A at 12 weeks followed by quadruple test in second trimester.

Data taken from the FASTER study (Malone *et al.*, 2005).

gestation. But some women book too late for this, and in others the nuchal translucency cannot be assessed because of the lie of the fetus or the mother's high body mass. These women may be offered a less accurate serum 'quadruple' test at 15–22 weeks, using hCG, uE3, AFP and inhibin A. Those with a composite risk over 1/150 on either test are offered a diagnostic test (Box 12.2).

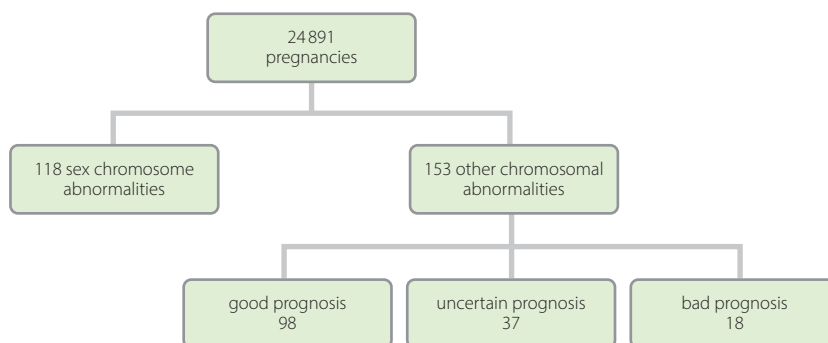
The developments in high-throughput DNA sequencing have led to a new alternative. Non-invasive prenatal testing (NIPT) has rapidly gained acceptance in many countries. This important development is discussed in *Disease box 12*.

What is the best prenatal diagnostic test for Down syndrome?

After antenatal screening by one of the methods described above, women whose risk of a Down syndrome baby is above the cut-off are offered a diagnostic test. Fetal cells are obtained by amniocentesis or chorion villus biopsy, as described in *Chapter 14*. What is the best way of analyzing these cells?

- *Traditional karyotyping*, as described in *Chapter 2*, or array-CGH (*Chapter 4*) can detect every chromosome abnormality, but arguably this is not desirable. While autosomal trisomies are always pathological, detecting some other abnormalities would present the couple with very difficult decisions that they might prefer not to face. Some, such as Turner syndrome or XYY, have well understood but relatively minor consequences, while the effects of others, such as *de novo* apparently balanced rearrangements or mosaicism for a small unidentified extra chromosome (a 'marker'), are unpredictable. Is it perhaps better not to know? *Box figure 12.1* shows some data found by full karyotyping.
- *Targeted molecular tests* check specifically for the common trisomies of chromosomes 13, 18 and 21. In most centers the sample is first tested by QF-PCR (see *Section 4.4*). A positive test for trisomy 13, 18 or 21 would lead to an offer of termination; some centers would first confirm that result by karyotyping, but others would not. If the QF-PCR result is negative, but the previous screening test showed abnormalities on ultrasound, the fetal DNA would most likely be tested by array-CGH (see *Chapter 4*).

In one UK survey, full karyotyping took 12 days and cost £253 per test; QF-PCR took 24 hours and cost £30 per test. But around 10% of the abnormalities missed by the rapid test have a poor prognosis. Some of these will be picked up on routine ultrasound examination, and some will abort spontaneously, but some will result in live-born abnormal babies. A common strategy is to get a quick result with QF-PCR and back this up with full karyotyping or array analysis, but some laboratories use just QF-PCR.



Box figure 12.1 – Chromosomal abnormalities found by full karyotyping of prenatal samples that would not have been detected by molecular trisomy testing.

Data on 24 891 pregnancies tested because of increased risk of Down syndrome, from Ogilvie *et al.* (2005).

12.3. Investigations of patients

CASE 2 BROWN FAMILY

- Baby Joanne, recurrent infections, poor growth
- Sweat test confirms she has cystic fibrosis
- Autosomal recessive inheritance
- Need for molecular test
- *CFTR* variants identified
- Molecular pathology
- Approaches to screening
- Possibilities for therapy

2

10

67

132

154

313

395

Both Joanne's parents David and Pauline come from quite large families (*Figure 1.8*). Their relatives are clearly at high risk of being carriers of cystic fibrosis. Since we now know what variants are present in David and Pauline (*Chapter 5*), it is fairly simple for any relative who might so wish to be tested for these specific variants. This is an example of **cascade screening**. Because cystic fibrosis variants are common in the general population, it is prudent also to test for a panel of the commonest variants, just in case anybody happens to be a carrier coincidentally, having inherited a variant from a different ancestor from the ones who were the source of David's or Pauline's variants.

As the most frequent severe recessive disease in many populations of European origin, cystic fibrosis has been considered for general population screening as well as cascade screening. This could include newborn screening to detect affected babies and/or screening adults to detect carriers. The argument for newborn screening is that early treatment improves the prognosis. In both the USA and UK this argument has been accepted, and universal newborn screening is recommended. *Figure 12.6* shows the flow-chart for the British scheme. The method is based on measurement of immunoreactive trypsin (IRT), which is raised in cystic fibrosis.

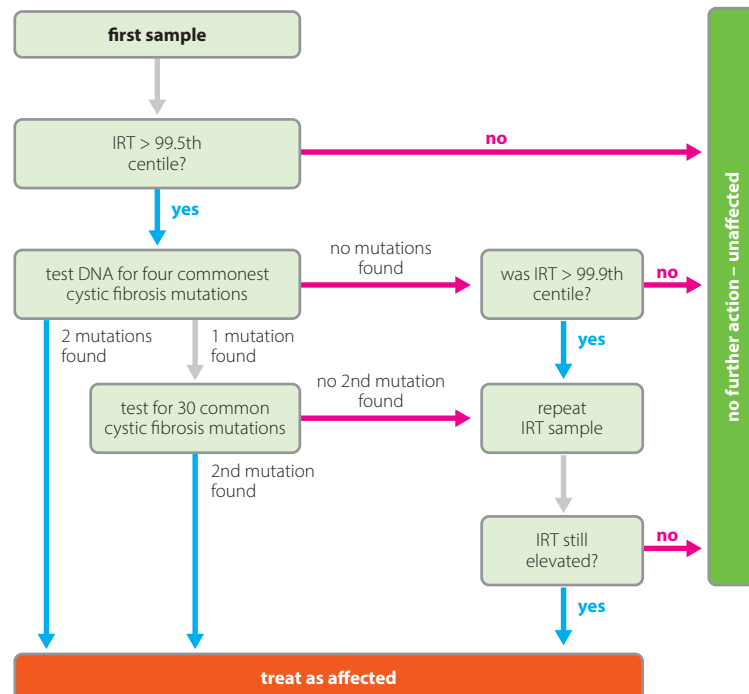


Figure 12.6 – Flowchart of the British scheme for screening newborn babies to allow early identification and treatment of those with cystic fibrosis.

IRT, immunoreactive trypsin. The 4 *CFTR* mutations checked are p.F508del, p.G542X, p.G551D and c.621+1G>T. These are estimated to account for 80% of severe CF variants in the native UK population. Adapted from *A Laboratory Guide to Newborn Screening in the UK for Cystic Fibrosis* (2014) under an Open Government Licence v2.0 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/397726/Cystic_Fibrosis_Lab_Guide_February_2014_v1.0_12_.pdf).

For carrier screening it is necessary to use a DNA test, because carriers are entirely normal biochemically. The problem here is the very large number of variants (over 1000) that have been described. As pointed out in *Section 5.4*, it is easy to screen for a specific variant but hard to test a gene for *any* possible variant. The *CFTR* gene has 27 exons (see *Figure 3.7*) and there is currently no way of scanning the entire sequence of such a gene at a cost low enough to use in population screening. Because of founder effects and heterozygote advantage (*Chapter 9*) a few specific variants make up the bulk of all cystic fibrosis variants in any particular community. Any carrier screening protocol therefore involves a trade-off between cost and sensitivity. Testing for the few most frequent variants is cheap but misses some proportion of carriers. Testing for many variants is expensive, and however many individual variants are checked, the sensitivity will never be 100%, unless the whole gene (including introns) is sequenced. *Table 12.2* shows the distribution of *CFTR* variants detected by one UK laboratory.

Table 12.2 – Distribution of 199 CFTR variants analyzed in Manchester, UK, 2007–2013

Variant	Exon	Number found	Percentage of mutations	Cumulative %
p.F508del	11	137	68.8	68.8
p.R117H	4	14	7.0	75.9
p.G551D	12	11	5.5	81.4
c. 3272–26A>G	Intron 19	6	3.0	84.4
c. 621+1G>T	Intron 4	3	1.5	85.9
p.N1303K	24	2	1.0	86.9
c. 1717–1G>A	Intron 11	2	1.0	87.9
p.G542X	12	1	0.5	88.4
c. 1898+1G>A	Intron 13	1	0.5	88.9
p.R553X	12	1	0.5	89.4
p.R560T	12	0	0.0	89.4
c.3659delC	22	0	0.0	89.4
c.2715 delT	15	0	0.0	89.4
Other detected		21	10.6	100
TOTAL		199		100.0

Note how small increases in sensitivity require large increases in the number of variants checked. The data does not include cases where no pathogenic variant was found, as it is unclear how many of those truly have cystic fibrosis. Numbering follows ENSEMBL transcript ENST00000003084. Data courtesy of Joanna Brock and Derek Barley, St Mary's Hospital, Manchester.

Inevitably, some couples who have tested negative will actually both be carriers, with a 1 in 4 chance of having an affected baby. Some of these couples will both be carriers of rare 'private' variants, particularly if they are a consanguineous couple with non-European ancestry. Most, however, will be couples where one partner was recognized to be a carrier, but the other carried a rare variant and was a false negative on the test. Given the smaller number of such people compared to the number initially screened, it

is possible to test the partner for a larger panel of variants to try to minimize the number of false negatives. Ideally one might wish that the risk for a positive–negative couple after screening should be no greater than the initial risk for any couple before screening. It could then be argued that the program has left nobody in a worse position than they were in before screening. For a carrier frequency of 1 in 23, this requires a 99.8% sensitivity for testing the partner of a known carrier (so that the risk of a false negative is 1 in 500). The figures in *Table 12.2* suggest that reaching 99.8% sensitivity would be a challenge.

CASE 4 DAVIES FAMILY

- Martin, aged 24 months, clumsy and slow to walk
- Family history of muscular dystrophy
- X-linked recessive inheritance
- Problems of testing dystrophin gene
- Exon 44–48 deletion identified by MLPA
- Molecular pathology
- Implications of X-inactivation
- Screen all newborn boys?
- Possibilities for therapy

4

11

68

98

156

285

315

395

Martin was not diagnosed with Duchenne muscular dystrophy until he was 2 years old, when his slow walking and clumsiness had become apparent. In some families, by the time the first affected boy is diagnosed there is already a second affected boy. This has led to the suggestion that all newborn boys should be screened for the disease. Technically this could be done reasonably cheaply by measuring the serum creatine kinase, which is strongly elevated in affected boys (see *Chapter 4*). Baby boys with a high CK could then be tested by a multiplex MLPA deletion screen (see *Figure 4.11*), which would pick up about two-thirds of cases. For the remainder, a muscle biopsy might be required, to demonstrate the absence of dystrophin protein.

The proposal has been controversial. A small number of repeat cases would be avoided; on the other hand, making so grave a diagnosis so early, and when there are only very limited prospects for treatment, robs the family of a couple of years of happy parenthood. There is also the question of how many unaffected boys whose CK was for some reason high would be subjected to muscle biopsy, and how serious the psychological trauma would be for parents whose baby initially tested positive, but was eventually shown to be unaffected. Neonatal DMD screening is not currently recommended by the UK National Screening Committee (see <https://legacyscreening.phe.org.uk/screening-recommendations.php>) but this decision will be reviewed in the coming years to take account of the results of treatment trials and the need for earlier diagnosis if effective treatment becomes possible.

CASE 8 HOWARD FAMILY

- Helen, newborn daughter of young parents
- Down syndrome confirmed
- 47,XX,+21 karyotype
- Options for prenatal testing
- Non-invasive prenatal test
- Possibilities for therapy

26

39

70

315

395

Women such as Helen's mother Anne, who had a previous child with Down syndrome, would normally be offered a diagnostic test regardless of their age or objective risk level, by-passing the biochemical screening procedures described above. Although the recurrence risk for a younger woman is low, their anxiety level is naturally very high after having had an affected baby, and it would be cruel to refuse them the test. The availability of NIPT has widened the options. Although NIPT is considered a screening test rather than a diagnostic test, the false negative rate is very low and for women who test negative it avoids the discomfort and risks associated with chorion villus biopsy or amniocentesis. Anne opted for this test (see *Disease box 12*). Reassured by the negative result, she went on to deliver a normal healthy boy.

CASE 13 NICOLAIDES FAMILY

- Spiros and Elena both carriers of β -thalassemia
- Need to define mutations for prenatal diagnosis
- Allele-specific PCR shows Spiros carries the p.Gln39X variant
- Restriction digest shows Elena carries the c.316–106C>G variant
- Molecular pathology
- Population screening for carriers
- Possibilities for therapy

117

129

159

316

395

Hemoglobinopathies (sickle cell disease and thalassemias) are the most frequent severe recessive conditions worldwide, particularly affecting people in or with ancestry from countries where malaria is or was prevalent. Population-wide carrier screening is implemented in many such countries. Carrier screening normally uses routine hemoglobin and red cell measures, supplemented as necessary by protein and/or DNA analysis focused on the particular variants that are frequent in that population. In the Nicolaides family, routine blood screening identified both Spiros and Elena as β -thalassemia carriers, then DNA testing for the specific frequent Cypriot variants identified the precise variant in each of them. In low-prevalence countries like the UK, carrier screening may be restricted to the antenatal clinic. Newborn screening is used for sickle cell disease but (in the UK) not for thalassemia because of the poor sensitivity of routine tests.

CASE 19 ULMER FAMILY

- Hannah, 6-month-old baby girl, Ashkenazi Jewish background
- Normal at birth but then increasing problems
- ? Tay–Sachs disease
- Enzyme test confirms diagnosis
- Test the sibs?
- Carrier screening
- Possibilities for therapy

231

239

316

395

Among Ashkenazi Jews the carrier frequency for Tay–Sachs disease is 1 in 30. Because of this, Ashkenazi-specific carrier screening programs have been established in the USA and many other countries, starting in the early 1970s. *Table 12.3* shows some statistics. In countries where the Ashkenazi community have embraced carrier screening enthusiastically, the majority of affected babies are now born to non-Jewish couples.

Table 12.3 – Tay–Sachs carrier screening among Ashkenazi Jews, 1971–1998

Country	Number tested	Carriers	Carrier–carrier couples
United States	925 876	35 372	795
Israel	302 395	7277	380
Canada	65 813	3301	62
South Africa	15 138	1582	52
Europe	17 725	1127	37
Brazil	1027	72	20
Mexico	655	26	0
Argentina	84	5	0
Australia	3334	102	4
TOTAL	1 332 047	48 864	1350

Data from Chapter 153 (Gravel *et al.*) of *Metabolic and Molecular Basis of Inherited Disease* edited by Scriver *et al.* (2001) and reproduced here with permission from The McGraw–Hill Companies.

Screening uses a biochemical test, measuring the ability of hexosaminidase A in serum to hydrolyze an artificial substrate MUG (an *N*-acetyl glucosamine derivative of 4-methyl umbelliferone). DNA testing is not used for population screening because not all carriers have one of the common Ashkenazi variants (see *Table 9.2*). A more definitive diagnostic

test is required after a positive screening result. About 2% of Jewish and 35% of non-Jewish people identified as carriers by the screening test in fact carry a so-called pseudodeficiency allele. Such alleles encode a variant form of the enzyme that is inactive against the artificial substrate used in the screening test, but retains enough activity against G_{M2} ganglioside to be non-pathogenic. Screening for Tay–Sachs disease is usually combined with screening for a variable selection of the other ‘Jewish diseases’ listed in *Disease box 9*.

There are various ways in which a Tay–Sachs carrier screening program could be implemented, and various actions that might follow a positive test result. Which one is adopted is very much a matter for the community concerned. Whatever protocol is adopted, screening should be voluntary and requires fully informed consent.

- **Testing babies or children** for carrier status (as distinct from testing newborns for being clinically affected) is unethical and inefficient. The child cannot give proper consent and the result has no implications until many years later, by which time it may well have been forgotten.
- **Testing school leavers** raises issues of just how far consent is truly voluntary, and requires careful handling to avoid stigmatization of carriers. Some Orthodox communities try to solve the stigmatization problem by giving the results to a match-maker and not to the individual tested (see below).
- **Testing young single adult volunteers** minimizes problems of consent but will miss some fraction of the target group.
- **Testing couples** identifies the small number of at-risk couples without worrying the much larger number of carriers whose spouse is a non-carrier (compare the last two columns of *Table 12.3*). But it removes the possibility of choosing not to marry a carrier. In some Orthodox communities where marriages are semi-arranged, the match-maker has the results of testing young single people for a range of recessive conditions, and will indicate if a proposed marriage would be risky.

CASE 20 VLASI FAMILY

- Valon, 6-year-old boy with serious learning problems
- Small, microcephalic, blue eyes, fair skin and hair, eczema; hyperactive
- ? Phenylketonuria
- Testing for subsequent baby?
- Newborn screening
- Possibilities for therapy

251 262 **317** 395 396

Phenylketonuria is the target of one of the most widely implemented newborn screening programs. Early detection allows early treatment (*Chapter 14*), which has excellent results and is highly cost-effective by avoiding the cost of lifelong institutional care.

The screening test is to measure the level of phenylalanine in the baby's blood. As mentioned in *Section 10.3*, the blood cannot be taken immediately after delivery of the baby, convenient though that would be, because until the placental connection is broken the mother (almost certainly a clinically normal heterozygote) will clear phenylalanine from the baby's circulation. Once the connection is broken, phenylalanine starts to accumulate. The optimum time for testing is 5 days after birth, but provided at least 24 hours is allowed to elapse the false negative rate is acceptably low.

Laboratories may use various methods to check the phenylalanine level, including a bacterial growth assay, chromatography, fluorimetry or tandem mass spectrometry. The latter has the advantage of allowing a number of other analytes to be measured at the same time (see *Table 12.4*). A DNA test is not used at this stage because there are many possible mutations in the *PAH* gene. A direct enzyme assay is not used because that would require a liver biopsy. Whatever the method, this is a screening test and not a diagnostic test. The cut-off for the screening test is usually set around 120 μM (normal

range $58 \pm 15 \mu\text{M}$). Babies whose blood level is above the cut-off are called in for more specific testing. This involves more careful measurement of the blood phenylalanine level. In PKU this is typically above $1000 \mu\text{M}$. A lower, but still elevated level is seen in babies with benign hyperphenylalaninemia. These babies develop normally without treatment. The screening test has a sensitivity of around 98–99% for PKU, provided it is not performed too soon after birth. Babies proven to have PKU are put on a special diet, as described in *Chapter 14*. Careful adherence to the diet ensures that the child grows up with no or minimal cognitive impairment.

Phenylalanine hydroxylase requires an essential cofactor, tetrahydrobiopterin (BH_4). A small percentage of phenylketonuric babies have a genetic defect in the production or recycling of BH_4 (OMIM 261640), rather than mutations in the *PAH* gene. These require a different treatment because BH_4 is required for several other amino acid hydroxylations. The laboratory work-up checks for these variant forms of PKU, and will often include DNA studies to define the mutations in the *PAH* gene.

CASE 24 SMIT FAMILY

- Sam Smit, familial hypercholesterolemia
- Identified through cascade screening
- LDLR mutation detected, treatment started
- Affected relatives, including a homozygote
- Conflict between privacy and cascade screening
- Possibilities for therapy

305

318

395

Familial hypercholesterolemia (FH; OMIM 143890) is an autosomal dominant condition. In many populations it is the most frequent of all clinically significant mendelian conditions, affecting up to 1 person in 250. Heterozygotes typically have serum cholesterol and LDL-cholesterol levels of 250–450 and 200–400 mg/dl (normal range 150–250 and 75–190 mg/dl, respectively), unrelated to diet. They develop tendon xanthomata (subcutaneous cholesterol deposits) and suffer coronary artery disease in mid-life, with many premature deaths from myocardial infarction. Untreated male subjects are at a 50% risk for a fatal or nonfatal coronary event by 50 years of age and untreated female subjects are at a 30% risk by 60 years of age. Rare homozygous affected people, like Pieter's grand-daughter Ana, have these features to a more extreme degree and, if untreated, die prematurely from heart disease.

Michael Brown and Joseph Goldstein won the 1985 Nobel Prize for Medicine for their work on FH that has led to a detailed understanding of cholesterol homeostasis (see *Useful websites* list at the end for further information). They demonstrated that FH was usually caused by mutations in the low density lipoprotein receptor (*LDLR*) gene. This cell surface receptor imports cholesterol-containing LDL into liver cells, where it represses endogenous cholesterol synthesis as part of a homeostatic mechanism. The mutant receptor is either absent or fails to bind LDL, depending on the particular mutation, leading to uncontrolled endogenous production of cholesterol. Sometimes FH can be caused by mutations in either of two other genes:

- some people produce LDL that is not recognized by the LDL receptor because a mutant *APOB* gene encodes an abnormal form of the lipoprotein
- occasional patients have gain of function mis-sense mutations in *PCSK9*, a gene encoding a protein-processing enzyme that is part of the homeostatic mechanism.

The normal action of internalized LDL is to repress 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase, which catalyzes the rate-limiting step in cholesterol synthesis. In FH patients, statin drugs are used to inhibit the enzyme. This is a very effective clinical management, which can normalize cholesterol levels and health risks. Because the condition is common, has serious health implications and can be effectively treated,

there is a good case for screening the whole population for FH. However, the resource implications are considerable. Screening based on cholesterol levels and tendon xanthomata has poor sensitivity, especially in younger people. DNA-based screening is expensive because the *LDLR* gene on chromosome 19p13.2 has 18 exons, and several hundred different mutations have been described.

Since this dominant condition affects large extended pedigrees, a very cost-effective method of ascertaining large numbers of affected people is to test relatives of known affected cases, a procedure known as **cascade screening** or cascade testing. The principle was mentioned earlier in connection with testing relatives of **Joanne Brown (Case 2)** for cystic fibrosis mutations. Index cases are ascertained through a variety of means. It has been argued that the most effective way of ascertaining index cases would be to test 1 year old children when they come for a routine vaccination.

In the Netherlands, a nationwide and government-subsidized cascade screening program started in 1994. Relatives of index cases were contacted directly by a genetic field worker. More than 28 000 individuals (an average of 5.6 per index case) were genetically diagnosed with FH and entered in a central national database. Of those who turned out to be affected, only 39% were already taking statins. However, new privacy regulations meant that family members could no longer be actively approached by the genetic field workers. Instead, index cases are asked to contact their relatives and, unsurprisingly, the number of affected family members identified has plummeted. Details of this and other programs are reviewed by Louter *et al.* (2017), while Sturm *et al.* (2018) discuss in depth the problems of FH screening.

Studies of the Dutch program have demonstrated very clear clinical benefits; 93% of affected people ascertained through the program subsequently took statins, thus greatly reducing their risk of premature death. Various ethical issues can be raised. The direct contacting of relatives is considered ethically problematic in some countries (including, now, in the Netherlands), and instead probands are left to suggest to their relatives that they might benefit from talking to a geneticist. There are questions about identifying affected children – does the benefit of early treatment outweigh the risk of stigmatization or impaired self-image? It is also necessary to think about insurance implications for people contacted out of the blue. Logically, underwriting should be based on the phenotype – the actual cholesterol level in the treated patient – and not on the genotype. Since statin treatment is effective, this should (but may not) largely dispose of the insurance issues. Careful studies of the Dutch program have shown none of the predicted ill effects, though it is possible that this partly reflects the admirable qualities of Dutch society, and might not be so easily replicated elsewhere.

12.4. Going deeper...

What conditions should we screen for?

Technically, the possibilities for genetic screening seem almost unlimited – certainly they are already vast, and are increasing every year. In practice, however, what is on offer is very much more limited. Apart from the inevitable time-lag in bringing new developments into service, there are four main technical reasons why genetic screening is not more widespread (social and ethical factors are considered later).

- If you work through *Self-assessment questions 1–3*, you will see that the positive predictive value of a test may be very low, even if the test performs well in the laboratory. It is technically difficult, as well as economically questionable, to screen for rare conditions. This is particularly an issue with direct-to-consumer (DTC) testing. A recent investigation (Weedon *et al.*, 2019) suggests that not merely some, but the great majority of *all* pathogenic rare variants reported by these companies are false positives. These matters are discussed further below.
- A DNA variant may be indisputably associated with increased risk of a disease, but may be responsible for only a small fraction of the overall risk. As described in *Chapter 13*, many DNA single nucleotide polymorphisms are associated with an increased risk of developing a common complex disease, but for almost all of them the odds ratio is so low as to be clinically insignificant. The newly developed Polygenic Risk Scores may be a way forward – see *Section 13.4* for further discussion. The **Population Attributable Risk** (*Box 12.3*) is an important consideration in any screening program. If a variant explains only 5% of the total risk, what is the point in screening for it? If that 5% is concentrated in a few individuals who have a very high risk, it might be valuable for those individuals to know – but this is approaching the situation for mendelian diseases, where testing is mainly family-based. If the variant is common, so that the extra risk is spread across a large proportion of the population, there seems little value to anybody in knowing.
- It is important to consider not just the *relative risk* but also the *absolute risk*. Even a high relative risk may not be too worrisome if it translates into a low absolute risk. An example is Factor V Leiden (OMIM 227400). This is a well-authenticated risk factor for venous thromboembolism, for example, among long-haul airline passengers and oral contraceptive users. The *relative risk* among oral contraceptive users is high (around 15) but the *absolute risk* is still quite low, because oral contraceptive users are generally young and the risk of embolism is very age-dependent.
- In general, there is no point in screening unless a positive result leads to some useful action. This might mean lifestyle changes, prophylactic drugs to reduce the risk, or maybe increased surveillance, for example, to pick up cancer at an early stage when it is still treatable. Sometimes genetic testing is done simply to provide people with information for planning their future as, for example, in predictive testing for Huntington disease, but normally a screening test should lead to some practical outcome.

The Office of Genomics and Disease Prevention at the Centers for Disease Control and Prevention in the USA has provided a framework that can be used for assessing any genetic (or other) test, not just a screening test. The ACCE framework suggests a test should be assessed against four sets of criteria.

- **Analytical validity:** how accurately does the test measure what it is supposed to measure? For a DNA test this might translate into asking how accurately chip-based assays genotype rare variants, or what proportion of all mutations in a gene are picked up by the test protocol.
- **Clinical validity:** how accurately does the test detect or predict the presence or absence of disease? For example, most people with hereditary hemochromatosis (OMIM 235200) have mutations in both copies of their *HFE*

gene, usually p.Cys282Tyr and/or p.His63Asp. But testing for these variants has low clinical validity because only around 5% of homozygotes or compound heterozygotes for the variants manifest clinical hemochromatosis.

- **Clinical utility:** how useful are the test results in providing clinical benefits? Everybody with Type 1 Waardenburg syndrome (OMIM 193500) has a pathogenic variant in the *PAX3* gene. The test has high analytical and clinical validity – but detecting the variants does not provide much obvious clinical benefit to most patients, beyond satisfying their curiosity.
- **Ethical, legal and social implications.**

The Population Attributable Risk

This is the proportion of the total disease risk in the population that is attributable to the factor in question. It is sometimes called the Population Attributable Fraction. If the PAR is low, it calls into question the value of screening for the factor.

If the overall risk in the population is r , but a proportion p of the population have a variant that gives them an additional risk of R , on top of the general risk, then the PAR for that variant is pR/r .

BOX 12.3

Population-based screening programs, where individuals opt out rather than opt in, need to be funded by the state or insurance companies. Funders will evaluate any proposed scheme critically against technical, financial and ethical criteria. The criteria used are usually based on a set formulated by Wilson and Jungner (1968) in a report for the World Health Organization. Box 12.4 shows a selection of the criteria used by the UK National Screening Committee. In the directly marketed sector, the arguments may look very different. Companies will offer a genetic test if they see a likely profitable market for it, regardless of whether or not it might result in useful actions or long-term benefits. This has resulted in various companies offering 'lifestyle' genetic screening: testing for variants associated with increased risk of some common disease, coupled with advice on how to mitigate any increased risk, as discussed below. Profitability may, however, depend on being able to pass part of the cost on to somebody else. Somebody in the UK who gets a disturbing test result from an internet-based company, will probably expect their family doctor and NHS genetics service to deal with their worries.

Newborn screening is the least ethically contentious area of population screening, because the targets are treatable conditions and the aim is to ensure early treatment to avoid irreversible damage. In every advanced country newborn babies are screened for a variety of conditions where early diagnosis has the potential to improve outcome. Screening for biochemical abnormalities uses a blood spot taken from a heel prick. Babies testing positive are then referred for a definitive diagnostic test. As previously mentioned, blood sampling has to wait until some time after delivery, because until the placental connection is broken, the baby's blood chemistry will be heavily influenced by the mother (see Figure 10.8).

In the USA, the Department of Health and Human Services has defined a recommended universal newborn screening panel (see Table 12.4). This lists 35 core disorders, plus another 26 that can be detected in the differential diagnosis of a core disorder. The list includes many very rare conditions that can all be diagnosed by tandem mass spectrometry.

Table 12.4 – 35 core disorders for which the US Department of Health and Human Services recommends universal newborn screening

Propionic acidemia	Classic phenylketonuria
Methylmalonic acidemia (methylmalonyl-CoA mutase)	Tyrosinemia, Type I
Methylmalonic acidemia (cobalamin disorders)	Primary congenital hypothyroidism
Isovaleric acidemia	Congenital adrenal hyperplasia
3-methylcrotonyl-CoA carboxylase deficiency	S,S disease (sickle cell anemia)
3-hydroxy-3-methylglutaric aciduria	S, β -thalassemia
Holocarboxylase synthase deficiency	S,C disease
β -ketothiolase deficiency	Biotinidase deficiency
Glutaric acidemia Type I	Critical congenital heart disease
Carnitine uptake defect / carnitine transport defect	Cystic fibrosis
Medium-chain acyl-CoA dehydrogenase deficiency	Classic galactosemia
Very long-chain acyl-CoA dehydrogenase deficiency	Glycogen storage disease Type II (Pompe)
Long-chain L-3 hydroxyacyl-CoA dehydrogenase deficiency	Hearing loss
Trifunctional protein deficiency	Severe combined immunodeficiencies
Argininosuccinic aciduria	Mucopolysaccharidosis Type 1
Citrullinemia, Type I	X-linked adrenoleukodystrophy
Maple syrup urine disease	Spinal muscular atrophy due to homozygous deletion of exon 7 in SMN1
Homocystinuria	

See www.hrsa.gov/advisory-committees/heritable-disorders/rusp/index.html.

In the UK the policy is more conservative and cautious. Only nine conditions are included in the national newborn blood-spot screening program (see www.gov.uk/government/collections/newborn-blood-spot-screening-programme-supporting-publications): congenital hypothyroidism, cystic fibrosis, sickle cell disease and six inherited metabolic diseases: glutaric aciduria type 1, homocystinuria, isovaleric aciduria, maple syrup urine disease, medium chain acyl-CoA dehydrogenase deficiency and phenylketonuria. This shorter list is dictated partly by cost, but also by the very stringent requirements used by the UK National Screening Committee for approving a screening program (*Box 12.4*). For most of the rarer biochemical disorders, the diagnosis is not made, and treatment not started, until the baby comes to attention because it is sick. Perhaps the universal free National Health Service reduces the risk of a sick baby slipping through the net and not being diagnosed.

Key criteria used by the UK National Screening Committee

For a complete list of the 20 criteria see: www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes.

The condition:

- (1) The condition should be an important health problem ... The epidemiology and natural history of the condition should be adequately understood ...
- (2) All the cost-effective primary prevention interventions should have been implemented as far as practicable.
- (3) If carriers of a mutation are identified as a result of screening, the natural history of people with this status should be understood, including the psychological implications.

The test:

- (4) There should be a simple, safe, precise and validated screening test.
- (6) The test should be acceptable to the population.
- (8) If the test is for particular mutations, the method by which they have been selected and the means by which they will be kept under review, should be clearly set out.

The intervention:

- (9) There should be an effective intervention for patients identified through early detection, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care.

The screening programme:

- (11) There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an 'informed choice' (e.g. Down syndrome, cystic fibrosis carrier screening) there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.
- (12) There should be evidence that the complete screening programme (test, diagnostic procedures, treatment / intervention) is clinically, socially and ethically acceptable to health professionals and the public.
- (13) The benefit from the screening programme should outweigh the physical and psychological harm (caused by the test, diagnostic procedures and treatments).
- (14) The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (i.e. value for money).

Implementation criteria:

- (16) All other options for managing the condition should have been considered (e.g. improving treatment, providing other services), to ensure that no more cost-effective intervention could be introduced ...
- (19) Evidence-based information, explaining the consequences of testing, investigation and treatment, should be made available to potential participants to assist them in making an informed choice.

Attitudes towards prenatal screening, or screening children or adults for carrier status or for risk of developing a late-onset condition are more engrained in local cultures. Criteria 6, 12, 13 and 19 of the UK NSC set (Box 12.4) cover ethical and social issues. Screening is an expression of the desires and values of the society in which it takes place. Antenatal screening will have little place in a society where abortion is unacceptable under any circumstances, even though the two are not necessarily linked. Some societies are more accepting of people with disabilities, or more fatalistic about abnormalities, while in others the determination to have a 'normal' child is very strong. Views on individual responsibility, and on the relative values of individual freedom and public health will affect the acceptability of specific programs. Attitudes to children are important – how far do parents own their children, and have the right to know about their genetic make-up, or how far must children's autonomy be respected, and no tests performed for later onset conditions or carrier status until the child is old enough to give properly informed consent? It is a general principle that screening should be voluntary and needs informed consent. People must in general be free to opt out of screening if they wish. This does not simply mean that they can decline testing; it also means that if, say, after declining prenatal testing they have an affected child, they should not be in any way blamed or penalized. But do parents who object to PKU screening have the right to risk condemning their child to a life of dependency on others?

In Britain, the Greek Cypriot community have taken up screening for β -thalassemia with enthusiasm, whereas the Pakistani community, also at high risk, have been slower to accept screening, although the take-up has increased in recent years. In Israel there is said to be a strong popular demand for every possible screening test, whereas in the UK the approach is much more cautious. For example, a chain of British high-street stores stopped offering a set of 'lifestyle' tests after public disquiet about the implications – even though many would argue that such tests are no more harmful than a horoscope and, at worst, only a waste of money. According to the stores, the reason for discontinuing the offer was lack of public interest which, if true, reflects better on the good sense of the British public.

Concerning the economics of screening, it is interesting to see how Criterion 14 in the UK NSC list is rather cautiously worded. Assessing the financial benefits of a screening program usually involves balancing the immediate costs of screening against claimed savings in the longer term. This is a difficult area, though of course a vital one in many areas of policy. It is handled using *discounted cash flow*. To compare a present cost with a saving in 10 years' time, the present cost is treated as an investment, whose value in 10 years is calculated using compound interest. The choice of interest rate to use (the discount rate) can have a major effect on the outcome. Also, different types of institution may have different views about the relevance of potential savings in the distant future. Moreover, these calculations have their limitations as guides to action. Applied to normal childbearing, using a high discount rate would suggest that reproduction is financially bad for society: it is unlikely that an individual's contribution to society in taxes as a working adult would repay the strongly discounted cost of his delivery, upbringing and education.

If the cost calculations look reasonable, the decision on whether to implement a screening program revolves around public demand and social acceptability. Since public opinion is never unanimous about such issues, decisions are in the end political.

Incidental findings – a form of opportunistic screening

Many clinical investigations have the potential to reveal findings that are unrelated to the primary reason for the investigation, but that may be clinically significant. A routine X-ray or MRI scan might reveal an unsuspected tumor. Although the investigation is conducted for a clinical reason and not as part of any population screening program, these extra findings do effectively constitute a form of screening. The potential for incidental findings exists in almost every clinical investigation. In clinical genetics it is particularly a concern when a patient's whole exome or whole genome is sequenced. This has often been seen as a novel problem – but in fact it is only novel in the context of mutation detection. Cytogeneticists checking a child's parents for a balanced translocation (as for the parents of **Elizabeth Elliot, Case 5**) have long lived with the possibility of discovering that the healthy father has an XYY chromosome constitution. Array-CGH study of a dysmorphic child might reveal a microdeletion that has nothing to do with the dysmorphism but renders the child at risk of cancer. Family studies of a mutation sometimes show that the mother's husband cannot be a child's father.

Arguably the extra concern surrounding genome sequencing is justified because genome sequencing has an exceptionally high potential to discover clinically relevant incidental findings. A major controversy erupted when the American College of Medical Genetics and Genomics recommended that *"laboratories performing clinical sequencing seek and report mutations in the genes listed ... [in Table 12.5]. This evaluation and reporting should be performed for all clinical germline (constitutional) exome and genome sequencing, including the "normal" of tumor-normal subtractive analyses in all subjects, irrespective of age but excluding fetal samples"* (see Green *et al.*, 2013). The list of genes and mutations was drawn up after extensive discussion as representing genes where reporting the findings to the referring clinician would likely have medical benefit for the patients and families. Only variants known or strongly suspected of being pathogenic were to be reported, but this was to be an integral part of every clinical sequence, with no option for the patient to decline it. The full report should be read to form a nuanced view of the recommendations.

Critics pointed out that some patients would have conditions that made the extra information irrelevant (elderly cancer patients, for example) and in all cases making reporting mandatory ignored patient autonomy and the right not to know. The patient's right was proposed to be limited to having the option of declining the whole procedure if they judged the risks of possible discovery of incidental findings to outweigh the benefits of the primary test. In the light of widespread criticism, the ACMG suggested the tests might be made optional, although not all members of the Working Group agreed with this concession. Meanwhile the European Society of Human Genetics recommended a much more conservative approach: whenever possible, testing should be targeted to genome regions linked to the patient's indications (de Wert *et al.*, 2020; van El *et al.*, 2013a,b). As exome and genome sequencing become ever more routine, the question of how to handle incidental findings can only become more acute. In litigious countries like the USA, the eventual policy will probably be determined by courts.

Table 12.5 – Genes for which the American College of Medical Genetics and Genomics recommended that laboratories proactively seek and report known and likely pathogenic changes

Gene(s)	Associated condition
<i>BRCA1, BRCA2</i>	Hereditary breast and ovarian cancer
<i>TP53</i>	Li–Fraumeni syndrome
<i>STK11</i>	Peutz–Jeghers syndrome
<i>MLH1, MSH2, MSH6, PMS2</i>	Lynch syndrome
<i>APC</i>	Familial adenomatous polyposis
<i>MUTYH</i>	<i>MYH</i> -associated polyposis
<i>VHL</i>	Von Hippel–Lindau syndrome
<i>MEN1</i>	Multiple endocrine neoplasia type 1
<i>RET</i>	Multiple endocrine neoplasia type 2, familial medullary thyroid cancer
<i>PTEN</i>	<i>PTEN</i> hamartoma tumor syndrome
<i>RB1</i>	Retinoblastoma
<i>SDHD, SDHAF2, SDHC, SDHB</i>	Hereditary paraganglioma–pheochromocytoma syndrome
<i>TSC1, TSC2</i>	Tuberous sclerosis complex
<i>WT1</i>	<i>WT1</i> -related Wilms tumor
<i>NF2</i>	Neurofibromatosis type 2
<i>COL3A1</i>	Ehlers–Danlos syndrome, vascular type
<i>FBN1, TGFB1, TGFB2, SMAD3, ACTA2, MYLK, MYH11</i>	Marfan syndrome, Loeys–Dietz syndromes, and familial thoracic aortic aneurysms and dissections
<i>MYBPC3, MYH7, TNNT2, TNNI3, TPM1, MYL3, ACTC1, PRKAG2, GLA, MYL2, LMNA</i>	Hypertrophic cardiomyopathy, dilated cardiomyopathy
<i>RYR2</i>	Catecholaminergic polymorphic ventricular tachycardia
<i>PKP2, DSP, DSC2, TMEM43, DSG2</i>	Arrhythmogenic right-ventricular cardiomyopathy
<i>KCNQ1, KCNH2, SCN5A</i>	Romano–Ward long QT syndrome, Brugada syndrome
<i>LDLR, APOB, PCSK9</i>	Familial hypercholesterolemia
<i>RYR1, CACNA1S</i>	Malignant hyperthermia susceptibility

They recommended this be done without seeking explicit consent whenever an exome or genome was sequenced, regardless of the indication for sequencing and the age or circumstances of the patient. See Green *et al.* (2013) for more detail.

‘Lifestyle’ genetic testing

Millions of people choose to send a saliva sample to a direct-to-consumer (DTC) genetic testing company. The company will extract the DNA and genotype it, usually for several hundred thousand SNPs using a SNP chip (see *Section 4.2* and *Figure 4.13*). They may

report on ancestry or relationships, or what concerns us here, susceptibility to diseases. The offer may cover health risks in general or it may focus on some particular aspect, producing a 'cardiac risk profile' or 'obesity risk profile', for example, in order to help you decide your most healthy lifestyle. All these analyses are based on reports of associations between particular SNP alleles and risk of specific diseases. In considering these analyses it is useful to distinguish between common and rare variants, each in the light of the ACCE framework described above.

Most analyses focus on common variants (minor allele frequency >5%). Applying the ACCE framework:

- **Analytical validity** – when properly used, SNP chips are highly reliable for genotyping common variants, with sensitivity and specificity typically >99%. In contrast to diagnostic laboratories, the laboratories performing DTC genotyping are unlikely to be enrolled in any external quality assurance scheme, but the larger and more established companies can probably be trusted to report true genotypes for common variants.
- **Clinical validity** – as described in *Chapter 13*, associations between specific common variants and diseases have been extensively investigated in genome-wide association studies (GWAS) for every imaginable common disease or phenotype. That field has moved on from its early days of unreliable and contradictory findings, and has now identified significant and well-validated associations between many hundreds of individual variants and specific phenotypes. At least for the more solid companies, the associations used are likely to have been reported by reputable scientists in peer-reviewed journals. However, associations may be specific to a certain population or ethnic group, and the great majority of all GWAS data comes from white Northern Europeans. It is important to check that any reported association applies to your ethnic group.
- **Clinical utility** – this is an extremely weak area. The common variants identified by GWAS almost always have very small effect sizes. If a variant is reported to increase your risk of some disease, always ask by how much? Does it increase your risk by a factor of 20, 5, 2 or 1.1? (Note also that odds ratios are not quite the same as relative risks – work through *Self-assessment question 4* to be clear on this.) Almost certainly the answer will be much nearer 1.1 than 20. Do you care about a 10% increase in your risk? It is important to distinguish between the statistical significance of an effect and its size. A large study can identify a highly significant association (that is, there is no doubt that it is real) with a quite trivial odds ratio. Note, incidentally, that it is well known that initial odds ratios almost always come down in subsequent studies, even when the risk is confirmed. Remember also the difference between relative and absolute risks, discussed above. Increasingly, companies may report polygenic risk scores, based on putting together the effects of variants across your genome. This is a promising development, but currently very immature. Polygenic risk scores are specific to ethnic groups; one quoted for you will only be valid if the underlying studies were all performed in your ethnic group (almost always white European). Also ask whether any prediction takes your family history and your own clinical data into account? Despite the technical success of GWAS, for almost every disease all reported DNA variants collectively account for only a small proportion of the overall heritability. Family history is at least as relevant as SNP genotypes.

- **Ethical matters** – you may wish to check how stringently a company guards your privacy and respects your ownership of your data.

The above was for common variants. There are additional concerns for rare variants, and these concerns become more important the rarer a variant is.

- **Analytical validity** – there is a major issue here. SNP chips are highly reliable for typing common variants, but they perform increasingly poorly as a variant gets rarer. This is not a problem of poor laboratory practice, it is intrinsic to SNP chip technology (see Weedon *et al.*, 2019). In that study, covering 50 000 individuals in the UK Biobank, for variants with a frequency <0.001% only 16% of heterozygous genotypes reported by the SNP chip analysis could be confirmed by sequencing. The chips generated many false positives: there were 425 pathogenic *BRCA1/2* variants reported in 889 individuals; of these, just 17 variants were confirmed by sequencing. There were also many false negatives: a further 43 pathogenic *BRCA1/2* variants were present in the sequencing data from the study group but were not detected by the SNP chip despite being assayed.
- **Clinical validity** – unlike the common variants described above, with their very small effect sizes, rare variants may be strongly pathogenic. The class of variants with frequency <0.001% includes most of those responsible for most mendelian diseases. That is why the genotyping problems described here are so important. Assuming a correct genotype, it is then crucial to know how strong is the evidence that it is pathogenic. It is well known that mutation databases contain many harmless variants that have been wrongly described as pathogenic. Gradually these are being weeded out; the ClinVar database is the best starting point for checking a variant, but when important decisions hang on the interpretation it is mandatory to do thorough research.
- **Clinical utility** – correctly identified rare pathogenic variants have high clinical utility.
- **Ethical matters** – the better DTC companies, perhaps appreciating the problems with rare variants, do not report them. However, customers are entitled to download their full raw data, and they can then submit this for analysis to third-party companies. It is hard to see any ethical framework within which supposed pathogenic rare variants detected by SNP chips should be reported to naïve customers, given that the great majority will be false positives.

As the cost of DNA sequencing continues to tumble, companies will increasingly offer whole exome or whole genome sequencing, rather than SNP genotyping. The paper by Ashley *et al.* (2010) gives a foretaste. Sequencing can reliably genotype rare variants, avoiding the problem of genotyping rare variants on SNP chips; because sequencing has the potential to reveal more significant risks than those revealed by SNP genotyping it would be crucial to be confident that a DTC laboratory was operating to the same standard as a properly certified diagnostic laboratory. Meanwhile, it would be prudent to read the paper by Janssens and van Duijn (2008) before parting with your money. Horton and colleagues (2019) offer advice to general practitioners on how to talk to clients who have taken, or are intending to take, a DTC genetic test.

Assuming a test is satisfactory by all these criteria, the question remains, so what? From a public health point of view, there is a risk that 'lifestyle' genetic testing may simply dilute universally valid messages about the need to eat sensibly, get some exercise and stop smoking. Companies looking at market opportunities may be less bothered by such

thoughts – but no company could risk saying to people who come out as low risk on their test that they can therefore cheerfully indulge in an unhealthy lifestyle. The lawyers would have a field day. A person's reported risk may be below the population average, but it is not zero. Some people given a low risk for a particular disease will nevertheless succumb to that disease. Therefore, quite regardless of the test result, the advice from the company must be the same routine healthy living advice, just delivered a bit more emphatically for high-risk people. It might therefore be argued that such testing has no ethical implications. The same advice may be available to everybody free of charge, but some people will only take notice of it if they have paid \$250 and received a scientific-looking report.

Regulators have started taking an interest in the burgeoning market for DTC genetic tests because of concerns about possible harmful effects. If somebody using a DTC ancestry company thereby discovers that their mother's husband cannot be their father, it would be hard to blame the company for the resulting fallout. But maybe companies giving health-sensitive results should carry the full costs of their business, and not expect publically funded services to provide any necessary counseling and diagnostic confirmation. In the USA the Food and Drug Administration (FDA) ordered the DTC testing company 23andMe to cease marketing its 'Personal Genome Service' without marketing clearance or approval. Clearly the genie cannot be put back in the bottle, and people have a right to know about their own genome; just, maybe they should not expect the taxpayer to support them in coping with the consequences of their own choices.

Non-invasive prenatal testing

NIPT is based on detecting cell-free DNA (cfDNA) derived from the placenta in the blood of pregnant women (Bianchi and Chiu, 2018). It can be used as a screening test for trisomy 21, 18 and 13 (Down, Edwards and Patau syndromes, respectively), as a diagnostic test to determine fetal sex, and for specific single gene disorders where the risk is known to be high. As a pregnancy progresses the amount of placental cfDNA increases in the mother's blood but falls rapidly after delivery, so it is specific to that pregnancy. From about 9 weeks of pregnancy there is usually enough cfDNA of placental origin in the woman's blood to get accurate results. There are many advantages of NIPT over other prenatal screening and diagnostic approaches: NIPT has no risk of miscarriage and is less invasive and time consuming than other prenatal tests. A report by the UK Royal College of Obstetricians and Gynaecologists provides a detailed discussion of the method and its likely implications (RCOG, 2014). Ethical questions are discussed by the Nuffield Council on Bioethics (Nuffield Council on Bioethics, 2017).

NIPT as a screening test for trisomy 21, 13 and 18

A trial (Bianchi *et al.*, 2014) showed cfDNA had significantly lower false positive rates and higher positive predictive values for detection of trisomies 21 and 18 than standard screening (serum biochemical assays with or without nuchal translucency measurement). As a screening test for common trisomies it is more accurate and less gestation-sensitive. Confirmatory invasive tests are usually recommended, but these are in fewer numbers than those occurring in longer established screening programs.

Analysis of circulating cfDNA in maternal blood can be carried out in two main ways. Maternal and fetal cfDNA molecules can be randomly sampled, sequenced and mapped to specific chromosomes. The numbers of DNA molecules matched to different human chromosomes are then counted. Where a pregnancy is affected with trisomy 21, 18 or 13, the proportion of cfDNA molecules derived from those chromosomes is higher than that in a reference data set based on samples from pregnant women carrying fetuses without trisomy (but not 50% higher, because the great majority of all the

DNA in the sample will be maternal). Alternatively, single-nucleotide polymorphisms (SNPs) on the chromosomes of interest are amplified and sequenced. Ratios between heterozygous SNP alleles are compared with those of other targeted chromosomes. When there is aneuploidy of a targeted chromosome sequenced, skewing of the ratios is expected.

As with other screening strategies there are false positives and false negatives. False positives may result from confined placental mosaicism for the trisomy although the fetus is normal, from an unrecognized twin which has died and been absorbed, and occasionally from a genetic disorder in the mother. A major advantage of NIPT for trisomies is that the false negative rate is extremely low. Women can be strongly reassured after a negative test. The main cause of rare false negative results is that there was insufficient cfDNA of placental origin in the sample of maternal blood.

The prior risk of a trisomy affects how reliably a 'positive' NIPT test can be interpreted. For those at high prior risk, positive screening tests result in a confirmed diagnosis in over 90% of cases, but in general population studies the rate of confirmed diagnoses is lower. This has resulted in the recommendation by most professional bodies for a confirmatory diagnostic test such as amniocentesis and direct chromosome analysis in those with a 'screen positive' result. This has also informed national screening programs in a number of countries, where only those identified as at higher risk by standard prenatal screening are offered NIPT. The number of chromosomal aberrations that can be reliably screened for is increasing as techniques and experience improve, so that cfDNA testing for sex chromosome anomalies and copy number variants is offered by some programs.

NIPT as a diagnostic test for single gene disorders where the risk is high

In couples where there is a known high risk for a single gene disorder because one parent is affected or there is a significant family history, or where other prenatal tests strongly suggest a specific syndrome, analysis of cfDNA in maternal blood can be used diagnostically (Zhang *et al.*, 2019). When a condition is *de novo*, or where the father is the parent affected by a genetic disorder, testing is more straightforward and variant DNA sequences can be sought in maternal blood. If the disorder for which the pregnancy is at high risk is X-linked, testing for Y chromosomal DNA sequences is undertaken to sex the fetus. Where the mother is affected or where an autosomal recessive disorder is concerned, cfDNA methods are used to assess whether there is proportionally more or less of the maternal mutant allele or haplotype in comparison with its normal allele. The allele or haplotype that is proportionately greater is the one the fetus has inherited.

12.5. References

- Ashley EA, Butte AJ, Wheeler MT, *et al.*** (2010) Clinical assessment incorporating a personal genome. *Lancet*, **375**: 1525–1535.
- Bianchi DW and Chiu RWK** (2018) Sequencing of circulating cell-free DNA during pregnancy. *New Engl. J. Med.* **379**: 464–473.
- Bianchi DW, Parker RL, Wentworth J, *et al.*** (2014) DNA sequencing versus standard prenatal aneuploidy screening. *New Engl. J. Med.* **370**: 799–808.
- de Wert G, Dondorp W, Clarke A, *et al.*** (2020) Opportunistic genomic screening. Recommendations of the European Society of Human Genetics. *Eur. J. Hum. Genet.* in press.
- Green RC, Korf BR, Grody WW, *et al.*** (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics Med.* **15**: 565–574.

- Horton R, Crawford G, Freeman L, et al.** (2019) Direct-to-consumer genetic testing. *Br. Med. J.* **367**: l5688.
- Janssens ACJW and van Duijn CM** (2008) Genome-based prediction of common diseases: advances and prospects. *Hum. Molec. Genet.* **17**: R166–R173.
- Louter L, Defesche J and van Lennep JR** (2017) Cascade screening for familial hypercholesterolemia: practical consequences. *Atherosclerosis Suppl.* **30**: 77e85.
- Malone FD, Canick JA, Ball RH, et al.** (2005) First-trimester or second-trimester screening, or both, for Down's syndrome. *New Engl. J. Med.* **353**: 2001–2011.
- Nuffield Council on Bioethics** (2017) Non-invasive prenatal testing: ethical issues. Nuffield Council on Bioethics.
- Ogilvie CM, Lashwood A, Chitty L, Waters J, Scriven PN and Flint F** (2005) The future of prenatal diagnosis: rapid testing or full karyotype? An audit of chromosome abnormalities and pregnancy outcomes for women referred for Down syndrome testing. *Br. J. Obstet. Gynaecol.* **112**: 1369–1375.
- RCOG** (2014) Non-invasive prenatal testing for chromosomal abnormality using maternal plasma DNA. Scientific Impact Paper 15. www.rcog.org.uk/en/guidelines-research-services/guidelines/sip15/ [accessed 26 June 2020].
- Sturm AC, Knowles JW, Gidding SS, et al.** (2018) Clinical genetic testing for familial hypercholesterolemia: JACC Scientific Expert Panel. *J. Am. Coll. Cardiol.* **72**: 662–680.
- van El C, Cornel MC, Borry P, et al.** (2013a) Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur. J. Hum. Genet.* **21**: 580–584.
- van El CG, Dondorp WJ, de Wert GMWR and Cornel MC** (2013b) Call for prudence in whole-genome testing (letter). *Science*, **341**: 958.
- Weedon MN, Jackson L, Harrison JW, et al.** (2019) Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing. www.biorxiv.org/content/10.1101/696799v2. Note that this is a preprint that has not been peer-reviewed; broadly similar conclusions have been reported by Van Hout et al., (2020) *Nature*, **586**: 749–757.
- Wilson JMG and Jungner S** (1968) *Principles and Practice of Screening for Disease*. World Health Organization, Geneva.
- Wright CF, West B, Tuke M, et al.** (2019) Assessing the pathogenicity, penetrance and expressivity of putative disease-causing variants in a population setting. *Am. J. Hum. Genet.* **104**: 275–286.
- Zhang J, Li J, Saucier JB, et al.** (2019) Non-invasive prenatal sequencing for multiple Mendelian monogenic disorders using circulating cell-free fetal DNA. *Nature Med.* **25**: 439–447.

Useful websites

For information on newborn screening in the USA: www.hrsa.gov/advisory-committees/heritable-disorders/rusp/index.html.

For information on screening programs in the UK: www.gov.uk/guidance/newborn-blood-spot-screening-programme-overview.

The Brown–Goldstein lab website at UT Southwestern Medical Center tells a fascinating story of their Nobel Prize-winning work on cholesterol regulation:

www4.utsouthwestern.edu/moleculargenetics/pages/brown/past.html

12.6. Self-assessment questions

- (1) A hypothetical disease is caused by mutations in the *IGNO* gene. 1 person in 100 carries a mutation. You have a genetic testing protocol that detects 80% of all mutations. You receive 10 000 blood samples from newborn babies – but 1% of the samples were taken into contaminated tubes that give a false positive result on your test. What is the positive predictive value of your test?
- (2) A scientist has been studying people who suffer severe adverse effects of a certain drug. In the general population one person in 10 000 suffers these effects. He has identified a DNA polymorphism that is strongly associated with the risk. In blind testing in the laboratory, 99 out of 100 people who had shown the adverse drug reaction tested positive for the variant, while only 1 out of 100 people who had taken the drug without ill-effects tested positive. He proposes to screen the entire 1 million population of his city for the variant. Calculate the positive predictive value of his test.
- (3) Repeat the calculation of the previous question, assuming the adverse reaction occurred in 1 person in 10 rather than 1 in 10 000. What does this tell us about the potential of screening in general?
- (4) Two DNA variants are each associated with a 50% increase in the chance of having a certain disease. For each variant, draw up a 2×2 table as in Box 12.1, with numbers from testing 1000 cases and 1000 controls and calculate the odds ratio, assuming variant A is present in 50% of the normal population and variant B is present in only 5%.
- (5) Assuming an incidence of neural tube defect of 1:100 pregnancies tested, use the curves in Figure 12.4 to estimate the sensitivity and positive predictive value of the maternal serum AFP test, using cut-offs of:
 - (a) everybody above the normal mean value
 - (b) everybody above the minimum abnormal value
 - (c) everybody above the maximum normal value
- (6) In cystic fibrosis carrier screening, some couples will have one partner test positive and the other negative. Calculate the sensitivity required of a screening test so as to ensure that such couples are at no greater risk of really both being carriers than they were before any screening was done. Assume the carrier frequency in this population is 1 in 40.
- (7) You are a health administrator charged with establishing a population screening program for carriers of cystic fibrosis (who make up 1 person in 25 in your population).
 - (a) Decide at what stage in their life and under what circumstances people should be tested; write a brief justification of your choice.
 - (b) You have proposals from two companies for performing the high-throughput genotyping. One offers to test a limited panel of mutations, covering those present in 70% of all carriers; the other tests a larger panel covering 90% of all carriers. Naturally the costs are different, but before you can make decisions you need to know the likely results. For each option calculate the expected outcomes from screening 1 million people, in terms of CF births avoided and CF births to couples who were not both identified as carriers on the screening test.

[Hints on questions 4, 5 and 6 are provided in the *Guidance* section at the back of the book.]

13

Should we be testing for genetic susceptibility to common diseases?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Describe the multifactorial nature of most human traits, both normal and abnormal, and the principles of multifactorial inheritance
- Explain how the heritability of a character can be estimated, and the uses and limitations of the concept of heritability
- Describe the process and achievements of genome-wide association studies
- Discuss the 'missing heritability' problem and possible solutions
- Discuss critically the generation and application of polygenic risk scores
- Discuss the present and future prospects for testing healthy people to identify their genetic susceptibilities to common complex diseases

13.1. Case studies

CASE 25 YAMOMOTO FAMILY

- Family history of dementia
- Alzheimer disease
- Test for ApoE4?

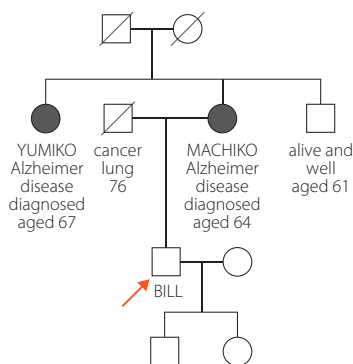


Figure 13.1 – Pedigree of the Yamamoto family.

333 344 395

Bill Yamamoto's mother, Machiko, had been getting increasingly forgetful. After a series of incidents, including one when she left a pan on the cooker and set the kitchen on fire, it became clear that she was not coping with life on her own. She moved into sheltered accommodation in the same Californian town as Bill and his wife. She never adjusted to her new surroundings and soon required residential care.

Over the next 3 years her dementia progressed until she seldom recognized Bill and was unable to do anything for herself. It was almost a relief when she died at the age of 71 – as Bill said, 'my real mother died several years ago'.

When a friend of his wife said that Alzheimer disease was hereditary, Bill started to worry. He knew that his aunt Yumiko – his mother's sister who had always lived in the old family home in Hawaii – had been diagnosed with Alzheimer disease at the age of 67. His wife suggested he should talk to his physician. The doctor told him that only the rare forms of the disease with onset before age 60 were inherited. Bill was still not entirely reassured.

Searching on the internet, he discovered that a genetic factor, ApoE4, was associated with susceptibility to the common late-onset form of the disease, and he came across companies offering to test for ApoE4. He wondered whether he should take the test, and decided to

CASE 26 ZUABI FAMILY

- Zafira, woman aged 52 years
- Overweight, sedentary lifestyle, insatiable thirst
- Type 2 diabetes
- Son's lifestyle and heredity put him at high risk
- Management of family

334

349

395

Zafira was 52 when she consulted her physician about a 3-month history of dizziness, headaches and blurry vision. She also mentioned that she had developed a terrible thirst, leading her to drink large amounts of water every day and produce correspondingly large amounts of urine. Her urine contained sugar, and a fasting glucose test showed a level of 9 mmol/l. This confirmed the diagnosis of Type 2 diabetes (T2D). She was given a thiazolidinedione drug as the first line of treatment. Enquiry showed that she had an entirely sedentary lifestyle, and her body mass index (BMI) was 32; she was enrolled on a program of graded moderate exercise. Both exercise itself and weight loss help minimize morbidity in T2D.

The shock of diagnosis made Zafira think about her family. Her brother had died of a heart attack aged 48. He had been badly overweight and completely inactive, though she did not remember him drinking especially large amounts of water. When she learned that first-degree relatives of T2D patients were at high risk of developing the condition (her physician quoted a risk of 38% for a child of an affected parent), her thoughts turned to her elder son Zahid. He did not have overt disease, but he shared several risk factors. He went to work by car, took the lift to his office, spent all day sitting at his desk, enjoyed a good meal then spent the evening watching television. Not surprisingly he was overweight. He agreed to come for some tests. These showed that his BMI was 30 with a waist measurement of 99 cm. His fasting plasma glucose was 6.4 mmol/l, below the 7.0 threshold for T2D, but above normal. He was hypertensive (blood pressure 142/90 mmHg) and had dyslipidemia (triglycerides 1.9 mmol/l). His combination of obesity, impaired glucose tolerance, dyslipidemia and hypertension defined him as having the ‘metabolic syndrome’ (Table 13.1). This is a rather loosely defined but well-known precursor of T2D and a major risk factor for cardiovascular disease (reviewed by Eckel *et al.*, 2005). He was prescribed exercise and antihypertensive drugs.

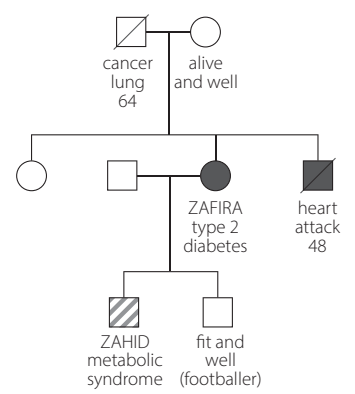


Figure 13.2 – Pedigree of the Zuabi family.

Table 13.1 – World Health Organization definition of the metabolic syndrome (1999)

Feature	Measure
Diabetes	
Or impaired fasting glycemia	Fasting glucose >7.0 mmol/l
Or impaired glucose tolerance	Fasting glucose 6.1–7.0 mmol/l
Or insulin resistance	Hyperinsulinemic
PLUS 2 or more of:	
Obesity	BMI >30 or waist-to-hip ratio >0.9 (male) or 0.85 (female) (different figures may be appropriate for non-white people)
Dyslipidemia	Triglycerides ≥1.7 mmol/l or HDL cholesterol (male) <0.9 or (female) <1.0 mmol/l
Hypertension	>140/90 mmHg
Microalbuminuria	Albumin excretion >20 µg/min

Other bodies have produced somewhat different but overlapping definitions. The list shows the complex of interacting characters that indicate susceptibility to T2D and cardiovascular disease. The prevalence of the syndrome rises with age, among US adults from 7% at age 20–29 to 44% at age 60–69. Prevalence is 50% in severely obese US youngsters, and is increasing at all ages and in most countries of the world.

13.2. Science toolkit

So far in this book we have been mainly concerned with rare inherited diseases, mostly with neonatal or childhood onset, where genetic variants – variants at the DNA sequence or chromosomal level – determine whether or not somebody suffers from the disease. But the major impact of genetic variation on risk of disease (other than cancer) is on the risk of the many common conditions like diabetes or Alzheimer disease, most of which are of adult onset. The role of genetic variants in these common diseases is subtle, and very different from their role in conditions like cystic fibrosis or Huntington disease.

No single variant causes the disease. Instead, a multitude of variants act independently to increase or decrease the risk that somebody will develop a condition. Environmental and lifestyle factors, and maybe simple chance, play a part, often a large part, in determining whether or not somebody will get the disease. These conditions do not give typical autosomal dominant or recessive pedigree patterns. They are non-mendelian; they are termed **multifactorial**, as distinct from mendelian or monogenic characters. They may nevertheless run in families to some extent. Relatives share genes, and so they may share variants that predispose to disease. The λ statistic describes the ratio of incidence of the condition in a relative of an affected person compared to the incidence in the general population. Different λ values can be calculated for different degrees of relationship, e.g. λ_s for sibs. Although multifactorial conditions can run in families, they do so to a much lower extent than mendelian conditions (*Table 13.2*).

Table 13.2 – Examples of risks for mendelian and multifactorial diseases

Disease	λ_s	Lifetime risk (to age 80)
Huntington disease	5000	0.01%
Cystic fibrosis	500	0.05%
Type 1 diabetes	18	1%
Type 2 diabetes	3	15%
Late-onset Alzheimer disease	3	17%
Celiac disease	10	1%
Multiple sclerosis	6	0.5%
Breast cancer	2	12%

λ_s is the relative risk for a sib of an affected proband, compared to an unrelated person. Compare the risks for the mendelian conditions (shaded rows) with the complex conditions. These latter are **empiric risks**, derived from surveys of families, not from theoretical calculations. They can vary between populations and also over time (the latter presumably because of environmental changes), so these figures are illustrative only.

The effect of previous history on recurrence risk is significantly different in these conditions compared to mendelian conditions. If a healthy couple have a child with cystic fibrosis, they must both be heterozygous carriers, and the recurrence risk is 1 in 4. It is still 1 in 4 even if they have had the misfortune to have three affected children. But for multifactorial conditions, the worse the previous family history, the greater is the empirical recurrence risk. The recurrence risk is higher for a couple who have three affected children than for a couple who have only one. If a condition is more frequent in males than females, the

recurrence risk is greater after birth of a female affected child. This is not because previous misfortunes increase the risk; it is because an unfortunate history alerts us to the fact that these people probably always did have a particularly high risk.

When studying the genetics of these common disorders we are looking for genetic susceptibility factors, not causative variants. Hopefully understanding the genetic susceptibility will complement epidemiological studies and lead to better understanding of why some people develop a condition while others do not, leading in turn to better prevention and maybe better treatment. A first task is to determine the overall role of genetic factors in determining susceptibility.

Estimating heritability

The **heritability** of a condition is the proportion of overall susceptibility that is due to genetic factors. It is a number between 0 (no genetic involvement) and 1 or 100% (no involvement of non-genetic factors). Estimates of heritability are central to the efforts of animal or plant breeders to improve stock by selective breeding. In human genetics heritability is rather a slippery concept. It is often misunderstood, either accidentally or willfully by people pushing a political goal. Visscher *et al.* (2008) review the meaning of this measure, and its uses and abuses.

Heritability is not a fixed property of a condition, like the mode of inheritance; it is a statement of the role of genetic differences in a certain population at a certain time. If a condition is influenced by social conditions such as deprivation or poverty, then its heritability will be higher in more equal societies, because more of the social variation will have been eliminated, so that genetic factors play a larger part in the remaining susceptibility. Equally, just because a condition has high heritability, that does not mean its prevalence cannot be reduced by social or environmental interventions; it just means that none of the current social or environmental variation in that society at that time has a major influence on the prevalence. It says nothing about the potential to reduce the incidence by some novel or unusual intervention.

Despite these reservations, estimating the heritability of a condition is an important step in investigating the causes of disease. Diseases are not abstract entities, they are things that happen to particular individuals in a particular society at a particular time, and before we dive into investigating the genetics, we need to know how much genetics there is to find. There are three main ways of estimating heritability of human conditions: family, twin and adoption studies.

- **Family studies** compare the incidence of a condition among relatives of an affected person to the incidence in the general population. The ratio is symbolized as λ , as in *Table 13.2*. The relatives of an affected person share genes with them (see *Box 9.3*). We can estimate heritability by comparing the extent to which relatives share genes with the extent to which they share a disease. However, a big problem with this approach is that relatives – especially parents, children or siblings – usually share features of their environment as well as their genes. If we do not take shared family environment into account, we will overestimate the role of shared genes. It is widely felt that many published estimates of heritability are rather too high because insufficient allowance was made for the shared environment.

- **Twin studies** start by ascertaining people who have the condition and are one of a twin pair. They compare how often the co-twin is affected (the **twin concordance**) depending on whether the pair are monozygotic or dizygotic. Monozygotic twins are genetically identical clones; dizygotic twins share on average half their genes, just like any other pair of sibs. The method rests on the assumption that twins should share their environments to the same extent regardless of their zygosity – which may not always be true. Identical twins may be more likely to be dressed and treated the same. MZ twins separated at birth and brought up in different environments form the perfect study, but their numbers are too small to provide anything more than fascinating anecdotes.
- **Adoption studies** seem a powerful way to separate shared genetics from shared family environment. One design is to ascertain people who have the condition of interest and who have been adopted in infancy, and to ask whether the condition runs in their biological or their adoptive family. Alternatively, one might seek people who have the condition and whose children have been adopted away, and ask whether being adopted away has saved the children from the parent's condition. The main problem with these designs is getting adequate data about the biological family; additionally, adoption agencies may seek to place children in adoptive families that resemble the biological family as closely as possible.

Each of these approaches has its limitations (*Figure 13.3*), but overall they have highlighted the large role of genetic factors in the causation of many common conditions – a finding of particular importance for psychiatric conditions like schizophrenia, where ‘nature–nurture’ arguments have been very contentious. Having decided that there are indeed genetic factors, it now remains to find them.

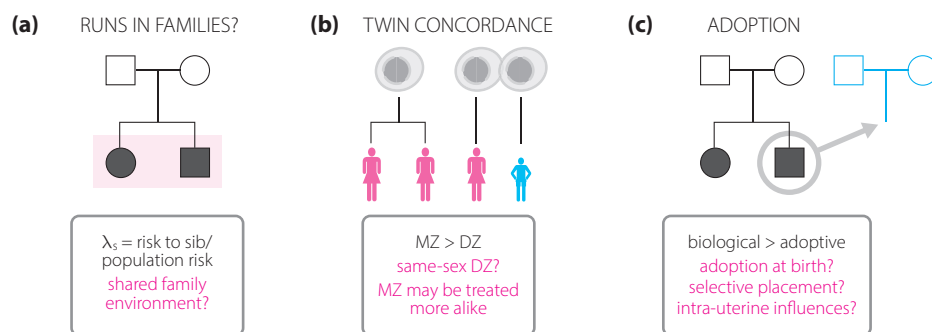


Figure 13.3 – Methods of demonstrating genetic effects in complex diseases, and (in pink) reasons for caution in interpreting the results.

Identifying genetic susceptibility factors

Early attempts to identify genetic susceptibility factors for common diseases used the same family linkage approaches, as described in *Chapter 8*, that had been so successful at identifying the causes of monogenic diseases. These were almost uniformly unsuccessful. Few susceptibility factors were identified, and the few positive identifications often could not be replicated in confirmatory studies. The basic problem was that the deterministic statistical models underlying classical linkage analysis do not fit the more

complex genetics of multifactorial diseases, where no single variant is either necessary or sufficient for the disease to develop. A related method (non-parametric linkage) used pairs of affected relatives, usually sibs, and compared the extent to which they shared alleles at a given chromosomal location with the extent to which they shared the disease. Increased sharing above the random mendelian expectation suggested that there was a susceptibility locus at that chromosomal location. Again, successful identifications were few and far between, and often could not be replicated. The problem here was that although the method was statistically impeccable, it had very low power, so that impossibly large numbers of relative pairs would have been needed to get reliable results.

Success came with a move from linkage-based to association-based methods (*Box 13.1*).

Linkage versus association

The difference hinges on whether *loci* or *alleles* are being considered.

- **Linkage** is a relationship between *loci*. It is a specifically genetic phenomenon. A marker *locus* is linked to the disease *locus*. It does not depend on which allele (disease / normal, different marker alleles) is present at either locus. Loci are linked because they lie close together on a chromosome.
- **Association** (in the present context) is a relationship between *phenotypes* and/or *alleles*. A particular marker *allele*, not the marker *locus*, is associated with a disease. The association is with the disease (the phenotype) not with a disease susceptibility *locus* that might have high-risk and low-risk alleles.

Association is a purely statistical phenomenon, and not specifically genetic. An association between a genetic variant and a disease may have several causes:

- the variant may directly confer susceptibility to the disease
- the variant may be on the same shared ancestral chromosome segment as a variant that directly causes susceptibility (see below)
- if the population studied is not homogeneous with random mating (see *Section 9.2*), but contains subgroups that are relatively isolated from one another, the variant may happen to be more frequent in a subgroup in which the disease is also more frequent, for some unrelated reason. Such **population stratification** must be guarded against in association studies.

Linkage does not imply a population-wide association. At a locus that is linked to a disease locus, the particular alleles may be different in different unrelated families. Within a family the same allele would be associated with the disease, but between unrelated families there would be no overall association.

BOX 13.1

Genome-wide association studies

After several false starts with underpowered studies, the Wellcome Trust Case–Control Consortium (WTCCC, 2007) set the template for successful genome-wide association studies (GWAS). The principle is very simple. A panel of people with the disease in question (the cases) and a matched panel of unaffected people (the controls) are collected. Both panels are genotyped for common genetic variants spread across the whole genome, and alleles are sought that are significantly more frequent in cases than in controls. The success of this and all subsequent GWAS depended on three developments.

- The development of high-resolution SNP chips that allow a person to be genotyped for up to 1 million SNPs in a single operation. As described in *Section 4.2*, a SNP chip is a microarray where different cells contain anchored oligonucleotide probes specific for the different alleles of a single nucleotide polymorphism.
- The recognition by researchers and funding agencies that successful studies require large sample sizes. This has prompted the formation of consortia able to recruit and genotype a thousand or more cases and controls for each disease. More recently, statistical methods have been developed to combine data from independent studies of the same condition, allowing meta-analysis of huge numbers (100 000 or more) of cases and controls.
- The realization that very stringent controls on data quality are necessary to avoid false positive results. Samples and genotypes to which the slightest suspicion could attach must be rejected, and the cases and controls must be very carefully matched, as slight inadvertent differences between the two groups can easily introduce spurious associations. Many populations do exhibit fine-scale structure, as expected since local people are more likely to be part of the same extended family. Statistical tests are used to check that cases and controls really do match.

The results of GWAS are typically displayed in a so-called Manhattan plot (*Figure 13.4*).

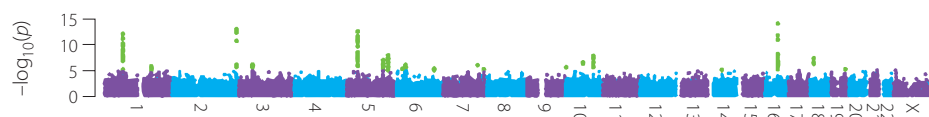


Figure 13.4 – A Manhattan plot displays the results of a GWAS.

The X axis shows the chromosomal location of each marker, the Y axis the p -value (as $-\log(p)$) for its association with the condition being studied. Data for markers that pass the threshold of significance are shown as green dots; data for the remaining markers are in blue, but because of the very large number of markers with non-significant associations, the blue dots mostly coalesce. The figure shows data from the Wellcome Trust Case-Control Consortium (2007) for Crohn disease. Reproduced from *Nature*, with permission from Springer Nature, © 2007.

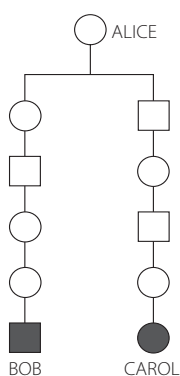


Figure 13.5 – Bob and Carol both have Type 2 diabetes, in part because they each inherited a susceptibility allele from their great-great-grandmother Alice.

Because GWAS look at very large numbers of markers, a stringent threshold of significance is needed in order to avoid false positive results. In a test of random non-existent associations, 5% will be significant at the $p=0.05$ level and 1% at $p=0.01$. To avoid false positives, the $p=0.05$ threshold is divided by the number of independent questions asked (the so-called Bonferroni correction). A modern GWAS might check 1 million markers, so the threshold of significance is $p=5 \times 10^{-8}$ (though the WTCCC used a threshold of 5×10^{-7}).

GWAS are based on the realization that our genomes are a mosaic of shared ancestral chromosomal segments. Consider the case of Bob and Carol (*Figure 13.5*) who both have Type 2 diabetes. In part this is because they both inherited a susceptibility allele on chromosome 9 from their shared great-great-grandmother Alice. Bob and Carol probably did not know they were related. They each have 32 great-great-grandparents. Even if they were both enthusiastic family historians, it is unlikely they would know about all 32 of them and would have identified all their many descendants.

They will share the segment of chromosome 9 that carries Alice's susceptibility allele, and so they will also each have the same allele for the various non-pathogenic SNPs that are located on that segment. It will be quite a small segment. During prophase I of meiosis there are an average of 60 crossovers per cell in males and 90 in females (though the figure varies widely, both between individuals and between different gametes from the same individual). An average chromosome might be split into three to six segments. Most recombination events take place at a limited (though large) number of recombination hotspots, so that the small segments between adjacent hotspots have a significant chance of being transmitted intact over many generations. But a segment that has a 90% probability of surviving one meiosis without being broken up by recombination would have only a $(0.9)^{10} = 0.35$ chance of remaining intact through the ten meioses that separate Bob and Carol.

Extending this example, N generations ago each of us had 2^N ancestors (*Figure 13.6*). Each ancestor in turn would have on average 2^N descendants (assuming, for simplicity, that the population size remained constant). Going back even 20 generations (say, 500 years, to the year 1500), 2^{20} is over 1 million. Ultimately we are all related. The word 'unrelated' is used in this book to mean people who do not share any great grandparent and are unaware of any other common ancestor. People who thought themselves unrelated nevertheless share small chromosome segments that are inherited from distant common ancestors. The more distant the ancestor, the smaller each shared segment will be, but the larger will be the number of people who share it.

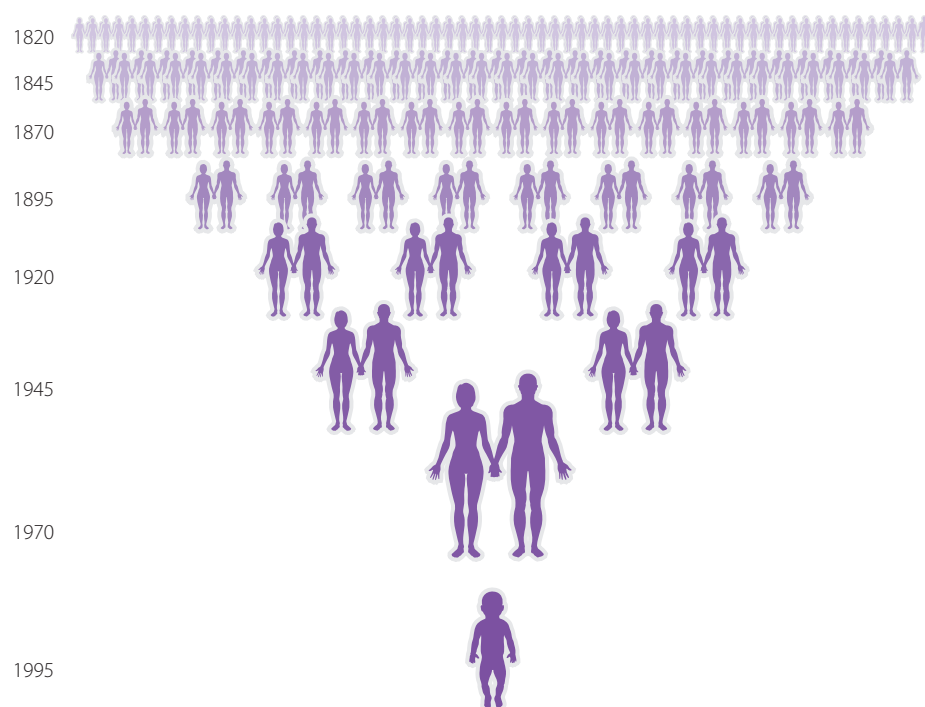


Figure 13.6 – N generations ago a person had 2^N ancestors.

This number quickly exceeds the population of most countries a few hundred years ago. Even allowing for inbreeding and the isolation of communities, we are all related through distant common ancestors. For a more nuanced treatment see Ralph and Coop (2013) and <https://gcbias.org/european-genealogy-faq/>.

Suppose one such ancestral segment contained an allele conferring susceptibility to diabetes. A collection of present-day 'unrelated' diabetics would tend to share that chromosome segment. The allele is neither necessary nor sufficient to cause diabetes, so not every diabetic will have this segment, and some non-diabetic people will have it – but insofar as having the allele increases susceptibility, affected people are more likely than unaffected people to have that ancestral segment. Along with the susceptibility allele, they will share alleles of the SNPs that mark that segment. Thus searching for shared SNP alleles can identify chromosomal segments carrying shared ancestral susceptibility factors.

How many SNPs would one need to test to obtain genome-wide coverage? Around one nucleotide in 300 is polymorphic, meaning that there might be 10 million SNPs available to genotype. However, the structure of our genomes as mosaics of shared ancient segments greatly reduces the problem. Starting with the HapMap project in the early 2000s (International HapMap Consortium, 2003), large-scale SNP genotyping projects have identified these shared ancestral chromosomal segments as conserved haplotype blocks. The blocks vary in size, but average around 5 kb. The precise number, size and identity of blocks depends on the statistical criteria adopted to define a block, but the overall structure is clear. The key finding from large-scale population genotyping is that at most chromosomal locations most genomes have one of only 3–5 different ancestral blocks. That does not mean that most of us are descended from only 3–5 different cavemen. At the next-door block, again most genomes may have one of 3–5 ancestral blocks, but they will be inherited from a different 3–5 remote ancestors. Our remote ancestry is with populations, not individuals.

The block structure, and the fact that we are all related, greatly reduces the amount of possible genetic diversity among humans. If there are 10 million common bi-allelic SNPs, in principle there are 2 to the power of 10 million possible genotypes – an unimaginably large number. The block structure reduces the GWAS problem to identifying which of the 3–5 common ancestral blocks a person has at each chromosomal location. Around 1 million carefully selected SNPs ('tag-SNPs') are enough to do this. Within a block all the variants go together – they are said to be in **linkage disequilibrium**. Not every person has one of the common blocks, and not every genomic location has this simple block structure. Nevertheless, genotyping 1 million tag-SNPs will identify most of the common genetic variation in a population. Layered on top of this common ancestral variation there will be many rare variants with more recent origins that may be significant susceptibility factors for some individuals, but will not be detected by GWAS that genotype only common variants.

Over the years since the WTCCC, genome-wide association studies have proliferated. Virtually every common disease or phenotype one can think of has been investigated by GWAS. Many thousands (over 100 000 according to Shendure *et al.*, 2019) of unique, robust associations between common variants and common diseases have been identified. A database of all reported associations between diseases and genetic markers is maintained by the US National Human Genome Research Institute (see www.genome.gov/gwastudies). Figure 13.7 gives an impression of the richness of the data.

GWAS identify associations between common genetic variants and diseases. How common is 'common' depends on the sample size. The WTCCC tested 2000 cases for each of seven conditions, plus 3000 unaffected controls, using SNPs with a minor allele



Figure 13.7 – An overview of genetic susceptibility factors identified by genome-wide association studies.

The diagram gives an impression of the range of conditions explored by GWAS over the period 2006–2014, and the many susceptibility loci identified. More recent data have been omitted in the interests of clarity. Data from *Welter et al. (2014)* and available, along with full current data, at www.ebi.ac.uk/fgpt/gwas/ [data for this figure accessed February 2015].

frequency (MAF) of at least 0.05. More recent studies, with larger sample sizes, have been able to test variants with MAF of 0.01 or lower. There is an argument that rarer variants might have larger effect sizes (*Box 13.2*). The search for rarer variants associated with disease has been the driver behind some huge meta-analyses, involving 100 000 or more cases and controls. There is now a feeling that such studies have run their course, and as the cost of sequencing continues to fall, the future of common-disease genetics lies with population sequencing studies that can detect every variant in the sample.

The hope driving all this work has been that identifying genetic susceptibility factors would lead to better prediction of risk for individuals – motivating those at high risk to change their lifestyle or take other preventive measures – and also provide the pharmaceutical industry with leads for developing more effective drugs. Early hopes of quick progress towards these objectives proved over-optimistic. These remain the objectives, but are now seen as long-term goals. We saw that the variants pinpointed by

Effect sizes, odds ratios and allele frequencies

Having identified a variant that is associated with risk of a disease, the next question is how strong is the associated risk? Ideally one would like to know the **relative risk**: the risk of disease for a person with the variant compared to the risk for a person who does not have the variant. Unfortunately it is not possible to calculate relative risks from GWAS data: to do that it would be necessary to recruit a representative cohort of young unaffected people, then follow them to see who developed the disease and who did not, and correlate that with their genotypes. So instead, we calculate **odds ratios**: the odds of being a case if you have the variant compared to the odds if you do not (*Box figure 13.1*):

	Cases	Controls	Odds of being a case	Odds ratio
V present	a	b	a:b or a/b:1	$(a/b)/(c/d) = ad/bc$
V absent	c	d	c:d or c/d:1	

Box figure 13.1 – The odds ratio.

Variant V has been shown by GWAS to be associated with the disease under study. a, b, c and d are actual numbers of individuals, either cases or controls, who have or do not have V.

Odds ratios for variants identified by GWAS are virtually always modest – for the most part below 1.1, and often below 1.05. Early researchers were disappointed that the effects they discovered with so much hard work were so small, but actually that was entirely predictable. GWAS look at common variants. For a variant to be common it must have persisted in the population for many generations (or have been subject to very strong positive selection). A variant that significantly predisposed to disease should face negative selection and not be able to persist in a population for sufficient generations to have any chance of becoming common. Thus it was always predictable that variants detected by GWAS would have small effect sizes. Any variants with stronger effects should have much lower allele frequencies (or have some other way of escaping negative selection, for example, by affecting people only long after reproductive age). The question is then whether the combined effects of many weak variants could have a sufficiently strong effect on disease risk to be clinically useful. We will consider that question in more detail below.

GWAS identify ancestral chromosome segments that carry a susceptibility variant – but there is no reason to suppose that the variant identified by GWAS is the one that actually causes the susceptibility. Sometimes it will be, but more often than not the variant is most likely non-pathogenic but in linkage disequilibrium with the true pathogenic variant. A typical haplotype block might carry 20 or 25 common SNPs in addition to a spectrum of rarer variants found on particular examples of the block. Any one of them might be the true causal variant. Moving from the linked variant identified by GWAS to the actual causal variant has been, and remains, a major challenge. GWAS variants are for the most part located outside coding sequences; probably most causal variants affect enhancers or other regulatory sequences. There is no easy large-scale method to recognize the true causal variants (or their regulatory targets: the target of an enhancer is not necessarily the gene nearest to it). Refined statistical analysis can help, but usually only laborious wet-laboratory work, specific to each variant, can complete the task. Thus GWAS data has not yet had much impact on public health or clinical management. The review by Shendure *et al.* (2019) includes a clear and balanced discussion of the current state and the achievements and limitations of GWAS.

13.3. Investigations of patients

For Alzheimer disease and Type 2 diabetes, the two diseases described in **Cases 25** and **26**, the scope for genetic advice is currently quite limited. For each disease it is important to identify the small minority of cases that involve a mendelian (single gene) form of the disease – but for the multifactorial majority, genetic services currently have little to offer. However, investigations of people like **Bill Yamamoto** and **Zafira Zuabi** are now the mainstream of clinical genetic research. In this section we will look at the progress of investigations into the genetics of these two diseases. *Section 13.4* will set these examples into a more general framework to address the question at the head of this chapter – should we be testing for susceptibility to common diseases?

CASE 25 YAMOMOTO FAMILY

- Family history of dementia
- Alzheimer disease
- Test for ApoE4?
- Genetic susceptibility to Alzheimer disease
- Possibilities for therapy

333

344

395

When Bill Yamamoto talked with the geneticist, she confirmed his physician's statement that only early-onset Alzheimer disease was strongly inherited. The condition in Bill's mother and aunt was the common late-onset form. Both forms are defined by the same post-mortem brain pathology, with abundant extracellular senile plaques and intracellular neurofibrillary tangles (*Figure 13.8*), but the late-onset form is not simply inherited. Bill pressed the geneticist about ApoE4. She confirmed that there is a statistical association with late-onset Alzheimer disease, and that this was valid in Japanese as well as in people of European origin. A number of studies had shown an E4 allele frequency of 0.25–0.3 in Japanese Alzheimer disease patients, compared to 0.10 in controls. She advised against testing. Several professional bodies have recommended against ApoE testing for predictive purposes. For example, joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors (Goldman *et al.*, 2011) state that for families in which autosomal dominant Alzheimer disease (the rare early-onset mendelian form) is unlikely, genetic testing for susceptibility loci (e.g. ApoE) is not clinically recommended due to limited clinical utility and poor predictive value.

Similarly, in 2010 the European Federation of Neurological Societies stated that ‘... there is no evidence to suggest ApoE testing is useful in a diagnostic setting’.

When Bill suggested this attitude was patronizing, she asked what he would do if the result was positive. “Ask you what I must do to avoid developing the disease” he replied. “But there’s nothing I could tell you. There is no proven way of preventing Alzheimer disease, though there are drugs that may slow the progression. The best you can do is keep your mind and body active – but I would say that to anybody, regardless of their circumstances. And what would you do if the result was negative?” “Celebrate!” “But that would be wrong – those Japanese figures imply that 50–60% of Japanese Alzheimer patients do not have an E4 allele. The association is only statistical. It is not predictive for an individual”. The consultation did not reassure Bill about his risk, but it did persuade him that spending money on testing would not provide the reassurance he wanted.

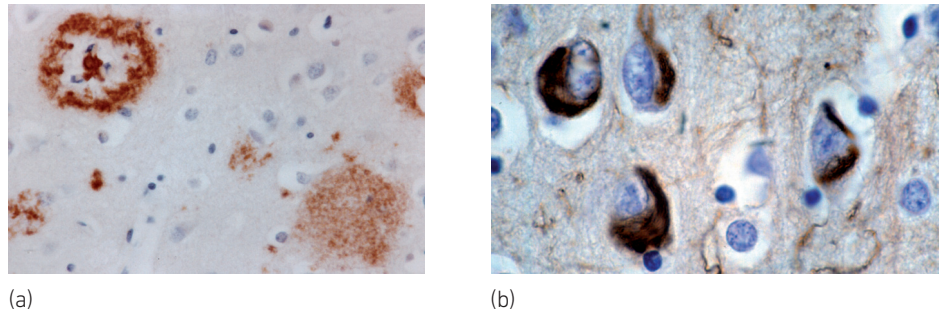


Figure 13.8 – The characteristic brain pathology of Alzheimer disease.

(a) Amyloid plaques and (b) neurofibrillary tangles. Photos courtesy of Dr Simon Lovestone, Oxford University.

A few percent of Alzheimer disease cases have onset before age 60. Most of them are just the tail of the age-of-onset curve for the typical late-onset disease, but around 10% are strongly familial and behave as mendelian dominant conditions. These are amenable to standard linkage analysis and positional cloning, as described in *Chapter 8*. These approaches have identified three causative genes (*Table 13.3*).

Table 13.3 – Known causes of early-onset Alzheimer disease

Gene	OMIM number	Location	No. of recorded families	Product
<i>APP</i>	104760	21q21	100	Amyloid precursor protein
<i>PSEN1</i>	104311	14q24	450	γ -secretase subunit
<i>PSEN2</i>	600759	1q31	30	γ -secretase subunit

The counts of families are from the UK Alzheimer Society (www.alzheimers.org.uk/about-dementia/risk-factors-and-prevention/alzheimers-disease-and-genes/). In total, these causes explain only around 0.5% of all Alzheimer disease; the great majority of cases are late-onset and non-familial.

The senile plaques in Alzheimer disease consist largely of β -amyloid protein. β -amyloid is derived from the amyloid precursor protein (APP) by proteolytic cleavage. APP is a 695 amino acid brain protein that is cleaved by the γ -secretase enzyme, producing $A\beta_{40}$ and

$A\beta_{42}$ peptides. $A\beta_{42}$ is thought to be the pathogenic variant. γ -secretase is a complex of five polypeptides including the *PSEN1* and *PSEN2* gene products. The rare mendelian forms of Alzheimer disease have clearly implicated amyloid β -peptides in the pathology, although their exact role remains unclear.

Geneticists have much to offer members of families affected by early-onset Alzheimer disease. They can try to identify a causative mutation in one of the known genes. If one is discovered, predictive testing can be offered, similarly to Huntington disease. A suitable protocol is described in the following chapter (Box 14.4). However, as Bill Yamamoto discovered, with the late-onset form we are still in the research phase and, at present, genetic services have nothing useful to offer.

Family studies suggest that 60–80% of the variance in susceptibility to late-onset Alzheimer disease is due to genetic differences between people, with a variety of environmental and lifestyle factors accounting for the remainder. One risk factor was identified as early as 1993. The *APOE* gene on chromosome 19q13 encodes apolipoprotein E. Many variants of *APOE* have been described (see OMIM 107741), but only three are common polymorphisms. *APOE**2, *3 and *4 are coding sequence variants producing ApoE proteins with either cysteine or arginine at positions 112 and 158 (Table 13.4). The allele frequencies have been studied in many populations. E4 is the ancestral allele, found in non-human primates, and it remains frequent in populations where foraging is still important. In settled agricultural populations the frequency of E4 is low.

Table 13.4 – The common apolipoprotein E alleles and their frequencies in various populations

	Residue 112	Residue 158	Spanish	UK	Chinese	Japanese	Native American	Khoi San
<i>APOE</i> *2	Cysteine	Cysteine	0.052	0.089	0.105	0.048	0.0	0.077
<i>APOE</i> *3	Cysteine	Arginine	0.856	0.767	0.824	0.851	0.816	0.553
<i>APOE</i> *4	Arginine	Arginine	0.091	0.144	0.071	0.101	0.184	0.370

Data from Corbo and Scacchi (1999; *Ann. Hum. Genet.* **63**: 301–310).

ApoE4 is a risk factor for both coronary artery disease and late-onset Alzheimer disease among people living in Westernized environments. People homozygous for E4 have 3–5 times, and heterozygotes about twice, the population risk of Alzheimer disease; E3 homozygotes have about the population risk, while E2 has a small protective effect. Several different hypotheses have been proposed for the mechanism. ApoE protein binds amyloid β -peptide, cholesterol, and many other molecules, with the different forms having different affinities. The E4 form enhances deposition of amyloid β -peptide. Neurons produce ApoE when stressed, and the E4 form is more subject to proteolytic cleavage, producing C-terminal fragments that may be toxic to mitochondria. Decreased mitochondrial function in neurons could lead to Alzheimer disease. Figure 13.9 shows the multiple ways in which ApoE protein may be involved in the pathogenesis of AD.

ApoE4 is a major susceptibility factor for Alzheimer disease but, as Bill Yamamoto learned, it accounts for only part of the overall susceptibility. Its estimated contribution is around 6%. Some 20 additional susceptibility factors have been identified through large GWAS (including a huge meta-analysis covering 74 046 cases), but collectively they explain only

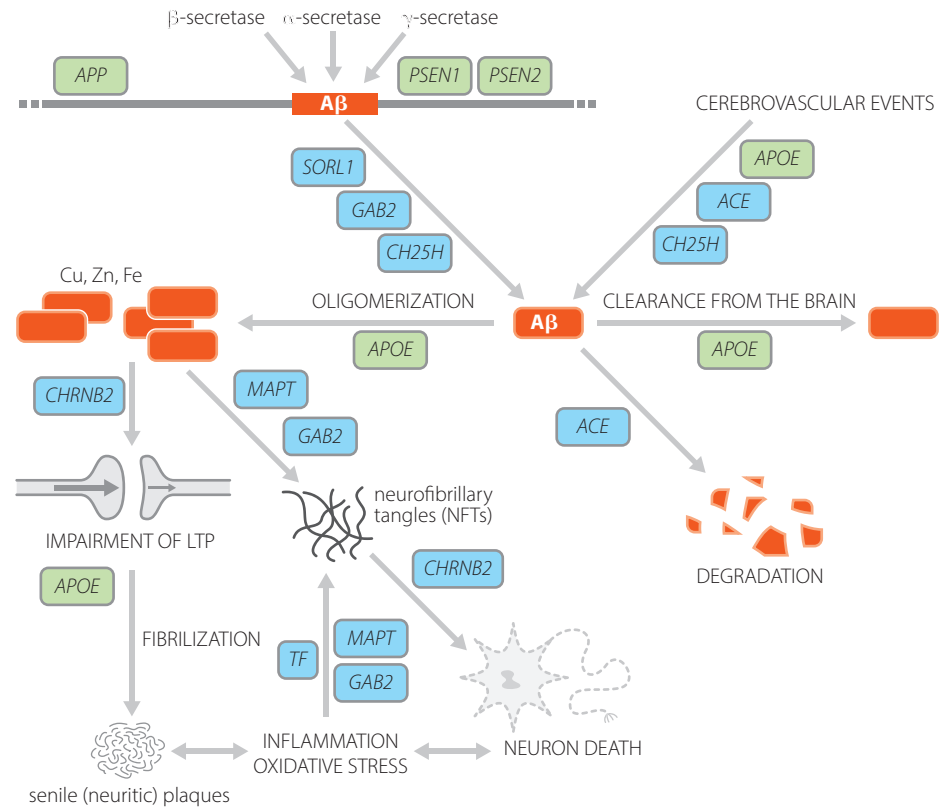


Figure 13.9 – Processes that may lead to Alzheimer disease.

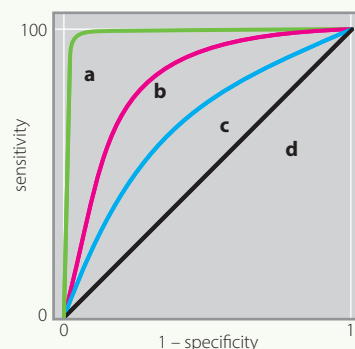
Note the multiple stages where ApoE protein may be involved. The possible roles of other AD candidate genes (confirmed, green, or suggested, blue) are indicated. LTP, long term potentiation. Reproduced from Bertram and Tanzi (2008) with permission from Nature Publishing Group.

a further 2% of the susceptibility (Karch *et al.*, 2014). As we will see in *Section 13.4*, this is a typical result for a common complex disease.

A measure of the ability of a test to identify cases is the AUC (area under the curve) statistic. The curve in question is the quaintly named ROC (receiver operating characteristic) curve (*Box 13.3*). This plots the sensitivity of a test against the specificity (see *Box 12.1* for definitions). As discussed in *Section 12.2*, there is usually a trade-off between sensitivity and specificity: if the action threshold is set very low so that no positive predictions are missed (high sensitivity), there will probably be many false positives (low specificity), and vice versa: setting the specificity high leads to low sensitivity. The ROC curve tracks the trade-off. AUC values range from 0.5 (no predictive value) to 1.0 (perfect prediction). To check the usefulness of a genetic test we ask how much the initial AUC (based just on clinical, etc. data) is improved by including the results of genotyping. The AUCs from two studies cited by Seshadri *et al.* (2010) were 0.826 and 0.670, respectively, based on age and sex alone. The difference between the two studies is presumably due to the younger average age and greater range of ages of cases in the first study (69 ± 9 years) compared to the second (80 ± 6 years). Incorporating the APOE genotype increased these figures to 0.847 and 0.702, respectively. Adding in two of the newly identified susceptibility factors,

Measuring the performance of a test using the ROC curve

The ROC curve plots sensitivity against (1–specificity) for a predictive test. A perfect test, that always predicts correctly, would give curve **a**. The area under the curve is 1. A completely useless test, whose predictions are no better than random, would give curve **d** (area under the curve = 0.5). Tests with varying degrees of usefulness might give curves like **b** or **c**.



BOX 13.3

CLU and *PICALM*, added only 0.002 and 0.003 to the respective AUCs. In other words, knowing the *APOE* genotype adds little, and knowing the *CLU* and *PICALM* types virtually nothing, to predictions based on age and sex alone.

All these findings lead one to ask, what should we tell the patients? The answer is, not much. Several expert panels have advised against using ApoE or any other alleged genetic risk factor for clinical purposes – the American College of Medical Genetics and the National Society of Genetic Counselors as mentioned previously (Goldman *et al.*, 2011), but also the UK NHS in 2009 and the European Federation of Neurological Societies in 2010. ApoE4 is neither necessary nor sufficient for Alzheimer disease. ApoE genotyping would not assist the diagnosis in a person with possible Alzheimer disease, and it would not usefully predict the likelihood of somebody developing Alzheimer disease. A British study concluded that the predictive power was too low to affect insurance underwriting, even for long-term care insurance (Warren, 1999). ApoE is quite properly tested as part of the investigation of dyslipidemias, and this raises the tricky question of whether patients should be told results that are irrelevant to their lipid problem and may be disturbing. Inevitably ApoE testing is also available over the internet. However, some research data suggest this is not a cause for moral panic. In one trial (Green *et al.*, 2009) some people in Bill Yamamoto's position welcomed ApoE testing, and were not particularly disturbed when the result showed they were E4 positive.

Thus the results of genetic research have not so far done much to identify people at high risk of the common late-onset form of Alzheimer disease (although this may change with the development of polygenic risk scores, see Section 13.4). However, part of the reason for pursuing these studies is to understand the pathology as a necessary prerequisite for developing effective treatments, and here the story is more optimistic. A true cure would of course be wonderful, but even a treatment that postponed the onset by a few years would have an enormous effect on the healthcare burden of late-onset Alzheimer disease. Although the newly identified susceptibility variants each contribute very little to the overall risk, their functions give clues to the pathology. In that respect it is interesting that they cluster in a few areas, namely immunity, inflammatory responses, lipid metabolism and endocytosis (when the cell membrane engulfs and internalizes a substance from outside). These must be giving us clues to the pathology underlying this complex disease.

CASE 26 ZUABI FAMILY

- Zafira, woman aged 52 years
- Overweight, sedentary lifestyle, insatiable thirst
- Type 2 diabetes
- Son's lifestyle and heredity put him at high risk
- Management of family
- Genetic susceptibility to Type 2 diabetes
- Possibilities for therapy

334

349

395

Diabetes mellitus, defined by hyperglycemia (blood glucose >7 mmol/l fasting, >11 mmol/l non-fasting), or by a glucose tolerance test, is a heterogeneous condition. In addition to various minor types, the two major types are:

- Type 1 (T1D) – a sudden onset disease in young people, the result of an autoimmune attack on the pancreatic β -cells, and not associated with obesity
- Type 2 (T2D) – normally with adult onset, associated with obesity and physical inactivity, without autoimmune features, and resulting from a combination of inadequate secretion of insulin and resistance to its effects; this is the type in this case.

These two separate diseases both involve genetic susceptibility and environmental factors. For T2D, evidence for environmental factors comes from the alarming recent increases in prevalence, and from intervention studies showing the efficacy of weight control and exercise in reducing progression from a pre-diabetic to the full diabetic state. Evidence for genetic factors comes from family and twin studies, and from the ethnic variations in prevalence. A positive family history confers an increased risk: 2–3 fold with one affected first-degree relative, substantially higher with more than one affected relative. Many studies have reported higher concordance in monozygotic compared to dizygotic twins. The prevalence varies greatly between different ethnic groups, even when members of the groups live intermingled in multiethnic communities.

The world faces an epidemic of T2D. In the USA, the prevalence doubled between 1990 and 2005. In 2017, 30.3 million adults in the USA (9.4% of the total US population, including 25.2% of those aged 65 years or older) had diabetes, 90–95% of which was T2D. 34% of all adults, and 48% of those aged 65 and over, had pre-diabetes (metabolic syndrome). The annual cost to the economy in 2012 was estimated as \$245 billion. (Data from 2017 National Diabetes Statistics Report, www.cdc.gov/diabetes/data/statistics/statistics-report.html.) These figures, and similar trends from many other countries, have stimulated intensive efforts to understand the causes of T2D. *Figure 13.10a* shows the self-reinforcing pathogenic cascade that produces the hyperglycemia and increased free fatty acids of T2D, while *Figure 13.10b* shows the complicated events controlling insulin signaling.

As regards genetic causes, it has long been known that 1–2% of T2D cases have a different condition, maturity onset diabetes of youth (MODY) that is mendelian. MODY affects all ages and is not associated with obesity or inactivity. It can be caused by mutations in any of seven or more genes, and identifying the cause is important because different forms respond well to different drugs. For the common complex T2D, large-scale GWAS have identified many susceptibility factors. Flannick and Florez (2016) review progress. A meta-analysis covering 34 840 cases and 114 981 controls brought the list of established susceptibility loci up to 64 (Morris *et al.*, 2012), and later work has moved the total to around 100. However, as is usual with complex diseases, the individual odds ratios are low, mostly in the range 1.05–1.2, and taken together, all identified factors account for only about 10% of the overall family aggregation. Fuchsberger *et al.* (2016) complemented the GWAS data with exome or whole genome sequencing of 6504 cases and a similar number of controls to explore the possibility that much of the 'missing heritability' is due to rare variants; their data, however, did not support a major role for rare variants. Few of the factors identified by GWAS map within coding sequences of genes, but taking the

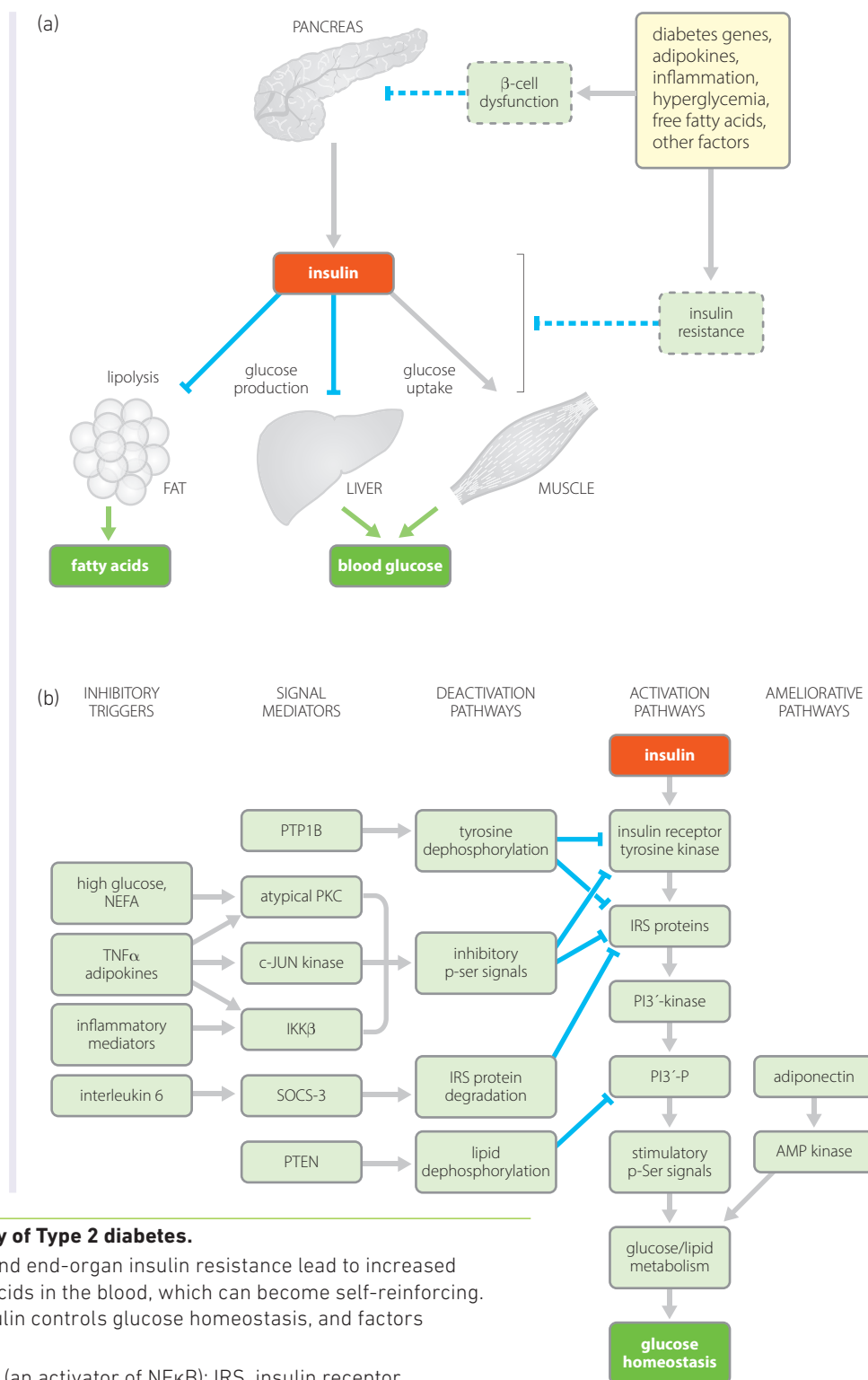


Figure 13.10 – Pathophysiology of Type 2 diabetes.

(a) Decreased insulin secretion and end-organ insulin resistance lead to increased levels of glucose and free fatty acids in the blood, which can become self-reinforcing.

(b) The mechanism by which insulin controls glucose homeostasis, and factors influencing it.

IKK β , NF κ B inhibitory unit kinase (an activator of NF κ B); IRS, insulin receptor substrate; NEFA, non-esterified fatty acids; PI, phosphoinositol; PKC, protein kinase C; p-Ser, phosphoserine; PTP1B, phosphotyrosine phosphatase 1B; SOCS-3, suppressor of cytokine signaling-3; TNF α , tumor necrosis factor α .

Both figures reproduced from Stumvoll *et al.* (2005) from *The Lancet* with permission from Elsevier.

closest gene as the most likely functional candidate, the distribution of functions makes it plausible why many of them should influence susceptibility to T2D (Figure 13.11).

Many studies have examined the potential of all this new genetic knowledge to predict a person's risk of developing T2D. Individual susceptibility factors identified by GWAS have effect sizes far too small for the genotypes to have any predictive value for individuals. However, as for many diseases, a large number of separate susceptibility factors have been defined – for example, Flannick and Florez (2016) list over 100 variants detected in GWAS of T2D. The question then arises, could combinations of such factors provide clinically useful risk predictions? Several studies of T2D have attempted to answer this question. The design is the same in each case.

- A cohort of healthy individuals is recruited, and a baseline clinical examination is used to predict their risk of developing T2D. Depending on the study the baseline examination might include age, sex, BMI, blood pressure, blood glucose, measures of insulin secretion, family history and so on.
- Individuals are followed up for many years, and the accuracy of the initial prediction checked by seeing who did and who did not develop T2D.
- Individuals are genotyped for a range of susceptibility variants, and the question is asked, how much better would the initial prediction have been if those genotypes had been part of the baseline assessment?

Table 13.5 shows the results of five such studies (summarized by Hivert *et al.*, 2014). The measure of the performance of the predictive tests is the AUC (area under the curve) statistic, described above (Box 13.3). In each study the question is how much the AUC of the initial assessment is improved by including the results of genotyping (do not compare the AUC from the different studies; those differences reflect the different recruitment criteria and stringency of the baseline examinations).

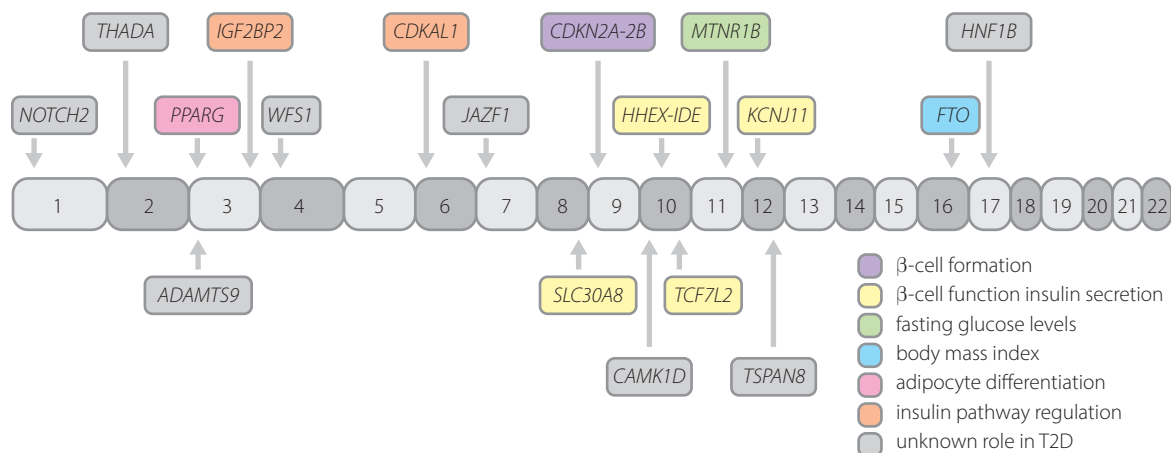


Figure 13.11 – Eighteen genetic susceptibility factors for Type 2 diabetes.

The figure shows the chromosomal locations of 18 genes in which variants have been associated with susceptibility or resistance to T2D. Their likely roles in the pathogenesis are shown by the color coding. Reproduced from Frazer *et al.* (2009; *Nat. Rev. Genet.* **10**: 241–251) with permission from the Nature Publishing Group.

Table 13.5 – Does genotyping improve the ability to predict whether somebody will develop Type 2 diabetes?

Clinical indicators	AUC from clinical indicators	Genetic predictor	AUC using combined clinical and genetic data	Reference
Age, sex, BMI	0.78	18 loci	0.80	Lango <i>et al.</i> (2008)
Age, sex, BMI	0.66	18 loci	0.68	van Hoek <i>et al.</i> (2008)
Age, sex, BMI, family history, liver enzyme levels, smoking, measures of insulin secretion and action	0.74	16 loci	0.75	Lyssenko <i>et al.</i> (2008)
Age, sex, family history, BMI, blood pressure, blood glucose, HDL cholesterol, triglycerides	0.90	17 loci	0.901	Meigs <i>et al.</i> (2008)
Age, sex, family history, BMI, blood pressure, blood glucose, HDL cholesterol, triglycerides	0.903	40 loci	0.906	de Miguel-Janes <i>et al.</i> (2011)

The AUC statistic measures the predictive power of a test; the higher the value, the better the prediction (see text for discussion and an explanation). The 16–18 loci used in the first four studies largely overlapped the 18 shown in *Figure 13.11*. BMI, body mass index.

Taken in isolation, the genotypes are predictive of risk; when combined with clinical examination and family history they do slightly improve the prediction, but only very slightly. The study of Meigs *et al.* (2008) concluded that their 17 genotypes would result in, at most, 4% of their subjects being reclassified to a different risk category. An alternative and more promising way of using GWAS data is described in *Section 13.4*.

As regards genetic advice, the Zuabi family is fairly typical. An overlapping set of problems – T2D, cardiovascular disease, coronary heart disease – cluster loosely in families. The metabolic syndrome is a clear predictor of risk, as are its individual components. While a strong family history predicts increased risk, there is little scope for specifically genetic advice: the general advice for every overweight and inactive person, regardless of family history, is to get some exercise and lose some weight. Except in MODY, the many possible drugs are prescribed according to the physiology and not the genetics.

13.4. Going deeper...

Why have GWAS told us so little that is clinically useful?

The specific examples of Alzheimer disease and T2D are typical of the great majority of complex diseases. GWAS have been technically very successful. They have identified thousands of confirmed susceptibility factors, but their low clinical utility has been a major disappointment. The poor predictive power shown in the five studies of T2D (*Table 13.5*) is fairly typical: similar disappointing results have been reported for other conditions. Is the problem our incomplete knowledge of the genetics of each disease, as reflected in the ‘missing heritability’?

The ‘missing heritability’ problem

In *Section 13.2* we saw how family, twin and adoption studies allowed the heritability of a condition to be estimated. Knowing the effect size of a GWAS variant, one can calculate its contribution to the heritability. For almost every condition studied, adding together the heritability calculated for each known GWAS variant, this ‘bottom-up’ heritability is far less than the heritability estimated from ‘top down’ family studies. Even for well-studied diseases, all known GWAS variants taken together account for only 20–50% of the heritability estimated from top-down studies. This has been called the ‘missing heritability’ problem. Are we missing something important?

Several hypotheses have sought to explain this problem.

- Much heritability may be due to variants with large effect sizes that are too rare to be detected by GWAS. These would be a class of variants intermediate between the common variants of small effect identified by GWAS and the ultra-rare variants with very strong effects that underlie monogenic diseases (*Figure 13.12*). They would be revealed by large-scale population sequencing.
- As explained above, GWAS estimates are based on variants that are not usually the actual causative variant. If the latter could be used, their contribution to the heritability would probably be greater.
- Top-down heritability estimates may be exaggerated, so that there is not actually so much heritability needing to be explained. One cause, alluded to previously, is that family studies often fail to take full account of the effects of shared family environment. A second cause may be interactions between genetic factors. Theoretical simulations show that genetic interactions could cause standard analyses to introduce ‘phantom heritability’.
- GWAS cannot identify variants with very weak effects, but these may be so numerous that their collective effect is large. Thresholds of significance in GWAS need to be very stringent in order to avoid large numbers of false positives – but necessarily these stringent thresholds will exclude some true but weak positives.
- Something quite novel and unknown – ‘genetic dark matter’ – might be operating.

Any or all of these explanations might apply to particular cases, though the case for mysterious ‘dark matter’ gets weaker by the year. However, as general explanations of missing heritability, the first three hypotheses have not fared well. Although sequencing has identified plenty of individual rare variants, studies have revealed no diseases where rare variants seem to be major causes of the missing heritability – see, for example, the paper by Fuchsberger *et al.* (2016) on T2D. Using causative rather than associated variants would no doubt improve estimates, but the present associations must already include most of the effect at each locus. Top-down heritability estimates may well often be rather too high, but not to a degree sufficient to account for much of the missing heritability. In fact, the cumulative effect of large numbers of individually very weak variants has emerged as the most promising general explanation. A statistical approach pioneered by Peter Visscher and colleagues (Yang *et al.*, 2011, 2015) allows the overall effect of these weak factors to be calculated. It does not allow identification of individual factors, but in several carefully studied cases it appears able to account for much of the missing heritability.

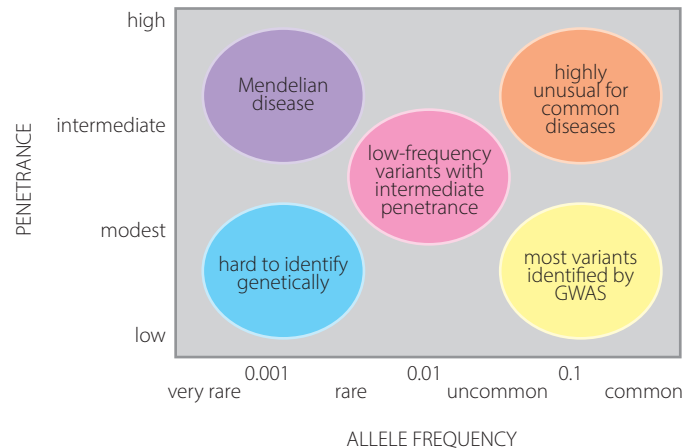


Figure 13.12 – One theory of the genetic architecture of disease.

Current experience suggests the intermediate category of variants are not numerous. Adapted from McCarthy *et al.* (2008; *Nat. Rev. Genetics*, **9**: 367) with permission from Nature Publishing Group.

In the past, before GWAS, the genetics of common diseases was explained by a **polygenic** model. Susceptibility was assumed to depend on the cumulative effects of a very large number of factors, each individually with a very small effect (the mathematical model assumed an infinite number of factors, each with an infinitesimally small effect). As GWAS identified individual factors, the polygenic model seemed less relevant. Ironically, it is now coming back. Not only does Visscher's polygenic model dispose of much of the mystery over missing heritability, we will see below that it also seems to hold the key to making useful predictions of individual risks. Just as the missing heritability problem was largely (although not necessarily completely) solved by using whole genome data rather than just panels of known susceptibility factors, a similar approach to risk estimation through **polygenic risk scores** (PRS) is proving promising.

Polygenic risk scores

Rather than basing a risk estimate on a limited number of known susceptibility factors, PRS use genome-wide genotypes, without asking whether or how individual genotypes contribute to risk. A typical procedure would be to:

- use analysis of the combined data (cases and controls together) of the largest available GWAS to produce a range of candidate algorithms to discriminate cases from controls
- test the algorithms on a large sample of individuals from a population biobank, whose genotypes and clinical data are known, to find the one that performs best
- use this algorithm to produce PRS on an independent sample from the same population to confirm its performance.

A much-quoted study by Khera and colleagues (2018) illustrates the process. Predictors for five common diseases (coronary artery disease, atrial fibrillation, T2D, inflammatory bowel disease and breast cancer) were developed using data from recent large GWAS. Data on 120 280 subjects in the UK Biobank were then used to identify the predictors that gave

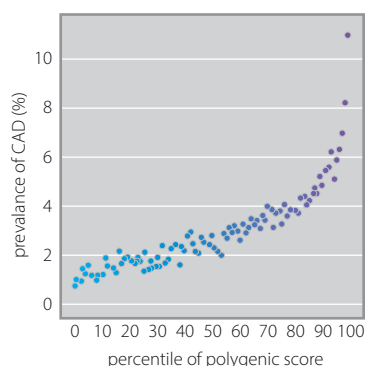


Figure 13.13 – Distribution of polygenic risk scores for coronary artery disease.

Figure reproduced from Khera *et al.* (2018) with permission from Nature Publishing Group.

the highest AUC. These were then used to compute PRS for a further 288 978 Biobank subjects. For each condition, individuals with PRS in the top end of the distribution had several times the likelihood of being affected compared to individuals at the lower end. *Table 13.6* and *Figure 13.13* show some of the data.

Table 13.6 – Proportion of the population at three-, four- and five-fold increased risk for each of the five common diseases

Odds ratio	Condition	Individuals*	%
>3	CAD	23 119	8.0
	AtrFib	17 627	6.1
	T2D	10 099	3.5
	IBD	9209	3.2
	BrCa	2369	1.5
>4	CAD	6631	2.3
	AtrFib	4335	1.5
	T2D	578	0.2
	IBD	2297	0.8
	BrCa	474	0.3
>5	CAD	1443	0.5
	AtrFib	2020	0.7
	T2D	144	0.05
	IBD	571	0.2
	BrCa	158	0.1

*Figures for CAD, AtrFib, T2D and IBD are based on 288 978 subjects; figures for BrCa are based on 157 985 subjects. AtrFib, atrial fibrillation; BrCa, breast cancer; CAD, coronary artery disease; IBD, inflammatory bowel disease; T2D, Type 2 diabetes. Data from Khera *et al.* (2018).

So should we be testing for susceptibility to common diseases?

These and similar results have sparked considerable optimism that GWAS data will finally find its clinical application. Polygenic risk scores are indeed very promising, but before we rush to incorporate them into routine clinical practice a few caveats are in order.

- Khera and colleagues created a total of 31 candidate estimators. The quoted results are for individuals in the UK Biobank, using the estimator that performed best on a test set of individuals from that same biobank. Risk scores that are not so carefully selected and targeted may not perform so well. On balance it seems likely that PRS can indeed identify some high-risk individuals, but probably usually with less precision than the results in *Table 13.6*. If bigger and better GWAS become available, the predictive power of PRS based on them should increase.
- Polygenic risk scores are specific to the population that was used in the original GWAS. They cannot be applied in a different population. This raises both a practical and an ethical point. The practical point is obvious; the ethical point is that only populations where large GWAS had been performed could benefit, so their application would not be equitable (Martin *et al.*, 2019).

- An analysis of Khera's data for coronary artery disease by Wald and Old (2019) showed that if the PRS was used as a population screening tool (even in the population in which it had been trialed and perfected) it could detect only 15% of cases, at a cost of 5% false positives (or 10% of cases with 3% false positives, if a more stringent threshold PRS were used).

The analysis by Wald and Old (2019) suggests that PRS would perform poorly if used as a population screening tool to try to identify all those individuals who will develop the disease. However, that would not be the best way to use them. They might be better used to give individuals an indication of their personal risk. Used in this way, and combined with clinical, family history and lifestyle information, they seem to have considerable promise for motivating some high-risk individuals to take preventive measures. As long as the preventive measures are not too onerous, expensive or risky, and fit in with general public health recommendations, it may not matter too much that many people given a high risk would never develop the condition. For example, PRS for breast cancer could be used to help decide at what age a woman should enter a routine mammography screening program (note, however, that in *Table 13.6* the proportion of women with high-risk scores is much lower for breast cancer than for the much-quoted case of coronary artery disease – evidently the predictive power of PRS will vary with different conditions). One worry is that this benefit would be counterbalanced by people given a low risk feeling false assurance and taking up risky behaviors. Limited evidence suggests this is not a major problem.

What if we knew everything?

A high PRS for coronary artery disease (CAD) of Khera and colleagues (2018) will identify only a small minority of the people who will actually go on to develop the condition (Wald and Old, 2019). How far is this because at present we do not fully understand the genetics of CAD? In 2012, Roberts and colleagues proposed an ingenious way of answering this question by using twins. If one of a pair of monozygotic twins has CAD, the co-twin has the identical genetic factors, and therefore the likelihood that the co-twin will develop the disease is a measure of how far perfect genetic knowledge would predict the outcome.

Roberts *et al.* (2012) considered 24 conditions that, between them, accounted for much of the morbidity and mortality in the US population. For each condition they used twin data to calculate the likelihood that interpreting a full genome sequence in the light of perfect knowledge of genetics – knowing every susceptibility factor and every genetic interaction – would predict the chance of an unaffected person receiving a clinically significant positive or negative test result, and also the chance that a perfect genetic test would have correctly identified a person who was actually affected. Some of their results are shown in *Figure 13.14*. The range shown for each condition reflects our *current* lack of knowledge; in the imagined future where we know everything there would be a single point for each condition somewhere within that range.

The perhaps surprising conclusion is that for every condition only a minority of people would get a clinically useful result, even if we knew everything there was to know about the genetics, and only a minority of affected people would have been correctly identified. The low percentage of people getting a significant positive prediction in part

(b) of the figure must be seen against the fact that most of these conditions are rare – hence it is only to be expected that few people would get a positive prediction, even if all predictions were perfectly accurate. The poor performance in part (a) of the figure is maybe less expected. At first sight this result seems counterintuitive. Surely perfect knowledge should lead to perfect predictions? On closer thought it is unsurprising. Stepping back from the clever mathematics of Roberts *et al.* (2012) which some may find contentious, a simple picture emerges.

The simple observation is that the average person has average susceptibility, and a perfect genetic test will only confirm that. Susceptibility depends on the cumulative

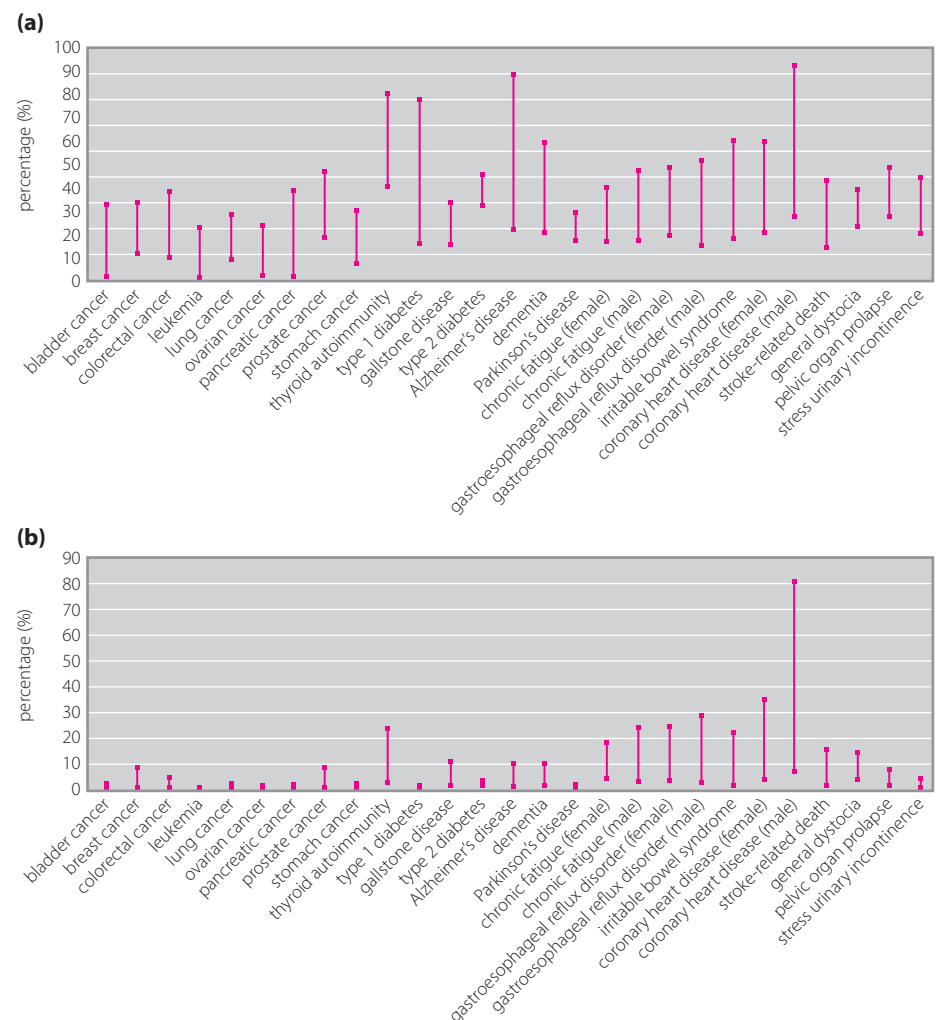


Figure 13.14 – The performance of predictions based on a person's whole genome sequence and perfect knowledge of the genetics of each condition.

(a) The percentage of affected people who would have been correctly identified. (b) The proportion of unselected people who would receive a clinically useful prediction of disease risk (defined here as a risk of 10% or double the population risk, whichever is the greater). Reproduced from Roberts *et al.* (2012) with permission from American Association for the Advancement of Science.

effect of many independent risk loci. At each risk locus one allele very slightly increases susceptibility while the alternative allele very slightly decreases it. The loci are independent and most people will have a mix of high- and low-susceptibility alleles. Only a small minority will have a strong preponderance of high-risk (or low-risk) alleles, leading to a clinically significant result. For most people perfect knowledge will just confirm that they are at more or less the population risk. However, this is not zero risk, and because there are far more people at roughly population risk compared to those with significantly increased risk, the majority of cases will come from the body of the risk distribution, and would not be predicted by the risk score. Hence the results of Wald and Old (2019).

Coming back to the question at the head of this chapter, 'should we be testing for susceptibility to common diseases?', the answer is nuanced. We should not be attempting to use genetic tests to identify everybody who will develop the condition. It is interesting to compare PRS for common diseases with non-invasive prenatal diagnosis of Down syndrome (*Disease box 12*). In both cases the result rests on the cumulative effect of a large number of individually weak effects. In the bloodstream of a woman whose fetus has Down syndrome, each individual chromosome 21 sequence is present in slightly greater numbers compared to sequences from other chromosomes (only slightly greater because only a small proportion of the total DNA is fetal). The effect of each individual sequence is blurred by random variation in the sequencing process. However, when the fetus has Down syndrome the effect goes in the same direction for every sequence, and so by considering a large number of different sequences from chromosome 21 we overcome the random fluctuations and achieve a firm prediction. In common disease susceptibility the many individual loci are independent, so for most people they more or less average out and there is no clinically useful prediction. Thus we should not be using DNA tests to try to identify everybody who will develop the condition, because even with perfect knowledge it will never work. It may, however, be valid and useful to use PRS together with clinical, family history and lifestyle data, to report individual personal risks, at least for conditions and populations where good GWAS data exist.

Educating people to understand these personal risks will be a major task. A UK Health Minister was invited to take a PRS-based test for his risk of prostate cancer. He said he was 'shocked' to discover his risk was 15%. He famously claimed the test might have saved his life, and called for an urgent roll-out of such testing across the NHS. In reality his baseline population risk before testing was 18%, and in any case, most of those who do develop prostate cancer do not die of it. And this was the minister for health!

Autism spectrum disorders

Autism is a strange and fascinating condition, the subject of a vast literature and a huge amount of research. Genetic aspects are reviewed by Woodbury-Smith and Scherer (2018) and by Iakoucheva *et al.* (2019). Diagnosis is entirely based on observed patterns of behavior. By definition, children with autism must manifest delays in 'social interaction, language as used in social communication, or symbolic or imaginative play' with 'onset prior to age 3 years', according to the *Diagnostic and Statistical Manual of Mental Disorders*. Children with classic autism generally have impairment of:

- social interaction, including poor eye contact, failure to develop age-appropriate peer relationships and to seek to share activities
- communication development, including delay or total lack of spoken language or use of repetitive language and lack of imaginative or imitative play

- behavior manifested by inflexibility of routines, obsessions and repetitive and stereotypical patterns of activity.

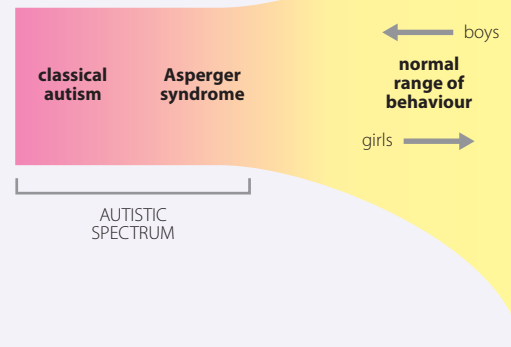
About 1% of children are labeled as autistic. The abnormal behaviors range from very severe, with no language, through to behaviors that seem just exaggerated versions of normal, and are often labeled Asperger syndrome. There is a 4:1 male preponderance – interesting because we all accept that ‘normal’ boys are more likely than girls to spend hours obsessively playing solitary computer games and to be socially inept (Box figure 13.2). 50% of autistic children have some co-morbidity, usually intellectual disability, and 10% have a specific syndrome such as Rett (OMIM 312750) or Fragile-X (OMIM 300624), in which autistic behavior is a known component.

No environmental factors have been identified as major causes of autism. Early reports of abnormal parenting, with ‘refrigerator mothers’, were not substantiated by more careful studies, and despite the publicity it has received, no serious study has implicated the MMR (measles – mumps – rubella) or any other vaccine. By contrast, many observations point to important genetic determinants. The sib risk is 10–20% and numerous studies have shown high heritability, typically around 80%.

Genetic investigations have revealed three types of susceptibility factors.

- Potentially pathogenic copy-number variants (CNVs) are present in 7–10% of patients with autism. Deletions at 15q13.3, 16p11.2 and 22q13.3 among many others, and duplications at 1q21.1, 15q11q13, 16p11.2, and 22q11.2 have been repeatedly observed. Often the variant is *de novo*, but in some cases it is also present in an apparently normal parent. In one large study, 4.7% of autism spectrum disorder (ASD) patients, but only 1–2% of controls, had a *de novo* CNV.
- Both inherited and *de novo* loss of function single gene variants are seen in cases at a much higher frequency than in controls (including unaffected sibs of cases). It is estimated that *de novo* mutations of genes contribute in approximately 30% of cases. Over 100 high-confidence susceptibility genes have been identified. As would be expected for variants with large effects, subjects that carry *de novo* mutations have lower nonverbal IQs than subjects who do not (Iossifov *et al.*, 2014). Some of the *de novo* CNVs involve the same genes as the *de novo* mutations (Pinto *et al.*, 2014). Variants with reduced penetrance may be inherited from a clinically unaffected parent.
- GWAS have provided evidence for common, low-effect susceptibility variants and led to the elaboration of PRS.

Many of the CNVs and susceptibility genes also predispose to other neurodevelopmental conditions, particularly intellectual disability and schizophrenia (Box figure 13.3). They seem to cause a general neurodevelopmental vulnerability, that may then manifest in various ways depending, perhaps, on genetic background, life events or chance. Moreover, the clinical severity of the known pathogenic CNVs or mutations is influenced by the presence of additional CNVs and rare variants, and also by the PRS. It seems that it is some sort of overall genomic burden that determines the phenotype. But this is not like the situation described previously for diabetes or other common conditions, where many variants of individually small effect cause disease by their cumulative effect in a person.

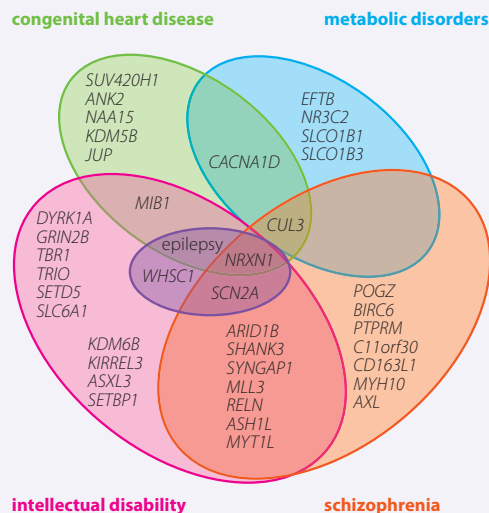


Box figure 13.2 – Autistic spectrum disorders merge into normal behavior with no natural boundaries.

Individual autistic children typically have only one of the major susceptibility factors, but the factor is different in different children.

The fact that there are many different ways of becoming autistic is intriguing. Non-autistic behavior seems to depend on correct functioning of a large number of genes. Looking at the functions of the gene products, they cluster around synaptic function, chromatin remodeling and regulation of transcription. There is evidence, discussed by Iakoucheva *et al.* (2019), that many of the different genes are part of regulatory networks involved in fetal brain development. Why dysregulation should specifically result in autistic behavior is far from obvious. Intellectual disability is in principle much easier to understand. A specific failure in very early language development might perhaps be a unifying factor. Achieving a full understanding of autistic behavior depends on closing the gap between biochemical study of synapses and psychological studies of behavior. That is a task likely to occupy scientists for many years to come.

Websites of the UK and US support groups are www.nas.org.uk/ and www.autism.org.uk.



Box figure 13.3 – Overlap between genes involved in ASD and other conditions.

Transmitted or *de novo* loss of function mutations in all the genes shown have been identified as likely causal factors in autism. Mutations in the same genes are seen in other conditions as shown. Reproduced from De Rubeis *et al.* (2014) with permission from Nature Publishing Group.

13.5. References

- Bertram L and Tanzi RE** (2008) Thirty years of Alzheimer disease genetics: the implications of systematic meta-analyses. *Nat. Rev. Neurosci.* **9**: 768–778.
- de Miguel-Yanes JM, Shrader P, Pencina MJ, et al.** (2011) Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. *Diabetes Care*, **34**: 121–125.
- De Rubeis S, He X, Goldberg AP, et al.** (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**: 209–215.
- Eckel RH, Grundy SM and Zimmet PZ** (2005) The metabolic syndrome. *Lancet*, **365**: 1415–1428.
- Flannick J and Florez JC** (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**: 535–549.
- Fuchsberger C, Flannick J, Teslovich TM, et al.** (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**: 41–47.
- Goldman JS, Hahn SE, Williamson J, et al.** (2011) Genetic counseling and testing for Alzheimer disease: joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors. *Genetics Med.* **13**: 597–605.

- Green RC, Roberts JS, Cupples LA, et al.** (2009) Disclosure of APOE genotype for risk of Alzheimer's disease. *New Engl. J. Med.* **361**: 11–20.
- Hivert M-F, Vassy JL and Meigs JB** (2014) Susceptibility to type 2 diabetes mellitus – from genes to prevention. *Nat. Rev. Endocrinol.* **10**: 198–205.
- Iakoucheva LM, Muotri AR and Sebat J** (2019). Getting to the cores of autism. *Cell*, **178**: 1287–1298.
- International HapMap Consortium** (2003) The International HapMap Project. *Nature*, **426**: 789–796.
- Iossifov I, O’Roak BJ, Sanders SJ, et al.** (2014) The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature*, **515**: 216–221.
- Karch CM, Cruchaga C and Goate AM** (2014) Alzheimer’s disease genetics: from the bench to the clinic. *Neuron*, **83**: 11–26.
- Khera AJ, Chaffin M, Aragam KG, et al.** (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**: 1219–1224.
- Lango H, UK Type 2 Diabetes Consortium, Palmer CN, et al.** (2008) Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes*, **57**: 3129–3135.
- Lyssenko V, Jonsson A, Almgren P, et al.** (2008) Clinical risk factors, DNA variants and the development of Type 2 diabetes. *New Engl. J. Med.* **359**: 2220–2232.
- Martin AR, Kanai M, Kamatani Y, et al.** (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**: 584–591.
- Meigs JB, Shrader P, Sullivan LM, et al.** (2008) Genotype score in addition to common risk factors for prediction of Type 2 diabetes. *New Engl. J. Med.* **359**: 2208–2219.
- Morris AP, Voight BF, Teslovich TM, et al.** (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**: 981–990.
- Pinto D, Delaby E, Merico D, et al.** (2014) Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**: 677–694.
- Ralph P and Coop G** (2013) The geography of recent genetic ancestry across Europe. *PLOS Biol.* **11**: e1001555.
- Roberts NJ, Vogelstein JT, Parmigiani G, et al.** (2012) The predictive capacity of personal genome sequencing. *Sci. Trans. Med.* **4**: 133ra58.
- Seshadri S, Fitzpatrick AL, Ikram MA, et al.** (2010) Genome-wide analysis of genetic loci associated with Alzheimer disease. *J. Am. Med. Assoc.* **303**: 1832–1840.
- Shendure J, Findlay GM and Snyder MW** (2019) Genomic medicine – progress, pitfalls, and promise. *Cell*, **177**: 45–57.
- Stumvoll M, Goldstein BJ and van Haeften TW** (2005) Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*, **365**: 1333–1346.

- van Hoek M, Dehgan A, Witteman JCM, et al.** (2008) Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes*, **57**: 3122–3128.
- Visscher PM, Hill WG and Wray NR** (2008) Heritability in the genomics era – concepts and misconceptions. *Nat. Rev. Genet.* **9**: 255–260.
- Wald NJ and Old R** (2019) The illusion of polygenic disease risk prediction. *Genetics Med.* **21**: 1705–1707.
- Warren V** (1999) *Report of Work Group on Genetic Tests and Future Need for Long-term Care in the UK*. Continuing Care Conference, London.
- Wellcome Trust Case–Control Consortium** (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**: 661–678.
- Welter D, MacArthur J, Morales J, et al.** (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associates. *Nucl. Acids Res.* **42**(D1): D1001–D1006.
- Woodbury-Smith M and Scherer SW** (2018) Progress in the genetics of autism spectrum disorder. *Dev. Med. Child Neurol.* **60**: 445–451.
- Yang J, Manolio TA, Pasquale LR, et al.** (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**: 519–525.
- Yang J, Bakshi A, Zhu Z, et al.** (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**: 1114–1120.

Useful websites

- 1000 Genomes Project: www.internationalgenome.org
- International HapMap project: <http://hapmap.ncbi.nlm.nih.gov/>
- National Human Genome Research Institute catalog of GWAS: www.genome.gov/gwastudies/
- UK 100 000 Genomes Project: www.genomicsengland.co.uk/the-100000-genomes-project/

13.6. Self-assessment questions

- (1) Which of the following would provide the strongest evidence for involvement of genetic factors in susceptibility to heart attacks?
 - (a) A raised incidence of heart attacks among sibs of index cases.
 - (b) Increased concordance of same-sex compared to opposite-sex twin pairs.
 - (c) For index cases who were adopted, an increased incidence among the biological, but not the adoptive, relatives.
 - (d) The observation that children tend to resemble their parents in healthy or unhealthy eating habits.
- (2) Two unrelated women each suffer from a complex disease where family, twin and adoption data suggest that genetic susceptibility is important. The disease is twice as common in men as in women. Anne is the only affected person in her

family; Betty's brother and son are also affected. For each of the following four comparisons, decide whether the risk is:

- (a) higher
 - (b) lower
 - (c) the same
 - (d) impossible to predict from these data
 - (1) The risk of a child of Anne's being affected, compared to the risk to the child of an affected man who, like Anne, is the only affected person in his family.
 - (2) The risk of a son of Anne being affected, compared to the risk to a daughter of hers.
 - (3) The risk of a son of Betty being affected, compared to the risk to a daughter of hers.
 - (4) The risk Anne's next baby will be affected compared to the risk to Betty's next baby.
- (3) A mutation arises on a chromosome bearing a particular haplotype of markers. What would be the half-life (in generations) of the association between the mutation and a marker showing (a) 1% and (b) 5% chance of recombination in each meiosis?
- (4) In the 1950s the statistician Ronald Fisher argued that the known association between smoking and lung cancer did not mean that smoking caused lung cancer. He suggested that incipient lung cancer causes an irritation in the lungs that drives people to smoke; or alternatively that people with a certain nervous constitution have a tendency both to develop lung cancer and, independently, to take up smoking. How would you prove him wrong?
- (5) Locus A has three alleles, A*1, A*2 and A*3 with gene frequencies 0.5, 0.4 and 0.1, respectively. The linked locus B has three alleles B*1, B*2 and B*3 with gene frequencies 0.6, 0.3 and 0.1, respectively. Which of the following would be evidence of linkage disequilibrium?
- (a) The frequency of the haplotype A*1, B*1 is 0.30.
 - (b) The frequency of the haplotype A*2, B*2 is 0.14.
 - (c) The frequency of the haplotype A*3, B*3 is 0.03.
 - (d) The frequency of the haplotype A*3, B*2 is 0.01.
- (6) Which of the following is evidence for linkage disequilibrium between cystic fibrosis and the *KM19* DNA polymorphism?
- (a) In a comparison of ten populations, the one with the highest prevalence of CF also had the highest frequency of *KM19* allele 2.
 - (b) *KM19* allele 2 is found on 91% of chromosome 7s carrying CF, but only 25% of those not carrying CF.
 - (c) The *CFTR* and *KM19* loci both map to chromosome 7q31.2.
 - (d) In family studies, the *CFTR* and *KM19* loci show close linkage.
- (7) In a bean-bag containing a very large number of black beans, one bean in 100 is red. Eyes shut, you put in your hand and take out a bean. If you are given ten tries, what is the chance you will pick out at least one red bean? Now apply this argument to a search for susceptibility genes. You test 1000 markers, each with 4 alleles, for association with a disease you are studying. What *P* value for an association will be significant at the 5% level? Would the answer be the same if instead of testing for association, you tested 1000 markers distributed across the genome for linkage?

- (8) One million SNPs are tested in a panel of 500 affected people and 500 controls. SNP 629 380 has 2 alleles that are present at frequencies of 50% each in the controls. What is the threshold number of affected people who must have SNP allele 1 for this to show significant association with the disease?
- (9) At a disease susceptibility locus the relative risk of disease for genotypes 1–1, 2–1 and 2–2 is 4:2:1. Calculate the expected proportion of affected sib pairs who are both 1–1, both 2–1 or one each 1–1 and 2–1 if the parents are:
- (a) $1-1 \times 2-2$
- (b) $1-1 \times 2-1$

[Hints on questions 1, 2, 3, 7 and 9 are provided in the *Guidance* section at the back of the book.]

14

What clinical services are available for families with genetic disorders?

LEARNING POINTS FOR THIS CHAPTER

After working through this chapter you should be able to:

- Understand how genetic services are organized
- Understand what clinical geneticists and counselors do and how they interact with other medical disciplines
- Understand the purpose of multidisciplinary teams (MDTs) and how they add value to clinical care
- Describe the common pediatric and adult indications for referral to a genetic clinic
- Describe the value of diagnosis and counseling for patients and families
- Describe the process of syndrome diagnosis
- Give examples of problems with puberty or reproduction that may be appropriate for genetic referral and investigations
- Give examples of common teratogens and the problems they cause
- Describe the main methods of prenatal diagnosis, their uses and risks
- Describe the approaches that are being made to the management and treatment of genetic diseases
- Understand the potential for better medical care stemming from advances in genetic knowledge

14.1. The work of the genetics service

There are no new cases for this chapter, which will instead draw on all the previous cases used throughout the book. Reflecting this difference, the structure of this chapter differs from that used in all the previous chapters.

Genetic medicine is a rapidly evolving specialty responding to changes in technology and new knowledge about the underlying genetic causes of disease, leading in some instances to targeted treatments. The provision of genetic services varies considerably, but almost all countries with well-developed healthcare systems now have specific provision for individuals and families with, or at risk of, genetic disorders. Most genetic

services are delivered from multidisciplinary centers where medically trained clinical geneticists work alongside genetic counselors, genetic nurses and laboratory scientists to serve a geographical population. Many centers offer subspecialty services including pediatric, cancer, cardiac, prenatal, neurology, ophthalmic and audiology genetics with associated research and training programs.

Genetic medicine can be distinguished from most other specialties in that it provides services to patients of all ages affected with disorders of any body system, and to their family members who may be healthy but at risk. Although the specialty was originally concerned only with rare disorders, now patients with common disorders such as cardiac disease and certain cancers which may have an inherited basis can benefit from the services.

As the cases illustrate, patients with genetic disorders will need to be seen by many doctors and other healthcare professionals over their lifetimes. It is important that care is integrated so that patients may benefit from advances in knowledge in all areas. Genetic medical care needs to be integrated with that delivered by other specialties to ensure optimum management. Increasingly, care for patients with rare disorders is delivered by multidisciplinary teams (MDTs) in centers of expertise who can develop best practice, management and treatment guidelines and liaise closely with colleagues working in local and general healthcare settings. With the introduction of exome and genome sequencing into routine care, interpretation of variants becomes ever more important. After initial analysis utilizing international databases, this is now mainly done in MDT meetings involving scientists, bioinformaticians and clinicians to ensure that the team agrees on pathogenicity of a variant and the links with the patient's phenotype and how this knowledge can be used in clinical care. Specific mutations may make a patient eligible for a particular treatment, and a decision can quickly be made by all appropriate clinicians in an MDT.

Increasingly, access to sequencing technologies is available to all medical specialists. This 'genome first' approach means that, for many patients, their case may be discussed by an MDT rather than being referred to a genetic clinic. Complex cases may be referred where the geneticist's role could be described as 'reverse phenotyping' and they may need to undertake functional genomic studies to determine pathogenicity.

Genetic centers are a major source of information for families and support groups, and for other professionals in health and social care and in education. Research and development of new services are an integral part of this fast-moving specialty as are training programs for the wider health community.

Reasons for genetic referral

Common pediatric indications for referral include:

- abnormalities of growth: overgrowth and short stature (including skeletal dysplasias)
- neurodevelopmental problems +/- seizures
- isolated learning disability
- multiple malformations +/- learning disability
- single malformation, e.g. cleft lip/palate, congenital heart defects, renal anomaly

- family history of Duchenne muscular dystrophy, cystic fibrosis, sickle cell disease
- disorders identified through screening, such as inborn errors of metabolism
- sensory abnormalities, especially of vision and hearing
- inherited skin disorders.

Common adult indications for referral include:

- family history of cancer, especially breast, ovary and bowel
- family history of a cancer-predisposing syndrome such as neurofibromatosis types 1 or 2, or von Hippel–Lindau syndrome
- to investigate the possibility of Marfan syndrome
- family history of cardiomyopathy or a cardiac rhythm disorder
- family history of neurodegenerative disorders such as Huntington disease
- family history of later onset neurological or neuromuscular diseases such as myotonic dystrophy, inherited ataxias
- family history of monogenic vision or hearing disease such as retinitis pigmentosa (OMIM 312600, 268000, etc.) or inherited late-onset deafness
- family history of other monogenic disorders such as adult polycystic kidney disease (OMIM 173900), familial hypercholesterolemia or disorders of connective tissue including Ehlers–Danlos syndromes
- reproductive genetic issues (*Box 14.1*) and worries about events during pregnancy (*Box 14.2*)
- increasingly, because a person has had a direct-to-consumer test and is concerned about the ‘results’.

Typically, when a patient or family is referred to a genetic medicine department by a primary care physician or another specialist they are seen in an outpatient clinic,

Reproductive genetic issues

Common reproductive issues that may trigger a referral to a genetic clinic are described below. Many genetic centers were founded to offer services to prospective parents wishing to know their risks of having a child affected with a genetic disorder, and to take advantage of emerging techniques for prenatal diagnosis. For many centers this is still the case, but for others much of the clinical work is carried out in obstetric departments, with samples being sent directly to genetic laboratories; only complex cases get referred to the genetic clinic.

Recurrent miscarriage

It has been estimated that 10–15% of all clinically recognized pregnancies end in a miscarriage. Recurrent miscarriage is defined as the loss of three or more pregnancies. Only a proportion of women presenting with recurrent miscarriage will have an identifiable underlying cause for their pregnancy losses. Based on a retrospective audit and cost–benefit analysis, the Royal College of Obstetricians and Gynaecologists now suggest cytogenetic analysis should be performed on products of conception of the third and subsequent *consecutive* miscarriage(s). Parental bloods should be karyotyped only where testing of products of conception reports an unbalanced structural chromosomal abnormality (RCOG, 2011).

Primary amenorrhea

Primary amenorrhea is the absence of menstrual periods, often in association with the lack of other signs of puberty in a woman. Usually baseline hormonal investigations and often genetic

tests will have been carried out by a gynecologist before referral to a genetic clinic. Diagnoses to be considered include:

- Turner syndrome. This disorder usually presents in infancy or childhood with short stature (see **Case 9, Isabel Ingram** in *Chapter 2*), but some women are not diagnosed until they present with primary amenorrhea in their teenage years. Around half the patients with Turner syndrome have a 45,X karyotype, with the rest having a structural abnormality of one X chromosome or a mosaic karyotype comprising a 45,X cell line plus one or more other cell line(s) including 47,XXX, 46,XX and 46,XY.
- Androgen insensitivity syndrome (AIS). This disorder, previously known as testicular feminization syndrome, is associated with a 46,XY karyotype and mutations in the androgen receptor gene on Xq11. Girls with AIS are phenotypically normal females at birth. They can present with inguinal herniae containing testes, or later as young adults with primary amenorrhea.

Other causes include:

- congenital absence of the uterus and vagina but with normal ovaries; these patients may have normal pubertal changes except for menstrual periods
- gonadal agenesis/dysgenesis: around 20% of XY females have mutations or deletions of the *SRY* gene, the master male-determining gene on the Y chromosome; they may have a uterus and streak gonads
- hypogonadotropic hypogonadism; the causes that are known are genetically heterogeneous.

Infertility

Infertility is defined by the failure to conceive after 12 months of unprotected intercourse. It is beyond the scope of this book to consider all causes, but genetic causes must be borne in mind, particularly those associated with male infertility.

- Klinefelter syndrome (47,XXY) has a prevalence of 1/600 – 1/800 male births and often presents in adult life with infertility. Boys enter puberty normally but by mid puberty the testes are smaller than normal, testosterone production is decreased and there is azoospermia.
- Congenital bilateral absence of the vas deferens (CBVAD; OMIM 277180) causes obstructive azoospermia and is usually due to mutations in the *CFTR* gene. Such males have a much higher incidence of partially functional alleles such as p.R117H or the 5T splice variant, but the other allele may be a common CF mutation such as delta-F508 (p.F508del); thus if such patients undergo sperm extraction procedures a child may be at risk of CF if the mother is also a carrier.

Other causes include:

- microdeletions or structural abnormalities of the Y chromosome; up to 15% of males with azoospermia have deletions of azoospermia factors on Yq
- 46,XX males, often due to a translocation of the *SRY* gene into the X chromosome
- Kallmann syndrome (OMIM 308700) comprising hypogonadotropic hypogonadism and anosmia.

Precocious puberty

This is defined as the appearance of signs of pubertal development at an abnormally early age. Generally for girls this is before 8 years and for boys before 9 years of age, but there are variations between populations; better nutrition and obesity are causing a drift towards earlier puberty. Rare genetic causes include:

- congenital adrenal hyperplasia (21 hydroxylase deficiency; OMIM 201910) in males
- McCune–Albright syndrome (OMIM 174800), a disorder comprising polyostotic fibrous dysplasia of bones and patchy skin pigmentation; it is due to a somatic activating mutation of the *GNAS1* gene in a mosaic form.

although some inpatients are seen for urgent consultations when there are pregnancy complications, or after the birth of a baby with abnormalities, or when a patient is acutely ill with a suspected genetic disorder. Prior to the consultation, information may be gathered from medical records and from the family and, particularly for cancer genetic referrals, verification of the precise diagnosis is obtained. Textbooks still have a place as important information resources to consult before a clinic, especially when the patient is known to have a rare disorder, and excellent books are listed in *Section 14.6*.

A consultation starts with construction of a pedigree (see *Chapter 1*), followed by the taking of a detailed history of the medical condition in the affected person(s), and then a physical examination. Investigations may then be ordered (*Section 14.2*). It is important to inform patients about any implications that results of tests may have for them and for other relatives, and to reassure them that if a DNA sample is stored, it will not be used for any purpose other than diagnosis without further consent. It may be possible at the first appointment to make a diagnosis on clinical grounds, or a follow-up appointment may be needed after results of investigations have been received. The clinical geneticist will usually write a summary letter for the patient after the consultation, with copies to the referring doctor and other specialists. If no diagnosis is made, particularly in children with a dysmorphic disorder, a review appointment may be arranged for a year or two later when new conditions may have been delineated or diagnostic technology has changed. This was the case for the **Meinhardt family (Case 12)** in *Chapter 4* where SNP arrays eventually revealed the cause the Madelena's problems. However, 'genome first' testing is short-cutting the traditional pathway for some patients. For those patients where no diagnosis was identified despite whole exome or whole genome sequencing, a review appointment may include a request for reanalysis of sequence data because new disorders are being reported all the time. *Disease box 8* illustrates one major collaborative approach to arriving at new diagnoses.

Concerns about a pregnancy

A common reason for referral in pregnancy to a genetic clinic is concern about possible effects upon a baby of maternal illness, of exposure to drugs and other substances, and of infections.

Maternal illness

- **Diabetes.** Diabetic mothers have an increased risk of miscarriage and their infants have around a 6–9% risk of major congenital malformations. Excellent diabetic control reduces the risks but even in very well controlled women the risks remain elevated compared to non-diabetic women. The main abnormalities seen include cardiac defects, neural tube defects and abnormalities of the skeletal system.
- **Phenylketonuria.** Previously, women with PKU had severe learning disability and rarely reproduced, but effective screening programs and dietary treatment means that such women are now normal and are having babies. Dietary control may lapse in adolescence: an untreated woman has a very high risk of having a baby with microcephaly, growth retardation and congenital heart defects and so it is important that women with PKU who are planning a pregnancy resume strict dietary control (see *Figure 10.8*).
- **Maternal genetic disorders.** In pregnancy, as well as the risk of the child inheriting the condition from a parent, the mother's genetic disorder may directly confer added risks. For example, a mother with myotonic dystrophy risks serious polyhydramnios (excess liquor) during pregnancy

and the child is at risk of the severe congenital form of the disease. An achondroplastic woman risks severe respiratory compromise during pregnancy, and because of the need for early delivery her baby may suffer effects of prematurity.

Drugs

These can be divided into drugs prescribed for maternal illnesses and so-called recreational drugs. A few examples are given below.

- Anticonvulsants. Around 0.4–0.7% of pregnant women have epilepsy. Overall there is a definite increase of developmental abnormalities, around two to three times, over the background risk for non-exposed children. Exposure to valproate seems to confer particular risks. The European Medicines Agency in 2014 and the Medicines and Healthcare Products Regulatory Agency (MHRA) in the UK in 2015 stated that evidence suggests children exposed *in utero* to valproate are at a high risk of serious developmental disorders (in up to 30–40% of cases) and/or congenital malformations (in approximately 10% of cases), and that valproate should not be given to women of childbearing age unless other treatments are ineffective or not tolerated.
- Warfarin crosses the placenta and if a fetus is exposed, particularly in the second half of the first trimester of pregnancy, there is a high risk that the baby will have severe skeletal effects, such as chondrodysplasia punctata, which manifests as short limbs with stippled epiphyses, and hypoplasia of the nasal bone.
- Exposure to other drugs has been linked to specific abnormalities in infants. Examples include heart disease, particularly Ebstein anomaly of the tricuspid valve, in babies exposed to lithium; central nervous system and heart defects and abnormalities of the first arch in babies exposed to retinoids; and choanal atresia, hypoplastic nipples and scalp defects in babies exposed to carbimazole.
- Alcohol. Fetal alcohol syndrome is a well-described condition which can be diagnosed confidently if there is a clear history of maternal alcohol ingestion and if the baby exhibits typical neonatal withdrawal symptoms such as poor feeding, constant crying and jitteriness, and clinical features such as low birth weight, microcephaly, short palpebral fissures and a long and flat philtrum. Both chronic exposure throughout pregnancy and episodes of binge drinking are associated with fetal alcohol syndrome. Most countries now recommend that pregnant women abstain completely from alcohol in pregnancy.
- Cocaine is known to induce vascular constriction and this is thought to be the basis of the observed increase in miscarriage and placental abruption in women abusing cocaine, and problems in their babies such as limb deficiencies, intestinal atresia and intracranial hemorrhage.

Infections in pregnancy

- Chickenpox – varicella is generally a mild illness in childhood but if a pregnant woman is one of the approximately 10% of non-immune adults there is a small risk, if she is exposed, of fetal varicella syndrome (FVS) in her baby. This involves damage to the brain and eyes, skin scarring and limb hypoplasia. The woman may also be at risk herself of severe pneumonia. Treatment with acyclovir soon after the rash appears may reduce the risks (RCOG, 2015).
- Rubella is also a mild illness in childhood and adult life, but congenital rubella can be a devastating condition. Maternal infection at 8 weeks' gestation results in severe effects in 90% of exposed babies, but this falls to around 10% by 20 weeks and very low risks thereafter. Features include growth retardation, learning disability, cataract, deafness and heart defects.
- Other infections associated with fetal effects include cytomegalovirus (a member of the herpes virus family where primary infection *in utero* is associated with growth retardation, microcephaly, hepatosplenomegaly, jaundice and thrombocytopenia) and toxoplasmosis caused by the parasite *Toxoplasma gondii* acquired by eating contaminated foods.
- Zika virus is a mosquito-borne flavivirus found in Africa, the Americas, Asia and the Pacific. Zika virus infection during pregnancy is a cause of microcephaly and other congenital abnormalities in the developing fetus and newborn.

Consanguinity

This was considered in more detail in *Chapter 9*. Box 9.3 showed how to calculate the proportion of genes that relatives share. In practice, couples who are related as cousins and planning to have children, and who come from communities where consanguinity is a customary practice, rarely request referral to a genetic clinic, although those from communities where this practice is uncommon sometimes do. More often referral follows the birth of an affected child.

- The empiric data usually given, where there is no family history of recessive disease, is that the birth prevalence of serious congenital and genetic disorders diagnosed by 1 year in unrelated parents is 2.0–2.5%, and for children of first cousin parents the risk is double that at 4.0–4.5%. For children of second cousins the risk is 3.0–3.5%. However, in some communities the risks are higher because a couple may be more closely related due to multiple consanguineous marriages in previous generations.
- The Birmingham birth study (Bundy and Aslam, 1993) found that the prevalence of recessive disorders in Northern European children, where only 0.4% of parents were related, was 0.28%, whereas in British Pakistani children, where 69% of parents were related, the prevalence of recessive disorders was 3.0–3.3%.
- In communities where there is a high prevalence of consanguineous marriages, some genetic services offer so-called cascade testing (see *Section 12.3*), where after the birth of an affected child other relatives are offered genetic screening. If both partners are found to be carriers of a trait they can be offered prenatal diagnosis.

14.2. Diagnosis and testing

The importance of a diagnosis

As the cases in the text illustrate, family history, clinical observations and examination and investigations have to date been the first steps in establishing a diagnosis. A precise diagnosis can make a huge difference to the lives of families affected by genetic disease that manifests in infancy or childhood. Patients and their families whose conditions are undiagnosed can feel very isolated and numerous studies have underlined the importance of a diagnosis for families, but also for clinicians and others involved in their care.

- A diagnosis is the cornerstone of clinical management. Without it, clinical management is unfocused and complications not anticipated.
- When the diagnosis is of a lethal condition, e.g. trisomy 18, the most appropriate management is usually supportive care, even if a structural malformation is present where surgery would normally be undertaken.
- Families will often consult many doctors and their child may be made to undergo multiple tests in search of a diagnosis. Getting to the right specialist is important, although the so-called diagnostic odyssey has been shortened greatly since the introduction of whole exome and whole genome sequencing.
- Establishing a diagnosis and providing information about a condition has real therapeutic value for many people, even if there is no cure. Links can be made with support groups, the families have something to write on forms, and educational and social services are more readily provided.

What if the diagnosis remains unknown?

A child should not be labeled as having a particular syndrome unless the physician is absolutely sure. It is more difficult to remove an incorrect diagnosis than to attach one. Where a syndromic diagnosis is still likely it is important for the child to be followed up, when new syndromes may have been delineated or improved technologies are available. If whole exome or whole genome sequencing has been undertaken, re-analyze the data in the light of new knowledge.

Dysmorphology

In the 1960s David Smith from the USA first used the term 'dysmorphology' to describe the study of human congenital malformations and patterns of birth defects. Dysmorphology is one of the subspecialties within genetic medicine, dealing with patients who have congenital malformations and syndromes. Terms used in dysmorphology are defined and illustrated in *Box 14.3*. To make a syndrome diagnosis the steps followed are essentially the same as for other clinical situations, i.e. history, examination, investigations and synthesis. However, a different emphasis is placed on several aspects compared to other clinical situations.

History

This concentrates particularly on:

- family history and past obstetric history
- maternal health – some maternal diseases, e.g. diabetes or systemic lupus erythematosus, may confer a higher risk of fetal abnormality; as described above, mothers with epilepsy, particularly those on anticonvulsant medication, may have a 2–3 times increased risk of fetal abnormality
- maternal vitamin supplements and drug use – are they likely to be teratogenic?
- pregnancy history – it would be relevant to know, for example, whether abnormalities were detected on scan, whether any invasive procedures were carried out, and whether there was any problem with liquor volume.

Observation

This requires the following to be taken into account; some diagnoses can be suggested by observing a child prior to examination:

- posture and tone – the characteristic flexed posture of the fingers and extended legs in trisomy 18, for example, or marked hypotonia in Prader–Willi and Down syndromes
- movements and behavior patterns – these are very characteristic in some syndromes: a girl with Rett syndrome will have repetitive hand movements (see *Disease box 11*) and individuals with Smith–Magenis syndrome (OMIM 182290) may hug themselves
- facial expressions – these may be typical in some syndromes: an individual with myotonic dystrophy has a mask-like face with poor facial movement, and the happy smiling face of Angelman syndrome is unmistakable
- personality – characteristic personality can be observed in patients with certain syndromes such as Williams syndrome where there is a friendly and talkative manner.

Physical examination

This should include documentation of:

- height and weight – these should be plotted on an appropriate growth chart; parental heights should be taken into consideration
- proportions – these can be altered in certain conditions, e.g. skeletal dysplasias or Marfan syndrome
- measurements of head circumference, facial features and other body parts where appropriate – these should be plotted onto charts for normal ranges and for specific conditions
- major and minor abnormalities – use correct terms; with minor anomalies be aware of what is abnormal and what is normal variation, e.g. minor 2/3 toe syndactyly.

It is useful to document general appearance and anomalies by taking photographs if the patient permits. Sequential photos of children at different ages are especially helpful in studying the evolution of phenotypes.

Synthesis

- Ask some basic questions; ‘Does the child have a single malformation or multiple malformations?’, ‘Is there intellectual disability or developmental delay?’, ‘Are there deformations that tie in with the pregnancy history?’, ‘Does the family history help?’.
- Think whether you have seen a similar pattern before. Do you recognize a “gestalt” which is familiar to you from a previous case or from the literature?
- When chromosomal syndromes have been ruled out and a monogenic cause is suspected, consider possible syndrome diagnoses in ‘syndrome families’ for which diagnostic panels have been developed including:
 - skeletal dysplasias
 - overgrowth syndromes.
- Seek help from the literature. There are numerous textbooks: *Gorlin’s Syndromes of the Head and Neck* (2010) and *Smith’s Recognizable Patterns of Human Malformation* (2013) are particularly comprehensive texts.
- Search databases. Use some of the emerging online facial recognition systems such as FACE2GENE (www.fdna.com/) which now incorporates the London Dysmorphology Database.
- Seek help from colleagues. Dysmorphic syndromes can involve all body systems and it is impossible for one person to be an expert in all areas, so it is often necessary to refer for a specialist opinion or to large scale research projects aimed at identification of rare diseases. Share information and photographs/ images with other colleagues within a department and specialists in the field. Present distinctive cases at dysmorphology meetings. If exome or genome sequencing has been undertaken, consider if any ‘variants of unknown significance’ are plausible candidates.

Terminology used in dysmorphology

Malformation: a morphologic abnormality that arises because of an abnormal developmental process (a primary error in morphogenesis), e.g. cleft lip.

Malformation sequence: a pattern of multiple defects resulting from a single primary malformation, e.g. talipes and hydrocephalus can result from a lumbar neural tube defect.

Malformation syndrome: a pattern of features, often with a unifying underlying cause, that arises from several different errors in morphogenesis ('syndrome' from the Greek 'running together').

Deformation: distortion by a physical force of an otherwise normal structure.

Disruption: destruction of a tissue which was previously normal.

Dysplasia: abnormal cellular organization within a tissue resulting in structural changes, e.g. within cartilage and bone in skeletal dysplasias.



(a)



(b)



(c)



(d)



(e)



(f)

Box figure 14.1 – Clinical photographs of the main types of dysmorphic features.

(a) Cleft lip, a **malformation** representing failure of fusion of components of the upper lip. (b) Meningomyelocele, talipes and hydrocephalus, a **malformation sequence** due to failure of closure of the neural tube and consequent effects. (c) Trisomy 13, a baby with a **malformation syndrome** consisting of holoprosencephaly, midline cleft lip and palate, polydactyly and heart defects. (d) Talipes, abnormal position of the feet, a **deformation** often due to extreme lack of liquor *in utero*. (e) Amniotic bands, **disruption** of a normal hand by constriction with strands of amnion leading to amputation and secondary fusion of finger tips (syndactyly). (f) Femur bones with multiple fractures and abnormal modeling due to osteogenesis imperfecta, a skeletal **dysplasia**.

Investigations

Like any other clinician, a clinical geneticist will call on a wide range of tests, many of which are not specifically genetic. Making a diagnosis can involve all the usual clinical skills, tests and investigations. If a specific disorder is strongly suspected a single test may be requested, although in many centers whole exome or whole genome sequencing are used routinely as a first-line test.

- Genetic tests, including cytogenetic, molecular genetic and metabolic tests are summarized below.
- Infection screening can be helpful where congenital infection is suspected from the history or from clinical signs such as a rash, hepatosplenomegaly or cerebral calcification.
- Radiological investigations: X-rays are crucial in the diagnosis of skeletal dysplasias. CT scans are useful to look for intracranial calcification; otherwise MRI scans provide more information and do not expose patients to radiation.
- Pathology/autopsy. Pathology investigations are useful for defining the full extent of abnormalities. With fetal pathology it is important to take into account the gestation of the fetus and the possibility of traumatic abnormalities sustained during delivery.
- Other miscellaneous investigations may be needed for specific disorders, e.g. Hb electrophoresis for the X-linked α -thalassemia–mental retardation syndrome (ATR-X, see *Disease box 11*) and testing for leucopenia and retinal pigmentation in Cohen syndrome (OMIM 216550).

Genetic testing

- Cytogenetic studies are indicated in babies with multiple congenital malformations, in patients of any age with a combination of intellectual disability and dysmorphism, and for otherwise unexplained infertility, recurrent miscarriages or intersex states. Techniques such as karyotyping, MLPA and array-CGH or SNP arrays are still used routinely in many laboratories. FISH is still utilized where a particular microdeletion syndrome is suspected. Mosaic chromosome disorders may not be detectable by analysis of lymphocytes, and buccal or skin tests (karyotype, array or FISH) may be needed. Chromosome breakage studies may still be indicated in some patients, particularly in those who are small, have microcephaly and other features such as radial aplasia and café-au-lait patches which suggest the diagnosis of Fanconi syndrome (OMIM 227650). Cytogenetic analysis of tumors (karyotyping or FISH, see *Figures 7.5, 7.10*) aids diagnosis and prognosis, although whole genome sequencing is increasingly being used as an alternative.
- Molecular testing is central to investigation of mendelian conditions and tumors. Next-generation sequencing has greatly expanded the scope for molecular testing, but various other options were discussed in *Chapter 5*. With the recognition that groups of disorders (e.g. the RASopathies, see *Disease box 3*) and many types of heart disease and eye and hearing disorders have a monogenic but highly heterogeneous basis, referrals to genetics services (clinical and laboratory) for these indications have greatly increased. For next-generation sequencing the relative advantages and disadvantages of sequencing gene panels, the whole exome or the whole genome were discussed in *Section 5.4*.
- Metabolic testing. Testing for disorders of metabolism involving amino acids, organic acids, mucopolysaccharides or peroxisomes should be considered in patients with symptoms such as hepatosplenomegaly, coarse facies, joint stiffness, severe hypo- or hypertonia, early onset seizures or deterioration of consciousness. Exome sequencing, rather than biochemical testing, is being used in many centers, especially for newborns or children with an acute presentation. Although most metabolic conditions present early, some such as Gaucher disease Type 3 (OMIM 231000) can present in adult life.

Genetic testing can be for diagnostic or other purposes, as discussed below. Diagnostic testing does not in general raise the same ethical issues as predictive testing, but clinicians need to be sensitive to possible implications for other family members. The special case of prenatal testing is discussed in more detail below. New technologies bring new ethical issues and there is much current discussion about reporting of so-called incidental findings after whole exome or whole genome sequencing (see *Section 12.4*). Incidental findings in this context are variants that do not cause the problem which was being investigated, but nonetheless can confer a risk of developing another condition.

Carrier testing is used for autosomal and X-linked recessive conditions, and also for balanced chromosomal abnormalities. It would normally be done at the request of the patient, and for a reason beyond mere curiosity. Cascade testing may be offered in certain situations where the risk to relatives is high (see the **Smit family, Case 24**). Counseling should always be part of the process, though when testing is done as part of a population screening program, pre-test counseling may be quite minimal and limited to written material. Children should not be tested unless there is an immediate benefit to the child.

Predictive testing is used in a clinical setting for late-onset conditions such as Huntington disease and familial cancers. Ideally the family mutation will have been identified through testing an affected individual. The laboratory procedures may be straightforward, but predictive testing needs to be undertaken within the context of detailed written protocols that define in advance the response to each possible outcome and allow individuals to have adequate information on which to base their decision about whether to proceed to testing. *Box 14.4* shows a protocol for predictive testing for Huntington disease, as offered to **John Ashton (Case 1)**.

Predictive testing for Huntington disease

Contributed by Dr Rhona MacLeod, St Mary's Hospital, Manchester

Predictive testing for Huntington disease by direct mutation analysis has been available in many countries since 1994. At the outset there was concern about how mutation-positive individuals would cope with their test result. Encouragingly, serious psychological sequelae are rare and most people adjust to their result (positive or negative) over time. This has led to a gradual shift from standardized testing protocols, necessary at the outset for research purposes, to a process more tailored to the individual. The scheme below provides an overview of the testing process. Molecular confirmation of the diagnosis of Huntington disease in the family is obtained where possible prior to the predictive test.

Predictive test session 1

Discussion of:

- Individual experience of Huntington disease
- Knowledge and understanding of Huntington disease
- Motivation and timing of predictive test
- Potential impact of test result on self and family (particularly partner)
- Discrimination/insurance issues
- Alternatives to having the test
- Options for participating in research
- How the test works

Current research including progress with disease-modifying treatment trials
Possible test outcomes (including intermediate and reduced penetrance range)

Predictive test session 2 (at least 1 month after predictive test session 1)

Further counseling session
Rehearsal of plans for disclosure of result/follow up
Neurological examination (optional)
Signing of consent form
Taking blood sample
Setting a date and time for result session

Result session (2–6 weeks after predictive test session 2)

Planned result session as face to face appointment
In cases where access to a genetic center is difficult, may be done remotely, e.g. by telemedicine, by prior arrangement
Plan for follow up agreed
Result should not be copied to the GP (or any third party) without the patient's prior consent
Where consent is given to inform the GP, the predictive nature of the test should be made clear (to distinguish from a diagnostic test for a mutation-positive result)

Follow up

Contact with patient following test result
Timing of follow up appointments guided by patient (and depending on protocols if participating in research); usually:

Follow up session 1–3 months post result

To see how individual and family are coping with result
For individuals found to carry the HD expansion offer annual follow up
Combined clinical/research appointment may be an option in some centers
Important to remember to offer follow up to individuals receiving a mutation-negative result

References

Losekoot M, van Belzen MJ, Seneca S, et al. (2013) EMQN/CMGS best practice guidelines for the molecular genetic testing of Huntington disease. *Eur. J. Hum. Genet.* **21**: 480–486.

MacLeod R., Tibben A., Frontali M, et al. (2013) Recommendations for the predictive genetic test in Huntington's disease. *Clin. Genet.* **83**: 221–231.

Direct to consumer (DTC) testing – many people are now responding to advertisements for genetic tests, either for health reasons or to trace ancestry (discussed in *Section 12.4*). The results relevant to health are on the whole limited to mildly worrying or mildly reassuring statements about the risk of various common multifactorial illnesses, and can mostly only recapitulate standard public health advice to take exercise, to stop smoking, to not drink too much alcohol, and to maintain a healthy weight. However, as discussed in *Section 12.4*, customers should be extremely wary about any reports of rare high-risk variants, such as *BRCA1/2* variants, if based on SNP genotyping. When researchers compared microarray-based genotypes with sequencing data from 50 000 individuals from the UK Biobank, they found that the great majority of *all* ultra-rare pathogenic variants reported by microarrays were false positives, and there was also a high rate of false negatives (Weedon *et al.*, 2019). This was not an isolated example of poor laboratory practice, but is inherent in the ways microarrays report variants. Microarrays (when properly used) report common variants

accurately, but are completely unreliable for ultra-rare variants such as those responsible for mendelian diseases. The outcomes could be devastating if clinical decisions were made on such results. Testing based on sequencing should avoid these problems.

Prenatal diagnosis merits separate discussion. Imaging is the only thoroughly established non-invasive method for routine fetal testing. With each passing year imaging becomes more powerful and sophisticated. 3D and 4D imaging can now give really clear images of fetal structure and even facial features. A detailed fetal anomaly scan (different from the scans used to establish gestational age and check the number of sacs) is often done between 18 and 20 weeks of gestation. Imaging is also employed as a screening test to detect micro signs of Down syndrome.

Fetal cells obtained by chorion villus biopsy or amniocentesis (see *Box 14.5*) have long been karyotyped for prenatal diagnosis of trisomies and other aneuploidies. DNA testing can detect smaller variants (microdeletions, microduplications and copy number variants) using microarrays, while standard DNA analysis techniques (see *Chapter 5*) have been used to diagnose pathogenic single gene variants. Much effort is being devoted to using fetal DNA obtained noninvasively from maternal blood for testing (see *Disease box 12*), and indeed a proof of principle experiment succeeded in reconstructing the entire fetal genome using such DNA – but that is far from being current clinical practice. Most tests use fetal DNA obtained by chorion villus biopsy or amniocentesis. The American College of Obstetricians and Gynecologists Committee on Genetics published a Committee Opinion in December 2013 (ACOG, 2013) which stated the use of microarray testing in pregnancy was most beneficial after the detection of structural abnormalities by ultrasound scan. Whole exome or whole genome sequencing of fetal DNA has been evaluated in several studies following ultrasound diagnosis of structural variants, and is being introduced into clinical practice in several countries. Two carefully designed studies found remarkably consistent results: the PAGE study (Lord *et al.*, 2019) reported a clinically significant genetic variant in 8.5% of fetuses, with an additional 3.9% harboring variants of possible clinical significance (12.4% in total), and the study by Petrovski *et al.* (2019) found a diagnostic genetic variant in 10.3% of fetuses.

Rather than leave the diagnosis until a pregnancy has been established, with the consequent risk of terminating the pregnancy, the diagnosis might be made on pre-implantation embryos obtained by *in vitro* fertilization, with only genetically normal embryos being returned to the mother (*Box 14.6*).

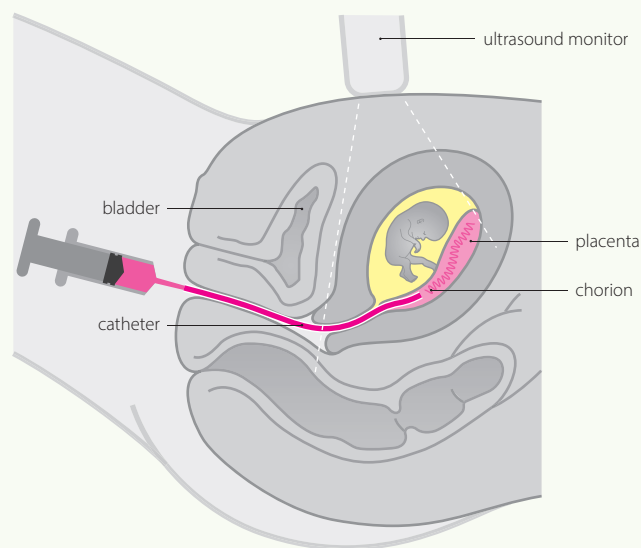
Decisions about prenatal diagnosis would be governed by several factors:

- the legal and ethical framework in the country – in some countries prenatal testing is legal only for a defined list of conditions; in others it is left to negotiation between parents and clinician
- the practical availability of testing, including whether somebody else – the state or an insurance company – will pay for it
- the severity of the condition – for example, many people would not regard hearing loss as an indication for termination of affected fetuses
- the availability and effectiveness of treatment – demand for prenatal diagnosis of cystic fibrosis, for example, depends on views about the current and future prospects for treatment
- the age of onset of the condition – many people do not see late onset conditions such as Huntington disease as appropriate for prenatal diagnosis

Obtaining fetal material

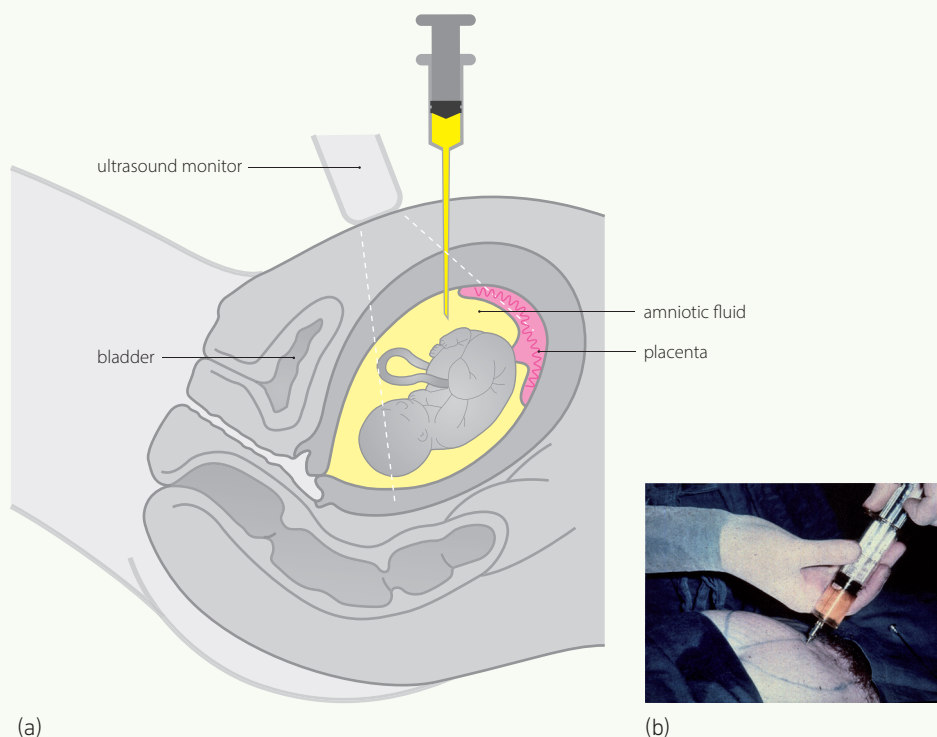
Small fragments of fetal DNA, shed by the placenta, are present in the maternal circulation, together with very occasional cells. As described in *Disease box 12*, these are being increasingly exploited as a means of noninvasive prenatal diagnosis. Initial applications used PCR for fetal sexing (checking for Y chromosome sequences) and testing for rhesus incompatibility. With next-generation sequencing it became possible to test for fetal trisomies; this is currently regarded as a screening test, with positive results needing confirmation by amniocentesis. Further developments of the sequencing procedure may allow routine detection of smaller chromosomal abnormalities and, potentially, any abnormality detectable by DNA sequencing. However, if undegraded fetal DNA or good numbers of fetal cells are needed these must be obtained invasively by amniocentesis or chorion villus biopsy.

- *Chorion villus biopsy.* The chorion is the outermost of the fetal membranes around the placenta. Biopsy is usually performed between 11 and 13 weeks of gestation under ultrasound guidance (*Box figure 14.2*). Either a transabdominal or transcervical route may be used; the sampling instrument should not penetrate the amniotic cavity. Once removed, the material needs expert dissection under the microscope to pick fetal material free of contaminating maternal tissue. Villi can be used for DNA extraction or for rapid cytogenetic analysis of dividing cells already present. Such short-term cultures need confirming with longer term cultures. For DNA testing, results should always be compared to a control sample of the mother's blood DNA, to ensure that the test result reflects the fetal genotype. Mosaicism detected in villi is difficult to interpret: retrospectively it often turns out to have been confined to the placenta. Chorion villus biopsy carries around a 2% additional risk of causing a miscarriage. There is a correlation between experience of the operator and success in obtaining chorionic villi at first attempt and with a lower complication rate.
- *Amniocentesis* provides material that comes directly from the fetus, rather than from the placenta as with chorion villus biopsy. It is performed between 14 and 20 weeks of gestation (*Box figure 14.3*) and carries a 0.5–1.0% risk of causing a miscarriage. Amniotic fluid consists mainly of fetal urine and washings from the lungs. It can be analyzed biochemically, or fetal cells can be isolated from the fluid and cultured for cytogenetic or molecular analysis. Culture of cells from amniotic fluid for cytogenetic analysis takes around 2 weeks to obtain good quality preparations.



Box figure 14.2 – Chorion villus biopsy.

Amniotic fluid is a poorer source of DNA than chorionic villi, because there are fewer cells. Techniques that do not require cell culture, such as QF-PCR (quantitative PCR with fluorescence-labeled primers, see *Section 4.4*) to detect specific trisomies, have now become standard.



Box figure 14.3 – Amniocentesis.

(a) Diagram of procedure. (b) Withdrawing amniotic fluid.

- the situation of the individual family – how well could they cope, emotionally, physically and financially with the birth of an affected child?
- the moral principles or religious beliefs of the individuals concerned – some people would not countenance a termination, whereas others would see it as the least bad option.

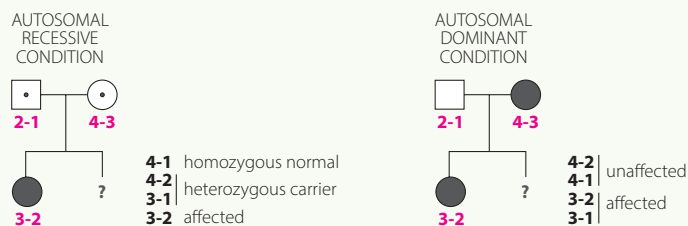
Benefits of testing

As well as hopefully establishing a diagnosis, testing is a crucial part of the overall aim of allowing families affected by a genetic condition to live as normal lives as possible. This is well illustrated by a study by Modell *et al.* (1980) on families at risk of serious forms of thalassemia. Before prenatal diagnosis was available, when the birth of an affected child made a couple aware of their 1 in 4 risk of having further affected children, they virtually ceased reproduction. Most succeeding pregnancies were accidental and they sought termination of 70% of them. Prenatal diagnosis for this condition became available in 1975. At-risk couples then resumed normal reproduction, with fewer than 30% of pregnancies being terminated for thalassemia. The effect of prenatal diagnosis was to reduce the number of terminations and to enable at-risk couples to have normal families.

Pre-implantation genetic diagnosis

Pre-implantation diagnosis seems an attractive option for people wanting prenatal diagnosis but unwilling to consider terminating affected pregnancies. Embryos are obtained by *in vitro* fertilization and one or more cells are removed for genetic testing. This is usually done at the 8-cell stage (day 3) by removing a single blastomere cell; alternatively, several trophoctoderm cells may be sampled from a blastocyst on day 5. Either of these procedures is far from simple and not always successful; there is also a risk that trophoctoderm cells may not be representative of the genetic constitution of the embryo. Possible genetic tests include:

- **FISH to check for chromosomal abnormalities.** A problem is the high frequency of mosaicism in pre-implantation embryos, much of which seems to self-correct during later development, so that FISH analysis of a single cell has a significant risk of giving a false positive or sometimes a false negative result.
- **PCR testing for a pathogenic variant present in the family.** A high level of laboratory expertise is required to perform PCR reliably on a single cell. The main problems are contamination or allele dropout, where an allele that was present in the cell is not present in the PCR product.
- **Gene tracking** uses linked non-pathogenic markers to check for transmission of a pathogenic variant (*Box figure 14.4*). The principle was described in *Box 8.1*, and an example was shown in *Figure 4.15* (though the latter case was complicated by John Ashton's unwillingness to know his own genetic status). *Box figure 14.4* shows how this would work for a couple who already have a child with either an autosomal dominant or autosomal recessive condition; similar logic can be used for other pedigree structures or modes of inheritance.



Box figure 14.4 – Gene tracking: using a linked non-pathogenic marker to follow the transmission of a pathogenic variant. The numbers represent different alleles of a multi-allele marker. The conclusion would be incorrect if there was recombination between the disease and marker loci.

In reality, to reduce errors, the test would use a panel of linked markers (genetic haplotyping) rather than just a single marker. The main advantage over direct PCR testing for the pathogenic variant is that for a given disease a suitable marker panel can be developed and validated in single cell assays well in advance of any live diagnostic test, and can be used for every case of that disease regardless of the actual pathogenic variant.

Couples would normally be advised to have standard prenatal diagnosis in any resulting pregnancy to exclude the risk of a false negative pre-implantation genetic diagnosis result.

A similar story can be told of Duchenne muscular dystrophy. Because many cases are due to new mutations, assessing carrier risks of female relatives is very uncertain when based only on pedigree data. Before the dystrophin gene was mapped in 1982, the only option for a woman who wished to avoid the birth of an affected boy was fetal sexing and termination of any pregnancy where the fetus was male – in the full knowledge that for a woman at 30% carrier risk, 85% of those wanted pregnancies would have produced a

healthy boy. Once the gene was mapped, gene tracking with linked markers (the principle was explained in *Box 14.6*) could be used to better define carrier risks. Direct mutation detection became possible when the gene was cloned in 1987. Provided the family mutation could be identified in an affected male, women could be accurately classified as carriers or non-carriers, and carrier women could be offered prenatal diagnosis, rather than just fetal sexing. *Figure 14.1* shows how the distribution of carrier risks for Duchenne muscular dystrophy among families known to one large genetics center has changed as new knowledge allowed more precise genetic testing.

These examples illustrate several aspects of genetic services. First, how rapidly scientific advances can be delivered into clinical service; secondly, how options for individual family members can change even when the disease remains incurable; and thirdly, the importance of long-term follow up of families. The girls at intermediate risk in *Figure 14.1* will be offered definitive carrier testing when they are of age.

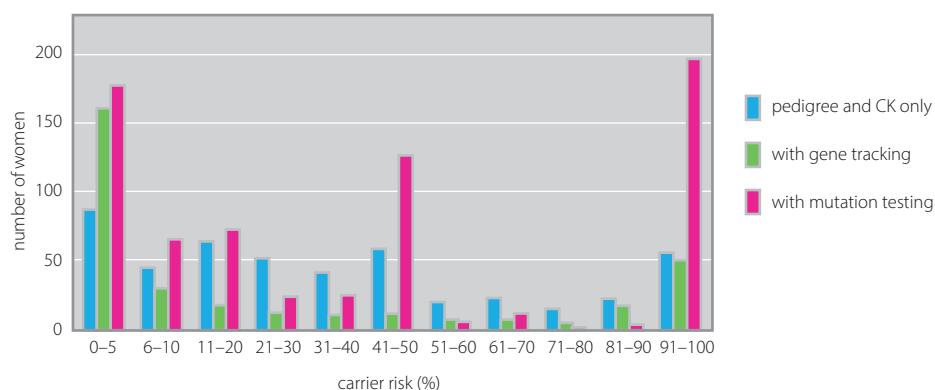


Figure 14.1 – Distribution of DMD carrier risks (%) among at-risk women in families on the North West Genetic Register, Manchester.

Blue bars: risks based on pedigree and creatine kinase (an indicator of muscle damage), before any molecular testing was available. Green bars: risk distribution in 1989 among women where the pedigree structure allowed gene tracking. Pink bars: risks when direct mutation testing became available. The peaks at intermediate risks in the latter series are mostly from girls currently too young to test. Data courtesy of Dr Elizabeth Howard, St. Mary's Hospital, Manchester.

14.3. Counseling and risk estimation

Genetic counseling is essentially an information-giving and communication process. A Task Force of the USA National Society of Genetic Counselors (NSGC) in 2006 (Resta *et al.*, 2006) developed the following definition of genetic counseling:

“Genetic counseling is the process of helping people understand and adapt to the medical, psychological and familial implications of genetic contributions to disease. This process integrates the following:

- *Interpretation of family and medical histories to assess the chance of disease occurrence or recurrence.*

- *Education about inheritance, testing, management, prevention, resources and research.*
- *Counseling to promote informed choices and adaptation to the risk or condition”.*

A diagnosis and counseling can help dispel the guilt and anger that can seriously impact the quality of life in families after the birth of a child with a malformation or disease. It is not enough simply to tell the parents that it wasn't caused by something one or other of them did. The professional skills of a counselor can help parents work through the natural reactions of shock, grief and anger.

- Counseling can relieve the burden of anxiety by dispelling exaggerated general estimates of recurrence risks and, especially in X-linked and recessive diseases, identifying family members who are at negligible risk. Contrary to popular belief, more people have good news and low risks given in genetic clinics than bad news and high risks.
- While counseling cannot abolish the real problems of living with an affected family member, it can help people focus on solutions rather than problems. Appropriate help can be given. Couples can be given a range of options for handling the recurrence risk.

Risk assessment

Genetic counseling involves much more than giving out recurrence risks – but an indispensable start is to get the recurrence risk right! Those giving risks need to have enough understanding of the science and the methods for risk calculation to be able to justify a quoted figure, even if they did not themselves calculate it.

- For *mendelian conditions* the increased availability and success of mutation testing have made this part of counseling much easier over the past two decades. Where risks are based just on the pedigree, the main difficulties come with serious dominant and X-linked conditions where new mutations are frequent and many cases may be mosaics. In these cases Bayesian methods are important tools (see *Box 14.7*). Counselors (including clinicians offering counseling) need to have enough understanding of Bayesian methods to be able to follow and justify a calculation, whether or not they themselves normally do the calculations.
- For *chromosomal conditions* such as trisomies, recurrence risks are empiric risks. Where a parent carries a balanced abnormality (as with **Ellen Elliot, Case 5**) cytogenetic colleagues should be consulted about the risks of the various unbalanced outcomes. Although each case is a one-off, cytogeneticists can give guidance based on the geometry of meiotic pairing, for example, in the quadrivalent in a carrier of a reciprocal translocation.
- For *complex diseases* risks are empiric. The most important point is to use data that are up to date (risks change with changing incidences) and relate to the appropriate ethnic group. The recent development of **polygenic risk scores** (see *Section 13.4*) has provided a tool for providing personalized risk estimates. The scores might be used to identify a small subset of individuals who are at particularly high risk, although there is still much controversy about their applicability in real clinical situations.

An introduction to Bayesian calculations in genetics

This method of combining probabilities was invented in the 18th century by the Reverend Thomas Bayes. It has turned out very useful for calculating genetic risks. It starts with a **prior probability** – how likely is a hypothesis in the first place? It then allows you to bring in relevant evidence supporting or opposing the hypothesis (conditional likelihoods) and to combine these to obtain an overall or posterior probability. To apply this method:

- (1) Set out each possible mutually exclusive hypothesis that you are testing. Cover all possibilities, so that one or other of your alternatives must be true. Usually there are just two alternatives – individual X either is or is not a carrier of this disease – but sometimes there are more (you might want to calculate the probabilities that X is aa, Aa or AA).
- (2) Assign a prior probability to each. In genetics these are usually the mendelian 1 in 2, 1 in 4, etc. probabilities. They must add up to 1.
- (3) Considering your first additional piece of evidence, for each alternative hypothesis in turn, write down the likelihood of having made that observation, *if that alternative were the true one*. These are the conditional likelihoods. They do not necessarily add up to 1.
- (4) If there are other relevant observations that are completely independent of the first one, repeat step 3 for each such observation.
- (5) When the list is complete, multiply the numbers down each column of the table. The results are called joint probabilities.
- (6) Since the final probabilities for each of the possible hypotheses must add up to one (i.e. one of them must be true), you must scale the joint probabilities so that they sum to 1. Do this by dividing each one by the sum of all the joint probabilities. The result is the final probability of each hypothesis, in the light of the prior probability and all the additional observations.

To illustrate it, here is how to calculate the risk that the healthy sister of a child with cystic fibrosis is a carrier. The steps are color coded.

Hypothesis – sister is:	AA	Aa	aa
Prior probability:	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Conditional: she's unaffected:	1	1	0
Joint probability	$\frac{1}{4}$	$\frac{1}{2}$	0
Final probability	$\frac{1}{4} / (\frac{1}{4} + \frac{1}{2} + 0) = 1/3$	$\frac{1}{2} / (\frac{1}{4} + \frac{1}{2} + 0) = 2/3$	0

1. Set out the alternatives. As the child of two carrier parents she could be AA, Aa or aa.
2. The prior probability is just the mendelian 1:2:1 probability.
3. This is the tricky bit. Remember, each conditional likelihood is the likelihood of the observation, *given that the particular hypothesis is true*. If she is AA then she will definitely be unaffected (likelihood = 1). Similarly if she is Aa. If she is aa there is no chance she would be unaffected (obviously you could introduce variable penetrances here, which is a major use of the method).
4. Multiply down each column.
5. Divide each joint probability by the sum of all joint probabilities, to make them sum to 1.

The *Guidance* for this chapter's *Self-assessment questions* give some more examples and discussion of Bayesian calculations. Bayesian approaches to questions of probability are attractive because they correspond to the way we decide whether or not to believe something in everyday life. As mentioned in the *Guidance* for SAQ1 of Chapter 8, you may well believe your friend if he tells you that he missed a lecture because he overslept, but not if he tells you it was because he had been abducted by aliens. You assess the overall credibility of his story in the light of its prior probability. Thus Bayesian calculations are a form of quantitative common sense.

Patients with rare disorders often feel very isolated, and indeed the majority of healthcare professionals they meet may know less about the disorder than the patients themselves. This is where support groups come in: most now have good websites, the groups also arrange support for newly diagnosed patients, organize meetings where families can share experiences, and promote research. A number of very helpful online information resources have been developed focusing on rare conditions, with sections for patients, professionals, researchers and industry. One of the largest is Orphanet (www.orpha.net) which has a base in every European country and is coordinated in France. It has extensive regularly updated summaries and reviews about rare diseases, in addition to regional information about clinics and laboratories. GeneReviews (www.ncbi.nlm.nih.gov/books/NBK1116/) is based in the USA and has similar information about clinics and laboratories as well as very helpful reviews on a range of rare conditions. Unique, a charity that set out to provide support and written information for families of people affected by rare chromosome disorders, have now extended their services to include information leaflets on many of the newly described developmental disorders that have been identified from large scale sequencing studies (www.rarechromo.org/).

14.4. Management and treatment

In medicine now, most treatments and management plans for common conditions are based on evidence from published studies, but for rare disorders evidence-based guidelines are few and far between. Studies to develop the evidence have not been done, mostly due to the rarity of the conditions and the general belief that genetic diseases are untreatable. A number of professional organizations and initiatives have now started to address this problem and to initiate studies to develop better levels of evidence using recognized methodology. While it is often not possible to reach the levels of evidence found in common diseases, there is benefit in drawing up guidelines for rare disorders based on expert opinion. The continuing development of centers and networks of expertise, together with much more emphasis on multidisciplinary and multispecialty working will allow faster collection of evidence for optimum management.

Treatment is available for a wide range of genetic diseases. Concentrating on the hopes for dramatic technologies such as gene therapy and stem cell therapy (see below) distracts attention from the incremental but very real advances that have been made in treating a whole range of genetic diseases. We are indebted to excellent articles by Munnich (2006) and Dietz (2011) for a list of some of these incremental advances. These articles should be consulted for more details and references. *Table 14.1* lists some approaches and examples.

Often knowledge of the mutation or even the gene is irrelevant to treatment. Munnich (2006) pointed out that patients do not suffer from their mutations, but from the functional consequences of those mutations. We do not need to know whether a genetic patient's kidney failure is due to Alport syndrome (OMIM 301050), polycystic kidney disease (OMIM 173900) or nephronophthisis (OMIM 256100) to know that a kidney transplant will greatly improve his life. We do not need to know which of the 100 possible genes has caused a child's deafness before we can consider the possible benefits of a cochlear implant. Even with Huntington disease, a famously 'untreatable' condition, life for the patient and his carers can often be improved by drugs such as neuroleptics and antidepressants, and attention to diet and the home environment.

Table 14.1 – Examples of genetic diseases where symptoms can be ameliorated or sometimes abolished by treatments based on knowledge of the malfunction, but not necessarily of the DNA sequence

Strategy	Example	Disease
Supply missing molecule		
	Insulin	Diabetes
	Growth hormone	Pituitary dwarfism
	Factor VIII	Hemophilia A
Replace defective enzyme		
	Acid beta-glucosidase	Gaucher disease
	Alpha-galactosidase	Fabry disease
	Alpha-glucosidase	Pompe disease
Dietary supplementation		
	High carbohydrate diet	Glycogen storage diseases
	Cholesterol	Smith–Lemli–Opitz syndrome
	Mannose	Carbohydrate-deficient glycoprotein syndrome 1b
	Biotin	Biotin-responsive carboxylase deficiency
	Pyridoxine	Pyridoxine-responsive homocystinuria
	Cobalamin	Cobalamin-responsive organic aciduria
	Alpha-tocopherol	Pseudo-Friedreich ataxia
	Creatine	Creatine synthesis deficiency
Dietary restriction		
	Low phenylalanine diet	Phenylketonuria
	Low protein diet	Maple syrup urine disease
	Low fat diet	Hypercholesterolemia
	Avoid phytanic acid	Refsum disease
Enhance residual enzyme activity		
	Fibrates	Fatty acid oxidation disorders
Remove a toxic product		
	Bleed regularly	Hemochromatosis
	Cysteamine	Cystinosis
Block a pathogenic process		
	Nitisinone	Tyrosinemia Type 1
	Bisphosphonates	Osteogenesis imperfecta

Data from Munnich (2006).

In other cases a closer knowledge of the gene function, though not necessarily of the gene itself, is the key to treatment.

- In carbohydrate-deficient glycoprotein disease type Ib (OMIM 602579) the underlying defect is an inability to isomerize fructose to mannose (deficiency of mannose phosphate isomerase). The symptoms of profuse diarrhea and severe liver disease are entirely due to lack of mannose. They can be completely rectified by oral mannose.
- The example of Type 1 tyrosinemia (OMIM 276700) was discussed in *Section 10.2*. A drug, nitisinone, can be used to manipulate the tyrosine catabolism pathway (*Figure 10.4*) so as to convert this lethal condition into a phenocopy of the milder Type II disease.
- Cystinosis (OMIM 219800) is an atypical lysosomal storage disease. The defect is not in a lysosomal enzyme, but in a lysosomal membrane transporter protein. Non-functioning of the transporter means that cystine is unable to get out of lysosomes. The accumulation produces symptoms including renal failure, pancreatic insufficiency, corneal erosions, central nervous system involvement and severe myopathy. Oral cysteamine is an effective treatment. Cystine consists of two molecules of cysteine linked by an S–S bridge (*Figure 14.2*). Cysteamine, given orally, enters lysosomes using a specific transporter. There it can displace

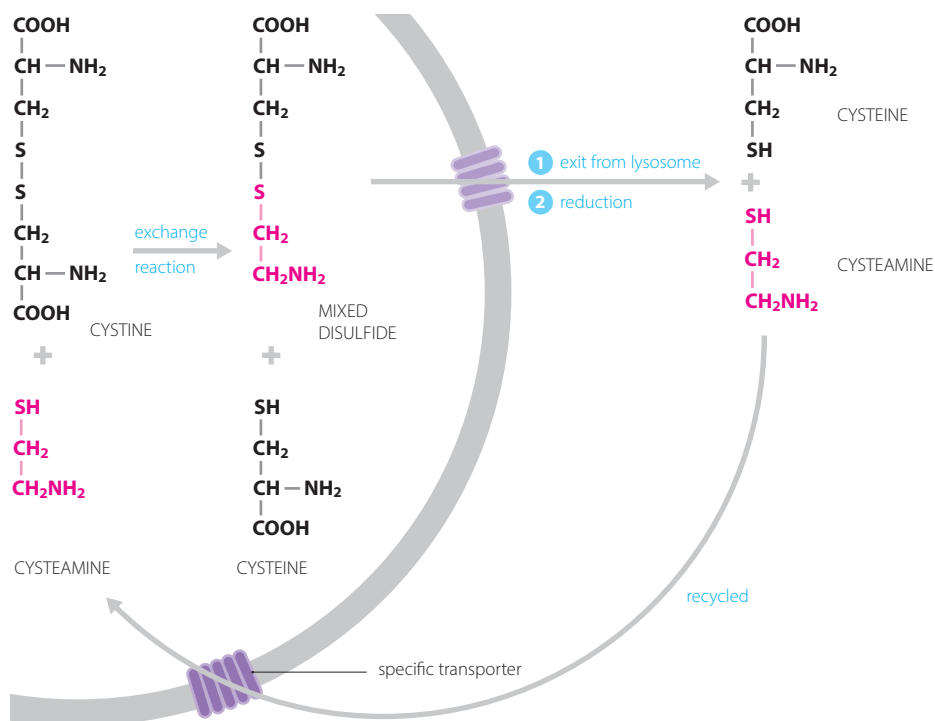


Figure 14.2 – Treating cystinosis with cysteamine.

Cystinosis patients lack the specific transporter protein that in normal people allows cystine to exit lysosomes. Their symptoms are caused by accumulation of cystine in lysosomes. Cysteamine and the mixed disulfide are able to cross the lysosomal membrane, and so relieve the symptoms.

one cysteine molecule from cystine to form a mixed cysteamine–cysteine disulfide, which is able to exit the lysosome, apparently using a lysine transporter. Once outside the lysosome, the mixed disulfide may be reduced, freeing the cysteamine to ferry another cysteine molecule out of the lysosome.

There is much excitement about the potential of using already approved drugs for new indications based on recent knowledge about the pathways in which certain gene products operate.

- **Losartan** is in a class of drugs known as angiotensin-receptor blockers, and has been used for the treatment of hypertension. It is known to act by blocking the action of the TGF β biochemical signaling pathway. Patients with Marfan syndrome (OMIM 154700) have raised TGF β signaling. Following promising results of a study using a mouse model of Marfan syndrome, clinical trials in humans were started. Results from the trial ‘Atenolol (a beta blocker) versus losartan in children and young adults with Marfan syndrome’ showed both drugs to be equally effective in slowing aortic root enlargement (a major cause of death in Marfan syndrome) and suggested there were benefits of starting treatment in childhood.
- **Sirolimus** is a drug used in transplant medicine. It acts by inhibiting the mTOR biochemical signaling pathway. Patients with tuberous sclerosis (OMIM 191100) have constitutive activation of this pathway. Following clinical trials in tuberous sclerosis patients with angiomyolipomas in their lungs, Sirolimus is now approved for clinical use with these patients.
- **Alpelisib** (BYL719) is an inhibitor of alpha-PI3K, developed as a treatment for hormone-resistant breast cancer. It had some success in trials for this indication, but has been repurposed for the PIK3CA-related overgrowth syndrome illustrated in *Disease box 11*, with spectacular results (Venot *et al.*, 2018).

Gene therapy offers a more radical approach to treatment of genetic diseases. Its promise rests on the existence of well-established laboratory methods for getting external genes into cells. It is surprisingly easy to get exogenous DNA into living cells in the laboratory, using any of a range of techniques (*Box 14.8*). The ultimate objective of gene therapy is to change gene expression in the relevant cells, such that the disease symptoms are cured or reduced without adverse effects. For many years hopes for effective gene therapy followed a manic–depressive course, with over-optimistic hopes of rapid progress dashed by major setbacks – but in recent years the field has been moving steadily forward in a much more realistic way. A search of the clinical trials database at www.clinicaltrials.gov using the search term ‘gene therapy’ produced 4073 results [accessed 10 August 2019]. The first gene therapy products have now been licensed for clinical use. *Box 14.9* describes using an adenoviral vector to treat spinal muscular atrophy (OMIM 253300). The whole field has been well reviewed by High and Roncarolo (2019).

However, the example of Glybera warns against simplistic belief in progress. This recombinant adeno-associated virus was approved in Europe for treatment of familial lipoprotein lipase deficiency (OMIM 238600) as early as 2012. But the extremely high cost (\$1 million per treatment) and the rarity of the targeted condition meant that almost nobody outside the successful clinical trials received the treatment, and the high cost to the manufacturer of maintaining the necessary cultures eventually drove them into bankruptcy.

Methods for inserting an exogenous gene into a cell

Broadly, these fall into physical methods and vector-dependent methods.

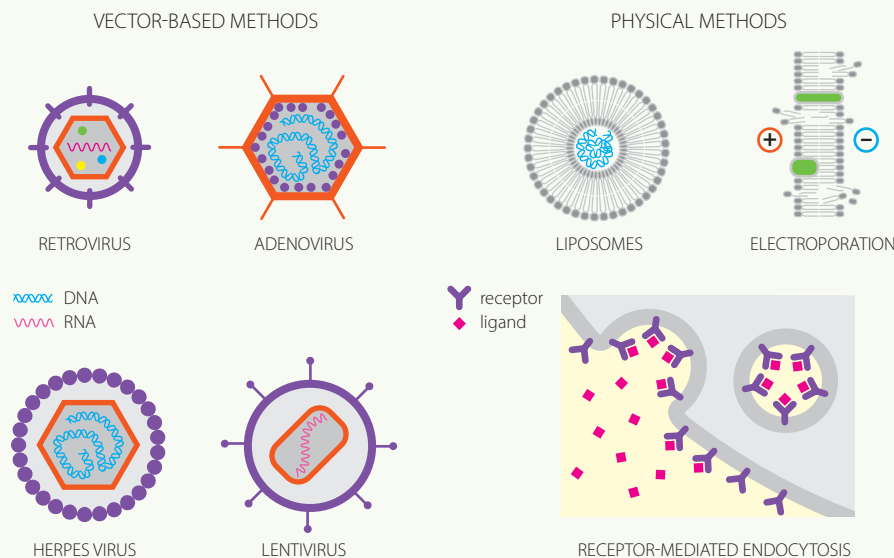
Physical methods include:

- *liposomes* – artificial membrane-bound vesicles that can fuse with cell membranes and release their contents into the cell
- *receptor-mediated methods* where the DNA is attached to the ligand for a cell-surface receptor which internalizes after binding the ligand
- *electroporation*, where a short high-voltage pulse temporarily alters cell membranes such that they take up naked DNA from the medium.

Vector-based methods use viruses engineered to be harmless and to carry a desired gene into target cells. Compared to physical methods, these are often much more efficient at getting the foreign DNA into a good proportion of the target cells. Many different viruses have been used. Factors governing the choice include:

- *capacity*: how long a piece of genetic material they can hold
- *tropism*: some viruses preferentially infect certain types of cell
- *ability to infect non-dividing cells*: retroviruses can only infect dividing cells
- *integrating or non-integrating vectors*: integrating vectors such as retroviruses integrate the transferred gene into a chromosome of the host cell, ensuring that all daughter cells will carry a copy. Non-integrating vectors such as adenoviruses remain as extrachromosomal **episomes**; they may remain in a cell for its lifetime but will be diluted out by cell replication.

Safety is of course a major consideration. With integrating vectors there is a risk of insertional mutagenesis (see the case of the **Portillo family, Case 21**, immunodeficiency), though this risk is mitigated by the use of safer lentiviral vectors. For non-integrating vectors the main risk is an immune response to the viral vector, since many people may have been exposed to the wild-type virus.



Box figure 14.5 – Methods of inserting an exogenous gene into a cell.

Depending on the nature of the condition it is designed to treat, gene therapy could have any of three aims.

Gene supplementation or augmentation aims to put a working gene into a cell that currently lacks it. This would be the aim of gene therapy for any loss of function condition, which includes many of the conditions covered in our *Case studies* (see below). Additionally, it could be used to introduce a novel gene into a cell, usually for the purpose of creating a vulnerability in cells that one wishes to eliminate. Cancer cells might be made to express a novel antigen that would trigger a cytotoxic attack by the immune system, or an intracellular enzyme that would convert a harmless pro-drug into a toxic metabolite. *Box 14.9* shows the very promising results of gene augmentation as a treatment for spinal muscular atrophy.

Gene silencing aims to prevent expression of a resident gene, either by inhibiting transcription or degrading the mRNA. This would be required for diseases caused by gain of function or dominant negative mechanisms. In most cases the silencing would need to be specific for the mutant allele, leaving expression of the normal allele unaltered. Gene silencing might also be used to prevent expression of viral genes in an infected cell. Synthetic oligonucleotides that can base-pair to the endogenous mRNA can be used to trigger degradation of the resulting double-stranded nucleic acid molecule. The oligonucleotides are usually chemically modified to make them resistant to nucleases.

Gene editing aims to correct the function of a malfunctioning endogenous gene or mRNA, rather than silence or replace it. The recent development of CRISPR–Cas technologies has made editing DNA sequences much more feasible than beforehand (Pennisi, 2013). These methods have quickly become standard tools in research laboratories, but before they can be used clinically some safety concerns must be addressed. Current protocols involve too many off-target effects to allow clinical use in humans. Rapid technical development is addressing that problem. At the mRNA level, synthetic oligonucleotides can be designed to bind to splice sites, changing the way the primary transcript is spliced and causing skipping or retention of specific exons. This is discussed below in relation to Duchenne muscular dystrophy, and in *Box 14.9* in relation to spinal muscular atrophy.

Any of these methods might be applied *ex vivo*, on cells taken from the patient which, after modification, will be returned, or *in vivo* by injecting or otherwise introducing the therapeutic construct into the patient's body. Equally, one could in principle treat either somatic or germ-line cells. Germ-line gene therapy has an appealing finality to it compared with somatic therapy – the problem is dealt with once and for all – but is generally regarded as ethically unacceptable. In any case, *Figure 14.3* shows that it would have little use for treating mendelian conditions. The only germ-line application currently under consideration is manipulation to avoid a woman transmitting a mitochondrial disease to her children (*Figure 14.4*).

DOMINANT CONDITION
1 in 2 embryos OK



RECESSIVE CONDITION
3 in 4 embryos OK

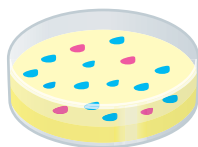


Figure 14.3 – The limited usefulness of germ-line gene therapy.

The targets for germ-line therapy would most likely be embryos created by *in vitro* fertilization. Typically IVF might result in 5–10 embryos, of which 1–2 are selected for implantation. A genetic test would identify which embryos would require the therapy. Depending on the mode of inheritance, either 1 in 2 or 3 in 4 embryos could be used without the need for any therapy.

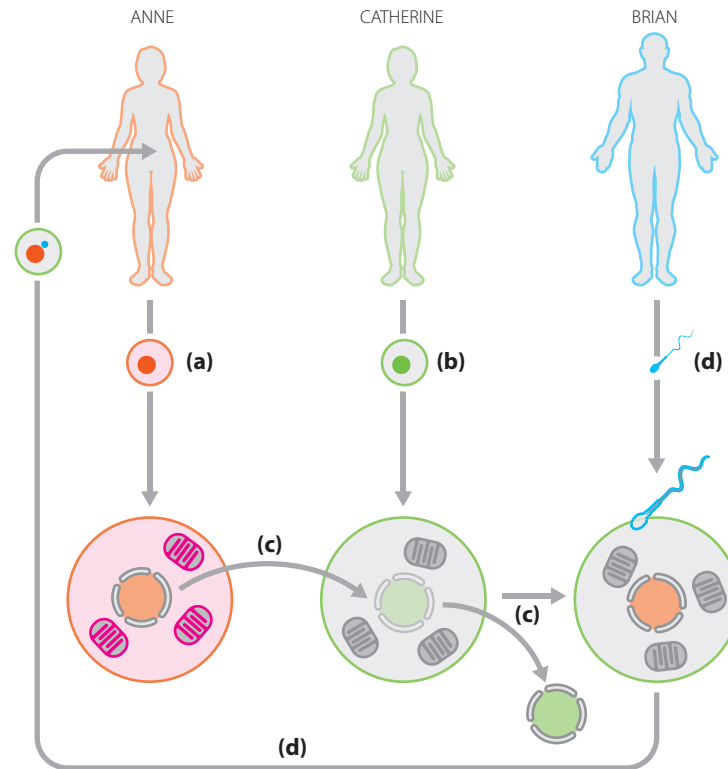


Figure 14.4 – Germ-line manipulation to avoid transmitting pathogenic mitochondrial DNA.

Anne and Brian wish to avoid having a child with Anne's mt-DNA mutation. (a) Anne provides an unfertilized egg, (b) Catherine donates an unfertilized egg, (c) the nucleus is removed from Catherine's egg and replaced with the nucleus from Anne's egg, and (d) it is fertilized with Brian's sperm. The resulting zygote is implanted in Anne. It will develop into a child inheriting all its nuclear genome from Anne and Brian, but its mitochondrial genome from Catherine.

Treatment of spinal muscular atrophy

Spinal muscular atrophy (SMA, OMIM 253300) is characterized by muscle weakness and atrophy resulting from progressive degeneration and loss of the anterior horn cells in the spinal cord and the brain stem nuclei. Symptoms of weakness can occur from before birth to adulthood. SMA used to be classified into four subtypes dependent on the age of onset, but in reality SMA is a continuum. The weakness is symmetric, proximal > distal, and progressive. In the most severe form it results in death before the age of 2 years.

SMA is an autosomal recessive disorder caused by loss of function of the *SMN1* gene on chromosome 5q13. As described in *Section 6.2*, the genetics is complicated by the presence of the highly homologous *SMN2* gene close by, which would appear to be able to encode a functional SMN protein, but has very low activity due to inefficient splicing. The commonest cause of SMA is homozygous intragenic deletion of the *SMN1* gene, although the genetic situation is often complex and complicated further by copy number or sequence variants in the neighboring *SMN2* gene. No correlation exists between the type of *SMN1* pathogenic variants and the severity of disease.

In the commonest and most severe form of SMA, which used to be called Werdnig–Hoffman syndrome, marked weakness and developmental motor regression are noted before the age of 6 months. Some infants gain head control but soon lose it and they do not learn to sit. There are reduced or absent deep tendon reflexes, and poor muscle tone. Fasciculation of the tongue is often seen and bulbar weakness results in problems with sucking or swallowing and recurrent aspiration. Weakness of the intercostal respiratory muscles with relative preservation of diaphragm musculature leads to the characteristic ‘bell-shaped’ chest (see Box figure 14.6). In some centers long term ventilator support along with gastrostomy feeding allowed longer survival, but without any prospect of improvement in the child’s longer term prospects. Cognitive function is normal.



Box figure 14.6 – Third child affected with SMA born to healthy parents.

He died at 6 months of age, similar to his siblings. Note finger contractures resulting from difference in rate of loss of anterior horn cells to flexor and extensor muscles, and bell-shaped chest resulting from degeneration of intercostal muscles.

Until recently the only management options were symptomatic, with ventilator and feeding support; however, two major therapeutic approaches given before or just after the onset of symptoms have improved the outlook enormously. These have led to the introduction of pre-symptomatic testing where there is a family history, and in some countries population neonatal screening is being proposed.

Nusinersen (Spinraza) is a modified antisense oligonucleotide designed to correct the splicing defect in the *SMN2* gene, thereby providing an alternative source of functional SMN protein. It has been used in all types of SMA due to *SMN1* variants. It is given by intrathecal injection in four loading doses and then by maintenance doses four-monthly. The cost in the USA is \$750 000 in the first year and \$375 000 per year thereafter, making it one of the most expensive of all drugs, and delaying its use in publicly funded healthcare systems.

A second promising approach (onasemnogene abeparvovec or Zolgensma) uses gene augmentation with a modified adenovirus to deliver functional copies of the *SMN1* gene. The target motor neurons are non-dividing, and a single dose of the non-integrating vector is used, delivered by intravenous injection. Clinical trials have been promising, and the drug is approved in the USA for treating children under 2 years of age with SMA. Cost, however, is a major concern, at \$2 125 000 for the single treatment.

Gene therapy is not the only way to change the level of gene expression. Alternative examples include the following.

- Symptoms of deficiency or abnormality of β -globin in sickle cell disease or β -thalassemia are milder in patients who continue to express some fetal hemoglobin during adult life (hereditary persistence of fetal hemoglobin). Patients in whom production of the fetal γ -globin has been shut down as part of normal development may be stimulated to re-express the gene by treatment with hydroxyurea.
- As mentioned in Section 1.3 and below, a clinical trial has demonstrated the feasibility of reducing the level of the mRNA as a possible treatment for Huntington disease.

- The antibiotic gentamicin can induce mis-reading of mRNA by ribosomes, such that the ribosome occasionally ignores a stop codon. Gentamicin has been used to improve symptoms in cystic fibrosis patients who have a nonsense mutation. Clearly it would be catastrophic to cause ribosomes to ignore a high percentage of stop codons, but even a very small degree of read-through may allow significant clinical improvement in cystic fibrosis. Other possible read-through drugs are also under development.

Cell therapy

Organ transplantation has been used successfully for decades, including for treating genetic diseases (e.g. kidney transplantation for polycystic kidney disease). Mending damaged tissues and organs by supplying new cells is more difficult. Cell therapy usually relies on a small number of transplanted cells taking root and multiplying. Most cells have a limited life and replicative potential, thus therapy requires the use of **stem cells**. It is believed that every tissue is maintained by a small population of stem cells. These cells have the potential to divide asymmetrically to produce another stem cell and a derivative cell that can go on to form the functioning cells of the tissue (Figure 14.5).

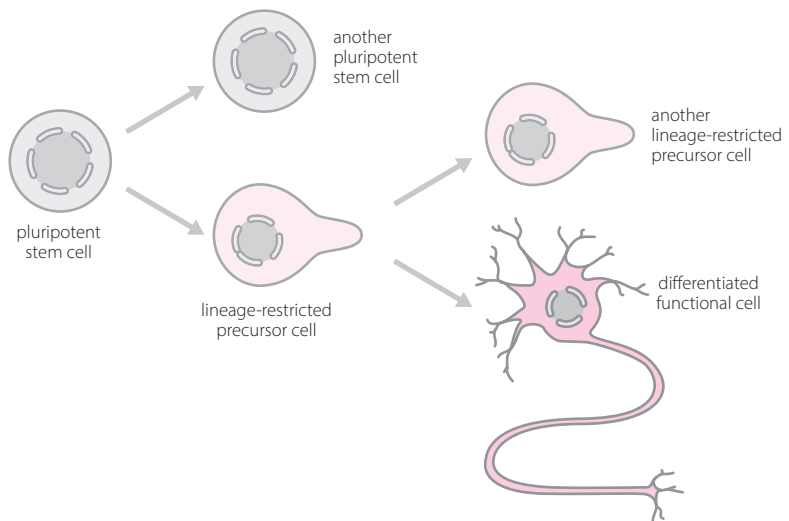


Figure 14.5 – Stem cells are capable of both self-renewal and differentiation into multiple lineages. Stem cells in the embryo can produce every cell of the adult body; other stem cells are more restricted in the range of cell types they can produce.

Stem cells differ in the range of cell types they can produce. Cells in the early zygote can produce all embryonic and extra-embryonic cells: they are *totipotent*. Cells from the inner cell mass of early blastocysts are *pluripotent*: they can produce all the cell types of the adult body, but not the extra-embryonic membranes or placenta. Later down the differentiation cascade are tissue-specific *multipotent* stem cells with more limited capabilities and *unipotent* types restricted to producing a single cell type.

Bone marrow transplantation represented the earliest clinical application of stem cells. More recently, cord blood has been widely used. It contains hematopoietic stem cells that can reconstitute every type of blood and bone marrow cell, and some other types

of cell too. It can be collected at every birth with no risk or pain to the mother or baby. Banks of cells, frozen in liquid nitrogen, have been set up and are routinely used in the treatment of leukemia and other blood disorders, and for bone marrow reconstitution in cancer patients who have had aggressive chemotherapy.

Specific tissue-restricted stem cells are being trialed in a wide range of applications, but the real excitement is over pluripotent stem cells. These can produce every cell type of the adult body. The two sources of human pluripotent stem cells are embryonic stem (ES) cells and induced pluripotent stem (iPS) cells (*Figure 14.6*). Mouse ES cells were first made in 1981; it took until 1998 to develop methods that produced human ES cells. ES cells can only be obtained by destruction of a very early embryo. This has made their production and use controversial. Most of the available ES cell lines have been derived from 'spare' blastocysts from IVF clinics. iPS cells avoid this problem. They were first produced in mice in 2006, and their human counterparts followed soon afterwards. A cocktail of transcription factors is used to reprogram adult cells to an embryonic-like state.

Already, iPS cells are proving immensely valuable tools for exploring molecular pathology and possible therapies. For example, fibroblasts from a patient with a neurodegenerative disease can be used to generate iPS cells that can then be differentiated into neurons carrying the patient's specific genetic mutations for functional study. Using iPS cells raises the prospect of being able to use a patient's own cells to make any required cell type for transplantation, thus avoiding problems of rejection. Their promise is immense, but before this becomes practical medicine much work is needed to define efficient ways to produce iPS cells and to reliably differentiate them into cells that can be safely implanted in patients. Meanwhile, dubious clinics in under-regulated countries are busy selling untested 'stem-cell' therapies to desperate patients.

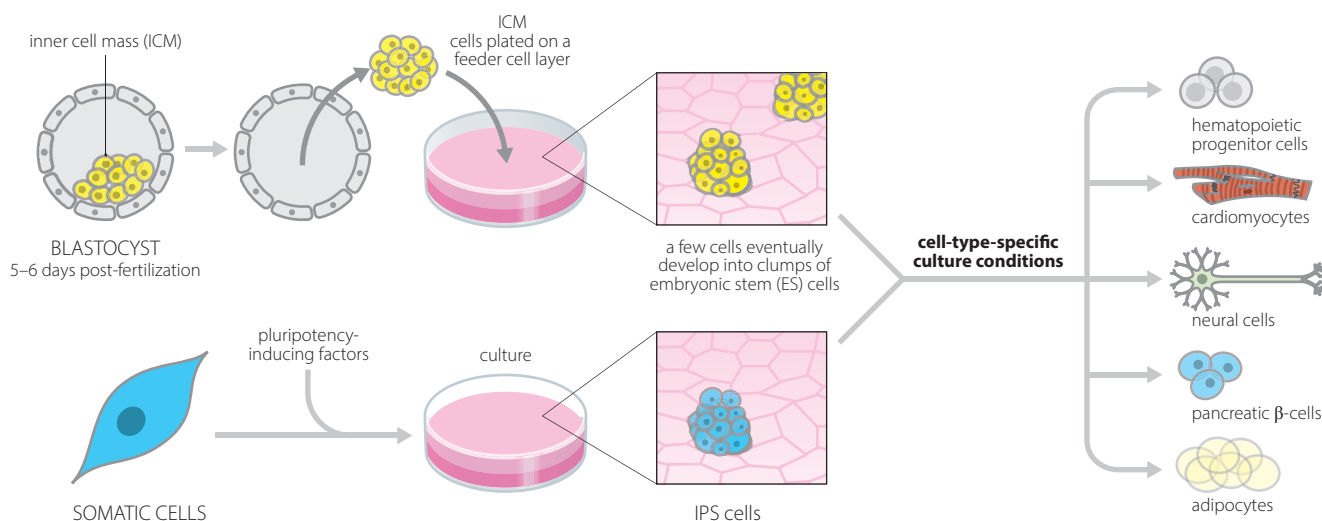


Figure 14.6 – Pluripotent stem cells.

Embryonic stem (ES) cells are derived from the 100–150 cell inner cell mass (ICM) of an embryo at the blastocyst stage, 5–6 days post-fertilization. ES cells have no exact natural counterpart in normal embryos. Induced pluripotent stem (iPS) cells are obtained by culturing adult somatic cells such as fibroblasts in the presence of a cocktail of transcription factors. Either type of stem cell can be induced to differentiate into any of a variety of cell types.

Management and treatment for Cases 1–26

Considering the 26 cases described in the previous chapters, we can distinguish those where the patient's current problems are the result of an irreversible developmental abnormality from those where the problems are the consequence of some malfunction here and now (*Table 14.2*). For the former group treatment is necessarily symptomatic; for the latter, in principle a complete cure might be possible if treatment is started early enough.

Table 14.2 – Summary of the 26 case studies used in this book

Case	Family	Conditions where the problem affected development	Case	Family	Conditions where the problem is a continuing malfunction
3	Kowalski	Intellectual disability	1	Ashton	Huntington disease
5	Elliot	Chromosomal anomaly	2	Brown	Cystic fibrosis
7	Green	22q11 deletion	4	Davies	Duchenne muscular dystrophy
8	Howard	Down syndrome	6	Fletcher	Leber optic neuropathy
9	Ingram	Turner syndrome	11	Lipton	Fragile X
10	O'Reilly	Stickler syndrome	13	Nicolaides	Thalassemia
12	Meinhardt	Chromosomal anomaly	15	Tierney	Acute lymphocytic leukemia
14	Jenkins	Achondroplasia	16	Wilson	Breast cancer
18	Choudhary	Hearing loss	17	Xenakis	Familial adenomatous polyposis
22	Qian	Angelman syndrome	19	Ulmer	Tay–Sachs disease
23	Rogers	Prader–Willi syndrome	20	Vlasi	Phenylketonuria
			21	Portillo	Severe combined immunodeficiency
			24	Smit	Familial hypercholesterolemia
			25	Yamamoto	Alzheimer disease
			26	Zuabi	Type 2 diabetes

For those in the left-hand column, treatment is necessarily symptomatic; for those in the right-hand, a cure is in principle possible.

Prenatal diagnosis could, in principle, be offered for almost all of these conditions. The exceptions would be leukemia (an acquired condition) and Alzheimer disease and Type 2 diabetes (complex diseases where genetic prediction is not possible). For most of the mendelian conditions it would only be possible if the causative mutation had been identified previously. Several of the chromosome abnormalities occurred *de novo* (in **Cases 7, 8, 9, 12, 22 and 23**). In these cases the recurrence risk is very low, and in general parents would be reassured rather than offered prenatal diagnosis in future pregnancies. However, we can never rule out the small possibility of germ-line mosaicism (except for **Case 23** where there was uniparental disomy). There is a finite though small recurrence risk for the other families. If the parents were particularly anxious it would be appropriate to offer prenatal testing, but also to point out that the chance of detecting a recurrence is probably smaller than the risk of losing the pregnancy because of an invasive procedure. Of course, the two risks are not necessarily equally important, and families will make their own judgment of the relative weights to put on them. In our cases, of those six families

only **Anne Howard (Case 8)**, Down syndrome) was tested in her next pregnancy. In her case she and her husband decided to have tests partly because non-invasive prenatal testing carried no risk to the pregnancy and partly because they felt they would not be able to give Helen all the attention she needed if they had another baby with Down syndrome.

Dietary management and enzyme replacement therapies for some inborn errors of metabolism have been around for many years. Dietary treatment of phenylketonuria (**Case 20, Vlas family**) comes nearest to being curative at present. Once diagnosed, the baby is put on a low protein diet to minimize its intake of phenylalanine. Some protein is of course necessary, as is some phenylalanine – it is an ‘essential’ amino acid that humans cannot synthesize but have to obtain from dietary protein. Careful titration of the phenylalanine level is necessary. Too low a level of protein or phenylalanine will cause malnutrition, retarded growth, etc., while too high a level will cause brain damage. Keeping a child on the diet is immensely demanding for the family, especially if there are unaffected siblings. Special phenylalanine-reduced flour is available so that the child can have its own cakes and biscuits. In most countries it is recommended that patients should remain on the diet all their lives, because MRI has documented white matter changes in teenagers and adults who discontinued the diet. Clinically, the consequences of non-compliance are less severe in adults; however, as shown in *Figure 10.8*, a phenylketonuric woman definitely needs to go back on the diet during pregnancy, or else the high phenylalanine level in her blood will damage her baby’s brain, even though genetically the baby is not phenylketonuric. Other examples of dietary or supplement treatments were given in *Table 14.1*.

Drugs are available to treat, if not cure, most genetic diseases. Of the cases in *Table 14.2*, noteworthy examples are:

- the recently developed treatments for cystic fibrosis (see discussion of **Case 2 – Joanne Brown**)
- the use of statins for familial hypercholesterolemia (**Case 24 – Smit family**)
- the major success of chemotherapy for acute lymphoblastic leukemia (**Case 15 – Jason Tierney**). The best centers report that 98–99% of children with ALL experience complete remission within 6 weeks of beginning treatment, and about 90% are leukemia-free for at least 10 years, at which time they are considered to be cured. Between 80 and 90% of adults with ALL will also have remissions following treatment. However, about half of those will experience a relapse, making the cure rate about 40% overall. The response to therapy is partly a question of the genetics of the leukemia cells – a variety of chromosomal rearrangements produce different chimeric oncogenes that influence prognosis – and partly a question of genetic polymorphisms controlling the pharmacokinetics (absorption, metabolism and clearance) and pharmacodynamics (response of the target) of the drugs used. Treatment response is measured by the level of residual disease after induction of remission. PCR of the specific chimeric oncogene is used to check the number of remaining cells with the leukemic genotype. In favorable cases this is 0.01% or less of the original level. Stem-cell transplantation is also often used, especially in adult cases. The cells are derived from cord blood from unrelated infants (see above). Jason Tierney had two prognostically favorable features: his *TEL-AML1*

translocation, and his thiopurine methyltransferase deficiency. Although the latter caused initial problems, it also meant that his cells had received very high effective doses of the drug.

For the *late-onset conditions* in *Table 14.2* there is the hope of prevention. The risk of type 2 diabetes (**Case 26 – Zuabi family**) can certainly be greatly reduced by physical activity and weight control. People with familial adenomatous polyposis (**Case 17 – Xenakis family**) are strongly advised to have their colon removed; unfortunately, there is still a risk of gastric and peri-ampullary tumors. An appropriate diet can reduce the risk of recurrence. Prophylactic mastectomy is one option for women who carry mutations in *BRCA1/2* (**Case 16 – Wilson family**); those with *BRCA1* mutations often also opt for oophorectomy. Regular enhanced surveillance is important for people at high risk of cancer.

In principle, gene therapy might be applicable to any of the conditions in the right-hand column of *Table 14.2*, where the problem is a continuing genetic malfunction. The easiest targets for gene therapy are loss of function diseases where the exact level of expression of the introduced gene is not important, and a low level could be clinically useful. This makes cystic fibrosis (**Case 2**), Duchenne muscular dystrophy (**Case 4**) and X-SCID (**Case 21**) the most appropriate targets from our list.

- A major problem with cystic fibrosis has been getting the therapeutic construct into enough target cells, especially in the lungs.
- For Duchenne muscular dystrophy, the huge size of the gene and the difficulty of getting a foreign gene into a high proportion of muscle cells make gene augmentation an unattractive option. However, promising preliminary results have been achieved with schemes to manipulate splicing. Two-thirds of DMD mutations are deletions of one or more exons and, as discussed in *Section 6.2*, the result is severe or mild disease depending whether or not the result is a frameshift. As mentioned above, oligonucleotides matching exon–intron boundaries can cause selective exon skipping. For patients with specific exon deletions, these can be used to restore the reading frame by inducing skipping of an extra out-of-frame exon. **Martin Davies (Case 4)** has an out-of-frame deletion of exons 44–48. If exon 43 could also be skipped the resulting larger deletion would be in-frame and probably have a milder effect (see *Table 6.2*).
- For X-SCID there is already major progress (*Figure 14.7*). Twenty children in Paris and London underwent a gene supplementation procedure during 1999–2006. They showed effective T cell recovery, but following the therapy, five developed leukemia; four went into remission but one died. It turned out that the retroviral vector used had integrated close to an oncogene, *LMO2*. The strong promoter used to cause high expression of the therapeutic *IL2RG* gene had the effect of up-regulating expression of *LMO2*. This is similar to the way the 8;14 translocation in Burkitt's lymphoma up-regulates the *MYC* oncogene by placing it next to the highly expressed immunoglobulin heavy chain gene (*Chapter 7*). Integrating vectors are required for long-term correction of dividing cells, but modified vectors have now been developed that greatly reduce the risk. To date, no adverse events due to insertional oncogenesis have been reported for this new generation of self-inactivating retroviral or lentiviral vectors for X-SCID (Rivers and Gaspar, 2015).

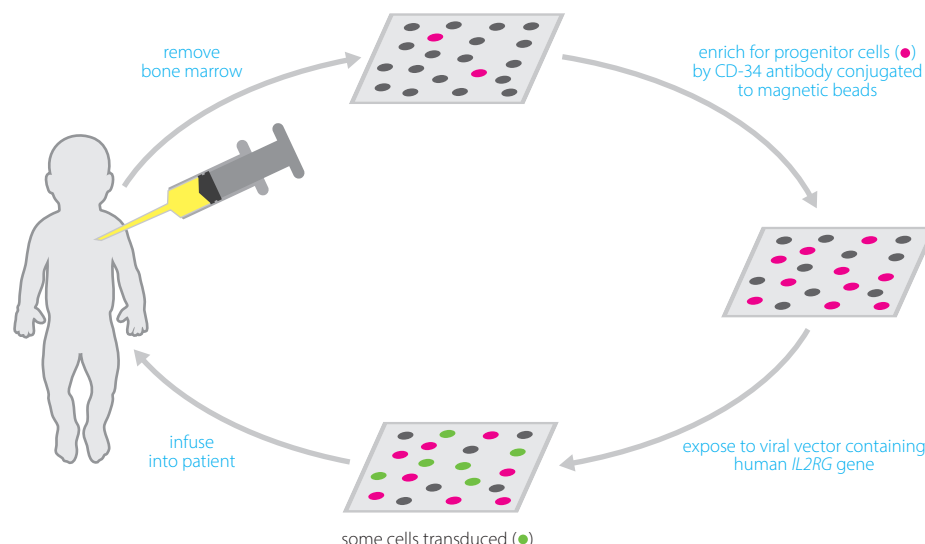


Figure 14.7 – Ex-vivo gene therapy for severe combined immunodeficiency.

The disease is caused by loss of function mutations in the *IL2RG* cytokine receptor gene. A viral vector is used to integrate a working copy of the gene into chromosomes of hematopoietic precursor cells. This gives them a selective advantage when re-infused into the patient, enabling them to reconstitute T-cell and NK-cell function in treated babies.

As a gain of function condition, Huntington disease (**Case 1**) would need a more sophisticated version of gene therapy than simple gene augmentation. A prospective treatment, currently in human trials, uses oligonucleotides to target the mRNA (Fischbeck and Wexler, 2019). Repeated treatments are needed by injections into the spinal fluid. The current trial is not specific for the mutant allele, but preliminary results suggests the overall reduction of HTT protein did not cause problems; whether it is clinically beneficial remains to be seen. Alternatively, an intervention might target the protein. Molecules have been designed that selectively increase the interaction between the mutant protein and the cell's general protein degradation machinery. Currently these are laboratory experiments and not ready for clinical application, but they show a hopeful direction.

For leukemia and cancers, any therapy would be directed at the neoplastic cells. There is great optimism surrounding the development of immunotherapy for cancer. Previous clinical trials have attempted to make neoplastic cells more immunogenic by expressing a foreign antigen on the cell surface, or to introduce drug-inducible 'suicide genes' into the cells. These were not very successful, but the newer approaches described in *Box 14.10* hold great promise.

Cell-based therapies probably hold greater promise than gene therapy for complex diseases where the pathology is due to loss of some cell population. Huntington disease and Alzheimer disease are high on the list of candidates. Regardless of the initial cause, in both cases it is loss of cells from specific brain regions that causes the clinical problems. The brain is an immunologically privileged site, and allografts (grafts from a different individual of the same species) are not usually rejected and animal studies have demonstrated long-term survival of transplanted brain cells. Bachoud-Lévi and Perrier (2014) review cell therapy trials in Huntington disease.

Immunotherapy of cancer

In theory the immune system should eradicate tumor cells due to the abnormal antigens usually present on the cell surface. Probably many incipient tumors are indeed eliminated in this way, but successful tumor cells have ways of evading or down-regulating immune responses. Some early gene therapy trials in cancer involved attempts to make the tumor cells more immunogenic, for example, by making them express a novel HLA antigen. More recently two approaches to immunotherapy have been showing great promise (Emens *et al.*, 2017).

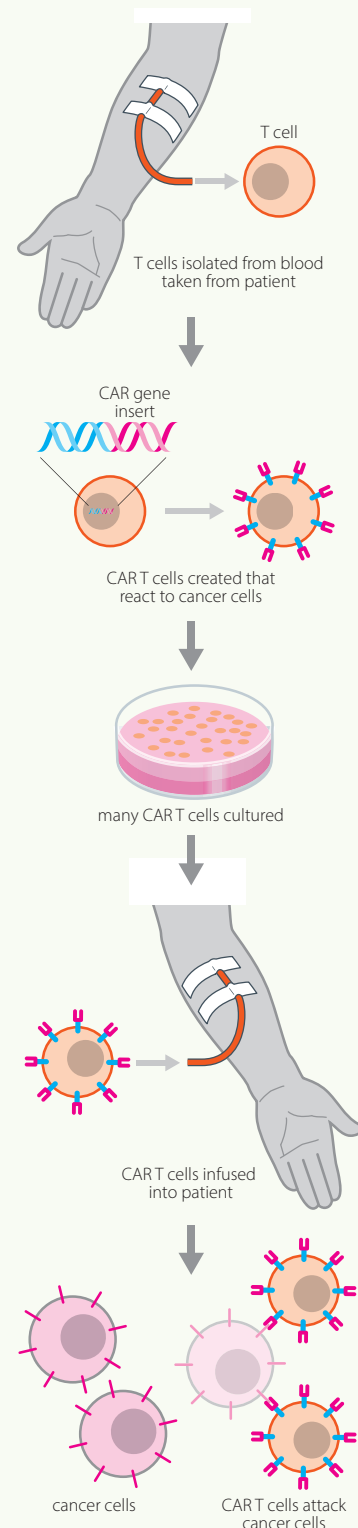
Immune checkpoint blockade involves blocking the PD-1 (programmed cell death protein-1) receptor or its ligand PD-L1 on tumor cells (Constantinidou *et al.*, 2019). When the PD-1 cell-surface receptor associates with PD-L1 this triggers a number of actions that reduce the intensity of immune reactions. This is part of a natural system that controls immune responses and helps prevent autoimmune disease. Many cancer cells over-express PD-L1, thereby protecting themselves against immune attack. Immune checkpoint blockade uses monoclonal antibodies against PD-1 and/or PD-L1 (often in combination with antibodies to another immunoregulatory protein, CTLA4) to overcome this. These are proving to be among the most effective drugs against a variety of cancers, and with fewer side-effects than many alternatives. James P Allison and Tasuku Honjo were awarded the 2018 Nobel Prize for Medicine “for their discovery of cancer therapy by inhibition of negative immune regulation”.

Chimeric antigen receptor (CAR)-T cells are T cells genetically engineered to attack cells expressing a tumor-specific antigen (June and Sadelain, 2018). Elaborate DNA manipulation is used to construct a gene encoding a synthetic T-cell receptor that is specific against a selected antigen on the patient's tumor. Specific features of the intracellular part of the receptor are designed to produce a very powerful immune response. In the laboratory, a retroviral vector is used to transfer the construct into T lymphocytes recovered from the patient's own blood. The engineered T cells are infused into the patient, whose own normal T cells have been depleted to make space (Box figure 14.7).

CAR-T therapy can be highly effective. It is, however, limited by the severity of side-effects; patients have died from the resulting ‘cytokine storm’ in clinical trials. Partly this happens because the selected antigen may be also present on some non-tumor cells of the patient; partly it is inherent to the intensity of the immune attack. A further limitation is the need for elaborate genetic manipulation of the T cells, specific to each individual case, which is both costly and time-consuming. Great efforts are being put into overcoming these limitations. Systems have been developed that allow the action to be switched off, for example, by transfecting the patient's T lymphocytes not only with the gene for the chimeric antigen receptor but also with genes that trigger apoptosis in response to administration of a harmless trigger drug. There are also hopes of by-passing the need for patient-specific T cells, allowing readily available all-purpose cells to be used.

Box figure 14.7 – CAR-T cell therapy.

Adapted from <https://medicalxpress.com/news/2018-02-car-t-cell-therapy-safe-effective.html> with permission from UT Southwestern Medical Center.



Objections to genetic interventions

Every new development in clinical genetics has raised anxieties and objections among some people. Much of this is just the natural anxiety about new interventions, especially in an area that touches on people's notions of identity. As with heart transplants, familiarity brings acceptance. Here we will consider some of the more principled concerns.

- *Helping people with genetic disease to live normal lives allows them to pass on their genes and shows an irresponsible lack of concern for future generations.* We dealt with this in relation to mendelian diseases in *Chapter 9*. In relation to complex diseases, this is really an objection to the whole of medicine – indeed to the whole of civilization. What is a civilized society but a collective attempt to limit the operation of natural selection?
- *Offering prenatal diagnosis for a genetic condition amounts to declaring war on people with that condition.* Understandably, people living with spina bifida, achondroplasia or Down syndrome are highly sensitive to the suggestion that they should never have been born. The desire of parents for healthy children does indeed conflict with the need to respect the lives of people with disabilities. The only general solutions to the conflict (allowing no choice or unfettered choice) are unacceptable to most people in most countries. Each of us probably has our own point at which we would draw the line. Perhaps the wisest course is to try as far as possible to circumvent the conflict by insisting that, whatever indications we may accept for termination of pregnancy, when an affected child is born we must value it as much as we would any other person and provide appropriate support and services to the family.
- *Curing certain conditions is another way of declaring war on people with that condition.* This has particularly surfaced in relation to cochlear implants for deaf children. If these are done on a very young child they can be quite successful in terms of enabling the child to understand speech. Most hearing people would not question the benefit of enabling a deaf child to hear, but some people in the Deaf Community see it otherwise. They contend that attempting to cure deafness is tantamount to trying to undermine their culture.

This is similar to the question whether people in small linguistic isolates should be educated in their own minority language or in the national language of their country. For an individual, learning the national language widens his opportunities, but it threatens the minority culture with extinction in two or three generations. Perhaps the best aim in cochlear implantation is to make the child bilingual, in signing and in spoken language. It is worth pointing out that most children with profound congenital hearing loss are born into hearing families and are not part of the Deaf Community in their formative years. The bilingual solution leaves them free to make their own decisions about where they belong when they are old enough to do so.

- *Advances in genetic testing and reproductive choice put us on a slippery slope.* The 'designer baby' catchphrase, so beloved of journalists, usually surfaces early in this discussion. It has been repeatedly used in connection with 'savior siblings'. These are babies born after *in vitro* fertilization, where embryos were selected to be HLA-compatible with an older child who has a serious disease that could be cured by a transplant of HLA-compatible cord blood stem cells. Allowing this selection, say the critics, opens the door to much more extensive selection of

embryos for non-medical reasons. The weakness of this argument can be seen by looking at *Figure 14.3*. Suppose Mr. and Mrs. Frank N Stein decide they want a tall blue-eyed blonde boy with an IQ of 150. They go through the trauma and expense of IVF and end up with 8 embryos in a dish. Four are female and are discarded. Blue eyes and pale hair are not simple mendelian characters, but they are usually recessive. Photos of the parents do not show us blue-eyed blonde people, so we must assume they are at best heterozygous for these attributes. Thus only 1 in 4 of their embryos will fit their pigimentary aspirations. Out go three, leaving them with just one. What a pity it happens to grow into a man 160 cm tall with an IQ of 95!

The 'designer baby' example just quoted shows that there is very little mileage in extensive embryo selection. However, if one day in the future all the limitations of gene manipulation had been overcome, so that any desired genetic alteration could be achieved safely and with high efficiency, might babies really be designed to order? Might ambitious parents be able to browse a catalog of desirable human characters and place their order? We cannot say this is impossible – but we note that most of the characteristics that ambitious parents might desire in their offspring are multifactorial, with innumerable genes making small contributions, and environment often making a large one. On balance, the 'slippery slope' does not seem very slippery. That is not to deny that all new developments need ethical scrutiny, and should not be implemented unless they pass that scrutiny. But developments in human genetics do seem to have a special propensity to trigger unfounded moral panics.

14.5. The evolving role of the genetics service

The rise of acute genetics

Clinical genetics services have historically focused on the diagnosis and management of rare diseases over the course of a patient's lifetime, with limited exceptions such as diagnosing trisomies in the neonatal period or informing the care of babies born with ambiguous genitalia. For many other childhood disorders and for adult onset genetic disorders such as inherited cancers, neurological disease and cardiac conditions, diagnosis was a lengthy process. Until relatively recently there were few genetic tests that could be offered to patients, and where a test was available it would normally take several months to return a result. As a result there has been little overlap between genetics and so called 'acute' clinical specialties, where management timelines are measured in seconds and minutes rather than months and years. This is changing rapidly as genetic technology continues to advance. Several studies in neonatal intensive care units have demonstrated the value of exome or genome sequencing to achieve a rapid diagnosis to inform clinical management, and have even pointed the way to treatments that have prevented further harm. Similarly, rapid genomic testing is making an impact in cancer treatment, where tumor testing may be available to help guide treatment choices. Significant changes have also been seen in the field of pharmacogenetics where, for example, before prescribing gentamicin in neonatal intensive care units, clinicians are now trialing point-of-care testing of neonates for the m.1555A>G mitochondrial DNA variant to avoid the risk of precipitating profound deafness.

Major areas of change in genetic services in the last 5 years

- Whole exome and whole genome sequencing and panel testing for the majority of groups of genetic disorders (see *Chapter 5*).
- Rapid testing for decisions on cancer treatments and pharmacogenetics.
- Whole exome and whole genome testing in neonatal and pediatric intensive care units.
- 'Mainstreaming', where genetic testing is requested widely by clinicians outside genetic departments.
- Establishment of MDT meetings for interpretation of variants and clinical management decisions.
- Use of international databases for variant interpretation (e.g. gnomAD: <https://gnomad.broadinstitute.org/>) and for identifying patients with similar variants to compare (e.g. Matchmaker Exchange: www.matchmakerexchange.org/).
- Increasing use of plasma tumor DNA tests – 'liquid biopsies' – for diagnosis and to monitor efficacy of cancer treatments (see *Box 7.5*).
- Use of polygenic risk scores to determine risk of, e.g. breast cancer (see *Section 13.4*).
- Prenatal screening using non-invasive techniques (see *Disease box 12*).
- Fetal whole exome sequencing as part of prenatal diagnosis.
- Adoption of therapies based on modification of pathways in which pathogenic variants are active.
- Introduction of gene therapies for a wider range of disorders, such as immunodeficiency (*Figure 14.7*), spinal muscular atrophy (*Box 14.10*), sickle cell disease, etc.

As can be seen from the list above there has been a virtual explosion in recent years in knowledge about the genetic basis of disease and in diagnostic and therapeutic possibilities, and there is no reason to think this will slow down soon. The main challenges are to evaluate the advances from scientific, clinical and ethical perspectives before rushing them into clinical services. There is also an urgent need to educate health professionals in all branches of medicine to ensure genomics is appropriately integrated into routine care. Personalized medicine is becoming a reality and it won't be very long before individual genetic variants are used in prediction, prevention, diagnosis and prescribing, and so public and patient engagement are essential too. Genetic variation is universal and it is important that differences are used to benefit individuals and not used for any form of discrimination. Health systems vary across the globe and it is inevitable some populations will have earlier access to newer treatments but it would be good to hope that eventually these treatments are available more widely since there is no doubt that genetic diseases cause economic, social and educational hardship wherever they occur.

14.6. References

American College of Obstetricians and Gynecologists (2013) The use of chromosomal microarray analysis in prenatal diagnosis. Committee Opinion No. 581. *Obstet. Gynecol.* **122**: 1374–1377.

Bachoud-Lévi AC and Perrier AL (2014) Regenerative medicine in Huntington's disease: Current status on fetal grafts and prospects for the use of pluripotent stem cells. *Revue neurologique*, **170**: 749–762.

- Bundy S and Aslam H** (1993) A five year prospective study of the health of children in different ethnic groups with particular reference to the effect of inbreeding. *Eur. J. Hum. Genet.* **1**: 206–219.
- Constantinidou A, Aliferis C and Trafalis DT** (2019) Targeting Programmed Cell Death-1 (PD-1) and Ligand (PD-L1): A new era in cancer active immunotherapy. *Pharmacol. Ther.* **194**: 84–106.
- Dietz HC** (2011) New therapeutic approaches to mendelian disorders. *New Engl. J. Med.* **363**: 852–863.
- Emens LA, Ascierto PA, Darcy PK, et al.** (2017) Cancer immunotherapy: opportunities and challenges in the rapidly evolving clinical landscape. *Eur. J. Cancer*, **81**: 116–129.
- Fischbeck KH and Wexler NS** (2019) Oligonucleotide treatment for Huntington's disease. *New Engl. J. Med.* **380**: 2373–2374.
- High KA and Roncarolo MG** (2019) Gene therapy. *New Engl. J. Med.* **381**: 455–464.
- Hoyme HE, May PA, Kalberg WO, et al.** (2005) A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: clarification of the 1996 Institute of Medicine criteria. *Pediatrics*, **115**: 39–47.
- June CH and Sadelain M** (2018) Chimeric antigen receptor therapy. *New Engl. J. Med.* **379**: 64–73.
- Kaufmann KB, Büning H, Galy A, et al.** (2013) Gene therapy on the move. *EMBO Mol. Med.* **5**: 1642–1661.
- Lord J, McMullan DJ, Eberhardt RY, et al.** (2019) Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet*, **393**: 747–757.
- Medicines and Healthcare Products Regulatory Agency** (2015) Medicines related to valproate: risk of abnormal pregnancy outcomes. www.gov.uk/drug-safety-update/medicines-related-to-valproate-risk-of-abnormal-pregnancy-outcomes [accessed 26 June 2020].
- Modell B, Ward RH and Fairweather DV** (1980) Effect of introducing antenatal diagnosis on reproductive behaviour of families at risk for thalassaemia major. *Br. Med. J.* **280**: 1347–1350.
- Munnich A** (2006) Advances in genetics: what are the benefits for patients? *J. Med. Genet.* **43**: 555–556.
- Pennisi E** (2013) The CRISPR craze. *Science*, **341**: 833–836.
- Petrovski S, Aggarwal V, Giordano JL, et al.** (2019) Whole-exome sequencing in the evaluation of fetal structural anomalies: a prospective cohort study. *Lancet*, **393**: 758–767.
- RCOG** (2011) Guideline number 17: the investigation and treatment of couples with recurrent first-trimester and second-trimester miscarriage. www.rcog.org.uk/globalassets/documents/guidelines/gtg_17.pdf [accessed 26 June 2020].
- RCOG** (2015) Guideline number 13: chickenpox in pregnancy. www.rcog.org.uk/globalassets/documents/guidelines/gtg13.pdf [accessed 26 June 2020].

- Resta R, Biesecker BB, Bennett RL, et al.** (2006) A new definition of Genetic Counseling: National Society of Genetic Counselors' Task Force Report. *J. Genetic Counseling*, **15**: 77–83.
- Rivers L and Gaspar HB** (2015) Severe combined immunodeficiency: recent developments and guidance on clinical management. *Arch. Dis. Child.* **100**: 667–672.
- Sagoo GS, Butterworth AS, Sanderson S, et al.** (2009) Array-CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet. Med.* **11**: 139–146.
- Venot Q, Blanc T, Rabia SH, et al.**, (2018) Targeted therapy in patients with PIK3CA-related overgrowth syndrome. *Nature*, **558**: 540–546.
- Vermeech JR, Voet T and Devrient K** (2016) Prenatal and pre-implantation genetic diagnosis. *Nature Rev. Genet.* **17**: 643–656.
- Weedon MN, Jackson L, Harrison JW, et al.** (2019) Very rare pathogenic genetic variants detected by SNP-chips are usually false positives: implications for direct-to-consumer genetic testing. *BioRxiv* preprint <https://doi.org/10.1101/696799>. Note that this is a preprint that has not been peer-reviewed; broadly similar conclusions have been reported by Van Hout *et al.*, (2020) *Nature*, **586**: 749–757.

Recommended textbooks

- Clinical Genetics and Genomics, (Oxford Desk Reference)** 2nd edn (2017). Firth HV and Hurst JA. Oxford University Press.
- Emery and Rimoin's Principles and Practice of Medical Genetics**, 6th edn (2013). Rimoin DC, Pyeritz RE and Korf BR. Elsevier Science.
- Gorlin's Syndromes of the Head and Neck**, 5th edn (2010). Hennekam RCM, Krantz ID and Allanson JE (eds). Oxford University Press.
- Harper's Practical Genetic Counselling**, 8th edn (2014). Clarke A. CRC Press.
- Management of Genetic Syndromes**, 3rd edn (2010). Cassidy SB and Allanson JE (eds). Wiley-Blackwell (4th Edition in preparation).
- Smith's Recognizable Patterns of Human Malformation**, 7th edn (2013). Jones KL, Jones MC, del Campo M (eds). Saunders.

14.7. Self-assessment questions

- (1) A woman has a son with Duchenne muscular dystrophy. There are no previous cases in the family, there are no other children and the mother is an only child. She might be a carrier, or the boy might be a new mutation. What is the chance that she is a carrier? (See the *Guidance* for two ways of doing this important calculation.)
- (2) This calculation uses the carrier risk calculated in the previous question. The woman has another child, a girl. What is the risk the girl is a carrier?
- (3) The woman in SAQ2 has two more children, both unaffected boys. Does this alter your estimate of her carrier risk? If so, use a Bayesian calculation to calculate her revised risk.

- (4) A man comes from a large family in which an autosomal dominant condition is segregating. His mother is affected but he is healthy. The condition is 90% penetrant, so he might either not have inherited the disease allele or he might be a non-penetrant case. He marries. Calculate the risk that his first child will be clinically affected by the disease.
- (5) Generalize the result of SAQ4 for a condition with penetrance x , and calculate the maximum risk that such a person would have an affected child, if x could have any value.
- (6) Your mother has Huntington disease. You are healthy at age 45. If half of people who have inherited the disease gene show symptoms by age 45, what is the chance that you have inherited the disease gene?
- (7) Score each of the following statements as true or false:
 - (a) Empiric risks are used in counseling for non-mendelian diseases.
 - (b) Empiric risks are based on mathematical simplifications.
 - (c) Empiric risks embody no assumptions about genetic mechanisms.
 - (d) Empiric risks are valid only for a particular population and time.
- (8) Outline proposals to develop gene therapy for (a) cystic fibrosis, (b) Duchenne muscular dystrophy, and (c) a rapidly growing brain tumor. In each case, consider the advantages and disadvantages of this condition as a target for gene therapy, the gene(s) and constructs you would use, which tissue or cells you would target and how.

[Hints on questions 1–6 are provided in the *Guidance* section at the back of the book.]

Guidance on self-assessment questions

Chapter 1

SAQ 1, 2 and 3. Following the two questions as suggested in the text should lead you to a convincing answer. We are told that each disease is rare, so the chance of an unrelated person who marries into the family being a carrier (if the disease is recessive) is small.

Pedigree 3. Unlike every other pedigree in this book, shows a real family. The disease is hemophilia and the family is Queen Victoria's. Queen Victoria is I-2. The affected people are Leopold (II-1), Frederick (II-8), Leopold and Maurice Mountbatten (III-18, III-19), Rupert (IV-2), Alexis, Tsarevitch of Russia, murdered with his four sisters by the Bolsheviks in 1918 (IV-8), Waldemar and Henry of Prussia (IV-9, IV-11) and Alfonso and Gonzalo of Spain (IV-14, IV-19). V-1 is the present Queen Elizabeth II of Britain. DNA analysis of the remains of IV-8 by Rogaev and colleagues (2009; *Science*, **326**: 817) demonstrated a mutation in the Factor IX gene. Thus the disease was hemophilia B (OMIM 306900), rather than the more frequent hemophilia A.

SAQ 4. This pedigree is ambiguous (you shouldn't be given one like this in an exam). Try each possible mode of inheritance. Don't be misled by the affected female, and don't jump to the conclusion that there is male to male transmission. An affected man has an affected son – but are you sure the son inherited his disease gene from his father? The risk for a child of IV-3 may depend on which mode of inheritance is correct.

Chapter 2

SAQ 1a. Consider events in oogenesis that could result in a missing X (egg fertilized by an X-bearing sperm) and events in spermatogenesis that could result in a sperm with either a missing X or a missing Y. Note that in reality, Turner syndrome is usually the result of anaphase lag rather than non-disjunction (see *Section 2.3*).

SAQ 2a. Use *Figure 2.5* to check whether the breakpoints are proximal (near the centromere) or distal (towards the telomere) on the chromosome arm. 2q22 is about one-quarter of the way down the long arm on chromosome 2 and 4q32 is about three-quarters of the way down the long arm of chromosome 4. Draw out the cross-shaped quadrivalent formed when the two translocated chromosomes pair with their two normal counterparts. Draw it very roughly to scale, using different colors for the chromosome 2 and chromosome 4 sequences. Check

that you are always pairing segments of matching color. Then draw out possible gametes. You could consider 3:1 segregation patterns as well as the 2:2 patterns shown in *Figure 2.17*.

Chapter 3

SAQ 2 CTCAAAGCACGCTCCAGTTCCTCCAGCTG
CAGCUGCAGGAACUGGAGCGUGCUUUUGAG

Remember that strands are anti-parallel but are always written in the 5'→3' direction.

SAQ 3. Transcription starts at the start of exon 1. The first codon is the AUG initiation codon. Intron 1 starts at the end of exon 1.

SAQ 4. You could use a word processor to interleave and align the two sequences, or just compare them manually. Sequences present in the genomic DNA but absent in the mRNA are introns.

Chapter 4

SAQ 1a. First ask yourself, to get a 50 bp product, will the primers flank the underlined 50 nt sequence or will they lie inside it? Write in the complementary strand for the regions where your primers will anneal. Then put in all the 5' and 3' directions and remember that the primers will extend from their 3' ends.

SAQ 4. Imagine you had one copy of each possible sequence n nucleotides long. There are 4^n such sequences, so the total length of all these would be $4^n \times n$. Look for a value of n such that this total $> 3\,000\,000\,000$. Of course, this exercise ignores the fact that the human genome is not a random sequence, and contains many repeated sequences – but it shows that a sequence does not need to be very long to be *potentially* unique within the human genome.

Chapter 5

SAQ 2. Hybridization is a property of single-stranded DNA. Double-stranded DNA doesn't hybridize to anything else.

SAQ 3. Remember that sequences must be read 5'→3'. CTTAAG is not an *EcoRI* site. If you write in the complementary strand, you will notice that GAATTC, like most restriction sites, is palindromic – that is, it reads the same on both strands.

Chapter 6

SAQ 1. Ask whether each change would specifically cause a failure to transport chloride ions across apical cell membranes.

SAQ 2. Two of the changes are deletions of one or a few nucleotides. One change affects a splice site.

Chapter 7

SAQ 3. A laboratory might report the last sequence variant as p.V29M but in fact it might very well affect splicing – it replaces the last nucleotide in exon 1, and a G at this position is part of the normal consensus splice site sequence.

SAQ 4. c.216C>G is a mis-sense mutation, p.172M – it creates an internal methionine codon. This is not an initiator codon: translation initiates at the first suitable AUG in the mRNA, and once initiated, further AUGs are just read as normal methionine codons.

SAQ 1. In regard to Statement 9, although it is generally true that mutations in oncogenes are somatic and not inherited, there are a few exceptions. Inherited mutations in the *RET* oncogene are found in familial thyroid cancer. If a gene product has more than one function it may be simplistic to talk simply of loss or gain of function. Equally, the classification of genes into oncogenes and tumor suppressor genes is a very useful tool for thinking about the molecular pathology of cancer, but it has its limitations.

SAQ 3. If there are n cells and the mutation rate (both for the first and second mutations) is μ , the incidence of sporadic cases is $n\mu^2$ and the penetrance is $n\mu$. Given the stated mutation rate of 2×10^{-5} and penetrance of 90%, $n = 45\,000$ and the expected incidence is 1.8 per 100 000.

SAQ 4. Using the scoring system of Box 7.4 and ignoring details of the pathology, simply adding up points across each pedigree gives 21 points for Family A and 29 for Family B. But all points in Family A come from the paternal side, while in Family B the points are split between the paternal and maternal sides. If the at-risk woman has a *BRCA1/2* mutation, she must have inherited it from either the paternal or the maternal side, not both (patients with 2 mutations in either *BRCA1* or *BRCA2* have been described, but they have a different phenotype, Fanconi anemia, from those with single mutations, see OMIM 605724 and 617883). Neither of these alternatives gives a risk as high as that in Family A. However, one could argue that her total risk should be seen as the sum of those two independent risks. In reality both probands would be offered testing. This is of course a rather contrived example, designed to provide points for discussion.

Chapter 8

SAQ 1. Although every meiosis may be recombinant or non-recombinant for a given pair of loci, only the meioses in the mother of the third generation are informative about this. The χ^2 result (3.6, 1 d.f.) is almost significant at the 5% level, while the lod score is well below the threshold for significance [$L1 = (1/2)^{10}$; $L2 = (1 - \theta)^8 \cdot (\theta)^2$; maximum lod score = 0.83 at $\theta = 0.2$]. The reason is that the lod score, but not the χ^2 test, takes into account the low prior probability of linkage. That is, given two loci picked at random, the chance they would show linkage is of the order of only 1 in 50 – they would probably be on different chromosomes, and even if they were on the same chromosome they might well be sufficiently far apart not to show linkage. Common sense tells us that we need to take the prior probability into account when deciding whether or not to believe something. For example,

you may well believe your friend if he tells you that he missed a lecture because he overslept, but not if he tells you it was because he had been abducted by aliens. See Strachan and Read *Human Molecular Genetics* (Chapter 11 in the 2nd edn, available on the internet, Chapter 17 in the 5th edn). for a brief explanation of how the lod score threshold includes the prior probability, or Ott, *Analysis of Human Genetic Linkage* for full details.

- SAQ 2.** Work out the haplotypes, starting with generation 2. The marker alleles have been chosen in this example to allow unambiguous haplotypes to be assigned to each individual. You can then identify which haplotype in individual II-1 is linked to the disease locus (remember we have been told that the disease is linked to this chromosomal region). When you examine the paternal haplotypes in generation III you can see that individuals III-3 and III-5 have inherited recombinant paternal disease-carrying haplotypes. The positions of crossovers show that the disease locus must map below marker A (from III-3) but above marker C (from III-5). III-6 has a recombinant maternal haplotype. The crossover might be anywhere between markers A and D. This recombinant does not provide any data for mapping the disease locus.
- SAQ 5.** The combination of deafness and diabetes is known sometimes to be caused by mutations in the mitochondrial DNA. NB Although it is sensible to attempt this sort of prioritization, when the causative gene is finally found, often it is not one of the more obvious candidates.

Chapter 9

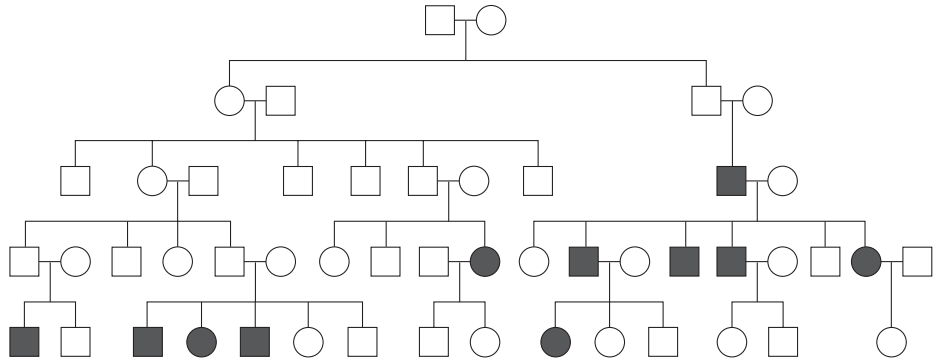
- SAQ 2.** Remember that only people who are homozygous at one or other locus are affected; people who are heterozygous at two or more loci are unaffected. In reality, as explained, it would be unwise to rely on Hardy–Weinberg figures for such a rare recessive condition.
- SAQ 3.** You could try solving this by saying $p^2 + 2pq = 0.64$; substitute $(1-p)$ for q , and you get a quadratic equation to solve for p . It can be done – but it is a lot easier to start by noting that $q^2 = 0.36$
- SAQ 4.** Treat the first part as a 3-allele Hardy–Weinberg problem, as in Box 9.1, with non-functioning, low functioning and normal alleles.
- SAQ 5.** Remember the difference between obligate carriers and possible carriers.
- SAQ 6.** Genetic counseling is non-directive.

Chapter 10

- SAQ 2.** Consider possible blocks in a multi-step pathway, or a block in post-translational processing (see Section 10.4).
- SAQ 3.** This SAQ is there to encourage you to do some reading around the condition.
- SAQ 5.** Eliza is the key – you can write down her two haplotypes directly, which must be two of the four parental haplotypes.

Chapter 11

SAQ 1. Here is one possible pedigree. The condition is autosomal dominant and so it affects both sexes and the mutant gene can be transmitted by both sexes – but it is always non-penetrant when it is inherited from the mother.



SAQ 3. The different sizes of product are from X chromosomes with different numbers of CAG repeat units. You can follow their transmission through the pedigree. Ignore the very small peaks; they are 'stutter bands', artefacts of the PCR process.

SAQ 7. There is an obligate crossover in every male meiosis between the Xp and Yp copies of the pseudoautosomal region. The crossover point might be proximal or distal to the location of the marker. You could try assuming that 50% of the crossovers are located between the centromere and the marker locus. In female meiosis crossovers within this region are quite uncommon.

Chapter 12

SAQ 4. You may be surprised that the odds ratios for the two cases are different:

- for variant A $(750 \times 500)/(500 \times 250)$
- for variant B $(75 \times 950)/(50 \times 925)$

This is a point to bear in mind when interpreting the relative risk conferred by a variant. The more common a variant is, the more extreme the odds ratio (that is, the further away from 1, whether >1 or <1). Odds ratios approach the intuitive relative risk only for rare variants. Most SNPs in genome-wide association studies are common.

SAQ 5. Estimate the relative numbers by counting squares under the appropriate part of each curve. The curves have been drawn having equal total areas, but remember that the curve for NTD should really have 1/100 the area of the normal curve.

SAQ 6. The risk before any screening that a couple are both carriers is $1/40 \times 1/40 = 1/1600$. We need to calculate the sensitivity of a test such that when the partner tests negative, there is only a 1 in 1600 chance that this is a false negative. You can see from *Table 12.1* that this is not readily achievable.

Chapter 13

SAQ 1. The familial eating habits might be learned or genetic. Opposite sex twins are all dizygotic, while like-sex twins may be either monozygotic or dizygotic. The first three observations all suggest genetic factors. The adoption data give the cleanest separation of genetic from environmental effects, though the Barker hypothesis (see *Chapter 11*) would sound a caution against drawing dogmatic conclusions.

SAQ 2. Because the disease is commoner in men than in women, the risk is always higher for a son than a daughter of a given person. The risk is higher for a child of an affected woman than for a child of the same sex of an affected man. Because Betty has two affected male first-degree relatives she is likely to have a higher susceptibility than Anne, so the risk for her child is greater.

SAQ 3. If the recombination fraction is θ , after n generation a proportion $(1-\theta)^n$ remain associated.

SAQ 7. This is the multiple testing problem and a Bonferroni correction is appropriate. Sometimes it is hard to decide how many questions were asked. If a marker has n alleles, and you check each for association, is that one question, or n or $n-1$? The Bonferroni correction is over-rigorous if the questions are not fully independent. If you look not only for associations with individual marker alleles but also with multi-marker haplotypes, are those extra independent questions? If you use linkage analysis and you know that there are susceptibility loci somewhere in the genome, each negative result reduces the area of the genome where the susceptibility factor(s) must be hiding – so are tests of the few remaining areas independent? In short, the multiple testing problem is severe in complex disease studies and requires expert statistical insight.

SAQ 9b. For the $2-1 \times 1-1$ parents the mendelian proportions among the sibs must be modified by the fact that you have selected pairs that were both affected. This is most easily done by a Bayesian calculation, with the mendelian proportions as the prior probability and the fact that both sibs are affected as a conditional:

Genotypes of sibs:	(1-1, 1-1)	(1-1, 2-1)	(2-1, 2-1)
Prior probability	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Conditional likelihood (both affected):	16	8	4
Joint likelihood	4	4	1
Final probability:	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

For an explanation of this method see *Box 14.7*.

Chapter 14

SAQ 1. This is an important basic calculation in genetic risk estimation so we show it in detail. There are two ways of arriving at the answer, which applies to any X-linked recessive condition where affected males never reproduce.

- (a) If the population contains equal numbers of males, who each have one X chromosome, and females who each have two, then one-third of all X

chromosomes are in males. Equally, one-third of all DMD X chromosomes are in males. Any DMD X chromosome that is in a male will not be transmitted to the next generation, thus one-third of all DMD X chromosomes are lost each generation. If the disease frequency remains constant over the generations, this must be balanced by new mutations. Thus one-third of cases are new mutations, so the chance that the mother of an isolated case is a carrier is $2/3$.

- (b) An alternative method starts by calculating the probability that any woman, picked completely at random, is a carrier of DMD. Call this probability P . Suppose she has a daughter. The probability that the daughter is a carrier is made of three parts:
- She might be a carrier because her mother is a carrier and she has inherited her mother's mutated X. The probability of this is $P/2$.
 - She might be a carrier because although her mother was not a carrier, the X she received from her carried a new DMD mutation. Call the mutation probability μ .
 - She might be a carrier because the X she received from her father carries a new DMD mutation. Again the probability of this is μ .

The daughter's overall carrier risk is $P/2 + 2\mu$. But our original woman was selected completely at random; exactly the same logic could have been applied to her and her mother. The carrier risk of the mother and daughter must actually be the same (if it weren't, by repeating the exercise over enough generations, you could get the risk to either 0 or 100%, which would be absurd). Therefore $P = P/2 + 2\mu$, from which we see that $P = 4\mu$.

Now we have worked out the prior probability that a woman is a DMD carrier, we can return to the original question and do a Bayesian calculation of the carrier risk of the woman who has a son with DMD. Either she is a carrier or she is not. The calculation goes as follows:

Woman is:	a carrier	not a carrier
Prior probability:	4μ	$1 - 4\mu \approx 1$
Conditional: affected boy	$1/2$	μ
Joint probability:	2μ	μ
Final probability	$2\mu/3\mu = 2/3$	$\mu/3\mu = 1/3$

SAQ 2. The daughter's risk is half the mother's risk – there is a 1 in 2 chance she will inherit her mother's 'risk' X rather than the other one.

SAQ 3.

- *Method 1:* repeat the calculation of [SAQ1](#), adding in an extra line of conditionals for the two unaffected boys. Given that she is a carrier, the chance of having two unaffected boys is 1 in 4. Given that she is not a carrier, it is $1 - 2\mu$, which is effectively 1.
- *Method 2:* start with her $2/3$ carrier risk as your prior probability. In that case, this has already used the information that she has an affected boy, so the only conditional likelihoods are for the unaffected boys.

This illustrates the fact that it doesn't matter which pieces of information you put as prior probabilities and which as conditionals, just as long as you use each piece of information once and only once.

SAQ 4. First you need to calculate the chance that the man is a non-penetrant carrier. Use a Bayesian calculation. The prior probability is his mendelian 1 in 2 risk, the conditional is the fact that he is clinically unaffected. Having got his risk, the risk his son has inherited the disease gene is half the risk the father carries it. Then remember that even if the child inherits the gene there is only a 90% chance he will be clinically affected.

SAQ 5. The formula is $\frac{1}{2} \times \frac{(1-x)}{(2-x)}$. The answer, 8.6% risk for 59% penetrance, is surprisingly and reassuringly low. If a condition has low penetrance, this increases the risk the unaffected father may carry the disease gene, but at the same time it reduces the risk the child will be clinically affected if it does inherit the gene.

SAQ 6. The 'obvious' answer, 1 in 4, is wrong. Use Bayes to get the correct answer.

Glossary

3' untranslated sequence – in a messenger RNA, the part downstream of the stop codon.

5' untranslated sequence – in a messenger RNA, the part upstream of the translation initiation (AUG) codon.

Acrocentric – of a chromosome, having the centromere close to (but not at) one end. In humans, chromosomes 13, 14, 15, 21 and 22.

Acute transforming retrovirus – a small RNA virus whose genome has accidentally incorporated an activated cellular oncogene.

Affected sib pair (ASP) method – a model-free method of linkage analysis that seeks chromosomal segments which pairs of brothers or sisters who have the same disease share more often than by chance.

Allele frequency (often loosely called **gene frequency**) – the frequency of allele A_n is the proportion of all alleles at the A locus in a certain population that are A_n .

Allele-specific oligonucleotide (ASO) – a short single-stranded oligonucleotide that under suitable conditions hybridizes to only one specific allele of a single nucleotide polymorphism.

Alleles – alternative forms of a gene.

Allelic heterogeneity – the situation where a clinical condition can be caused by any of several (often very many) different mutations within a certain gene. Characteristic of loss of function conditions (cf. **locus heterogeneity**).

Alternative splicing – alternative choices of which segments of an RNA primary transcript are retained in the mature messenger RNA (see *Figure 3.10*).

Amniocentesis – an invasive prenatal test in which 10–20 ml of amniotic fluid is removed with a transabdominal needle for testing fluid biochemistry or fetal cells contained in the fluid (see *Box 14.5*).

Analytical validity – of a test, the extent to which it measures that which it claims to measure.

Anaphase – the phase of cell division (mitosis or meiosis) in which chromosomes or chromatids separate and are pulled to opposite poles of the cell.

Aneuploid – of a cell, not **euploid**; having missing or extra chromosomes.

Anneal – of complementary single strands of nucleic acid, to hybridize forming a base-paired double helix.

Annotation – the process of working out the function and biological meaning of a DNA sequence.

Anticipation – the tendency of a disease to become more severe, more frequent, or to start at an earlier age, in successive generations. A feature of conditions caused by expanding nucleotide repeats, but often an artifact of biased ascertainment.

- Apoptosis** – a specific mechanism by which a cell kills itself.
- ASO** – an **allele-specific oligonucleotide** q.v.
- Association** – the statistical tendency of two things to go together more often or less often than by random chance. The combination occurs at a frequency that is not equal to the product of the individual frequencies.
- Assortative mating** – choosing a mate who is genetically similar to oneself. Can be based either on having a similar phenotype or on being related.
- Autosome** – any chromosome that is not the X or Y sex chromosome.
- Autozygosity mapping** – mapping a recessive condition in inbred kindreds by finding a shared ancestral chromosome segment that is homozygous in all the affected people (see *Figure 8.7*).
- Balanced** – of a chromosomal constitution, having nothing extra or missing. Also used loosely of Robertsonian translocations, even though these lack part of the short arms of the acrocentric chromosomes involved.
- Bayesian method** – a method of assessing the probability of an event by combining the likelihoods of all individual factors that affect the overall probability (see *Box 14.7*).
- Bias of ascertainment** – collecting a sample that is statistically unrepresentative of the larger population.
- Bisulfite sequencing** – a method of identifying unmethylated cytosines in DNA; after treatment with sodium bisulfite they are converted to uracil, which is scored as thymine on sequencing (see *Figure 11.16*).
- Bonferroni correction** – a rigorous statistical correction for multiple testing, consisting of dividing the threshold *P* value for statistical significance by the number of different questions asked.
- cDNA** – complementary DNA; a DNA copy of messenger RNA made using reverse transcriptase. Human cDNAs represent only 1–3% of the genomic DNA but contain most (but not all) of the clinically relevant sequences. Unlike genomic DNA, cDNAs are tissue-specific.
- Cascade screening** – ascertaining gene carriers by systematic testing of the extended family of an affected person (see discussion of **Case 24 (Smit family)** in *Section 12.3*).
- Cell cycle checkpoint** – a regulatory interaction that prevents a cell progressing through the cell cycle unless certain conditions are met (see *Figure 7.7*).
- Centromere** – the point at which the sister chromatids of a replicated chromosome are joined, and the location of the kinetochore, to which spindle fibers attach during cell division. Marked by heterochromatin containing a special histone H3 variant, CENP-A protein.
- Channelopathy** – a disease caused by malfunction of an ion channel.
- Chimera** – of a person, having cells derived from two different zygotes – a rare condition, the opposite of twinning. Of a gene, made by a chromosomal rearrangement that brings together exons of two different genes to form a novel gene. A common occurrence in cancer (see *Table 7.2*).
- Chorionic villi** – outgrowths of fetal origin on the external surface of the chorion, the outermost of the fetal chorion villus biopsy is used to obtain first-trimester fetal material (see *Box 14.5*).
- Chromatid** – a single DNA double helix packaged into a chromosome. Chromosomes normally exist as a single chromatid, but when they are seen during cell division they consist of two sister chromatids joined at the centromere.

- Chromatin** – a general term for the DNA–protein complex of which chromosomes consist.
- Chromatin disease** – a disease caused by faulty regulation of chromatin structure (see *Disease box 11*).
- Chromosomal instability** – accumulation of structural and/or numerical chromosome abnormalities in abnormal (e.g. cancer) cells.
- Chromothripsis** – in cancer, an extensive rearrangement of a single chromosome or part thereof.
- Cis-acting** – of a regulatory element, regulating a gene that is on the same DNA strand. Typical of enhancers; cf **trans-acting**.
- Clinical validity** – of a test, the extent to which it predicts a clinical outcome.
- Coefficient of inbreeding** – the probability that a person is homozygous at any given locus because of the inbreeding of his parents. Equal to half the coefficient of relationship of the parents.
- Coefficient of relationship** – the proportion of their genes that two people share by virtue of having identifiable common ancestors (see *Box 9.3*).
- Common disease–common variant hypothesis** – the hypothesis that most genetic susceptibility factors for common complex diseases are ancient variants that are common in the general population. The hypothesis underlying attempts to identify susceptibility alleles by association studies. The contrary hypothesis is that susceptibility is due to a heterogeneous collection of recent mutations.
- Comparative genomic hybridization – (CGH)** – a technique for detecting sequences anywhere in the genome that are present in an abnormal number of copies (see *Figures 4.7 and 4.12*).
- Complex** – of a disease, having a variety of possible mechanisms in different patients.
- Compound heterozygote** – a person with a recessive condition whose two copies of the relevant gene carry different variants.
- Congenital** – present at birth; not necessarily genetic.
- Consensus sequence** – of a family of related sequences, the sequence having the most common nucleotide at each position (which may or may not be the most common actual whole sequence).
- Conserved sequences** – sequences that are unchanged or little changed in related species.
- Constitutional abnormality** – an abnormality present at conception, and hence in all body cells.
- Copy number variant (CNV)** – a form of DNA variation in which a certain sequence (which may be anything from a few bp to megabases) is present in differing numbers of copies in different individuals. The copies are usually present as tandem repeats, but may be dispersed. Many CNV are non-pathogenic common variants.
- Cousin** – in genetics, used specifically to mean first cousin (q.v.).
- CpG dinucleotide** – a cytosine immediately upstream of a guanine in a DNA sequence. The target of DNA methylating enzymes, and hotspots for CpG→TpG mutations.
- CpG islands** – short chromosomal regions (typically less than 1 kb) where the usual genome-wide depletion of CpG dinucleotides has not taken place. See *Section 11.4* for their significance.
- Cryptic splice site** – a sequence in an exon or an intron that resembles a splice site but not sufficiently to be used as one; a mutation may increase the resemblance so that it does get used as a splice site ('activating a cryptic splice site').

- Denaturing** – separating the two strands of a double helix by heat or high pH; also called melting.
- Denaturing high performance liquid chromatography (dHPLC)** – a method of testing a PCR product or other double-stranded DNA fragment for changes, compared to a reference fragment, by checking the speed with which it passes through a column.
- Diagnostic test** – a test that establishes the diagnosis in a patient affected by a disease (cf. a **predictive test**, a **screening test**).
- Dichotomous character** – a character such as a disease, that some people have and others do not (cf. **quantitative** or **continuous characters**, that everybody has).
- Dideoxynucleotide (ddNTP)** – a chemically modified nucleotide used in DNA sequencing to terminate growing DNA chains (see *Figures 5.3 and 5.4*).
- Digital PCR** – PCR using limiting dilution of the sample to obtain a direct measure of the number of molecules of the test sequence (see *Figure 4.18*).
- Diploid** – of a cell or organism, having two genomes. The normal condition of human somatic cells.
- Dominant** – a character is dominant if it is manifest in a heterozygote. Dominance and recessiveness are properties of characters, not of genes or alleles.
- Dominant negative** – a variant where the product of the variant allele interferes with the function of the normal product in a heterozygote.
- Dosage-sensitive** – of a gene, where different non-zero copy numbers have an effect on the phenotype.
- Dot blotting** – a hybridization test in which either the test DNA or the probe is spotted on to a solid support to immobilize it.
- Double first cousins** – Fred and Joe are double first cousins if *both* Fred's parents are sibs of *both* Joe's parents.
- Downstream** – on a nucleic acid strand, in the 3' direction (of the sense strand in a gene).
- Driver mutation** – a mutation that contributes to the evolution of a tumor and is subject to positive selection during tumor development.
- Dysmorphology** – the study of congenital malformations and syndromes.
- Embryonic stem (ES) cell** – a pluripotent cell derived from the inner cell mass of a blastocyst (see *Figure 14.6*).
- Empiric risks** – risks defined by survey data, as distinct from risks worked out by applying genetic theory.
- ENCODE Project** – an international collaborative project (Encyclopedia of DNA Elements, www.encodeproject.org) that aims to identify all functions of human DNA (see *Section 3.4*).
- Enhancer** – a DNA sequence that regulates expression of a nearby gene by binding activating proteins and looping round to contact the promoter. Enhancers are often active only in specific tissues.
- Epigenetic** – making heritable (from cell to daughter cell, or sometimes from generation to generation) changes in gene expression without changing the nucleotide sequence (see *Section 11.2* for discussion of this definition). Effected by DNA methylation and/or changing chromatin structure.
- Epimutation** – a mutation that causes an epigenetic change but not a DNA sequence change.
- Euchromatin** – chromatin with a relatively open structure in which genes can be active if suitable transcription factors and co-activators are present; the opposite of **heterochromatin**.

- Euploid** – of a cell, containing some number of complete chromosome sets, without any extra or missing chromosomes. The opposite of aneuploid.
- Exome** – the totality of all exons in a genome.
- Exon** – a segment of genomic DNA that corresponds to sequence in a mature mRNA. Exons include the 5' and 3' untranslated regions of a gene as well as the coding sequence.
- Expression array** – a microarray of oligonucleotides or cDNAs that will hybridize to individual mRNAs or cDNAs. When hybridized to bulk cDNA from a cell or tissue the pattern of hybridization reflects the repertoire of mRNAs in the source material.
- Familial** – tending to run in families; not necessarily genetic.
- First cousins** – Jack and Jill are first cousins if one of Jack's parents is the sib of one of Jill's parents.
- Fluorescence *in situ* hybridization (FISH)** – *in situ* hybridization using a fluorescently labeled DNA or RNA probe (see *Figures 4.5 and 4.10*).
- Founder effects** – an unusually high frequency of a particular allele or haplotype in a population that is descended from a small number of founders, one or more of whom happened to carry that sequence.
- Fragile site** – in a chromosome preparation, a region that appears relatively uncoiled and extended. Usually only seen under specific culture conditions, e.g. after treatment with bromodeoxyuridine or aphidicolin. Most fragile sites exist as non-pathogenic polymorphic variants. The FRAXA and FRAXE fragile sites (see *Disease box 4*) are unusual in being pathogenic.
- Frame-shift mutation** – a mutation that alters the reading frame of a coding sequence (see *Box 3.3 and Section 6.2*).
- Functional genomics** – studying the functions of all the genes in a genome or all the genes expressed in a cell or tissue (see *Section 8.4*).
- G-banding** – a standard procedure in which chromosomes are treated so that they stain in a characteristic and reproducible pattern of dark and pale bands, as shown in *Figure 2.5*.
- Gene conversion** – a process in which a short stretch (typically 100 bp) of DNA sequence in a gene is replaced by sequence from the other allele in a heterozygous person. A recombination-like process but non-reciprocal – the donor gene is unchanged (see *Box 10.4*).
- Gene frequency** – the gene frequency (strictly, the allele frequency) of allele A_n is the proportion of all alleles at the A locus in a certain population that are A_n .
- Gene pool** – the totality of alleles at a particular locus in a certain population.
- Gene tracking** – using linked polymorphic markers to follow the segregation of a chromosomal segment through a pedigree. Used to follow a pathogenic mutation when, for any reason, it is not possible to check for the mutation directly by sequencing (see *Box 14.6*).
- Genetic drift** – a change in allele frequencies between generations because of chance differences between the allele frequencies in one generation and the alleles in the gametes that go to make the next generation. Only happens if the number of gametes is small, i.e. the breeding population is very small.
- Genome** – the totality of genetic material of an organism.
- Genomic** – of DNA, the DNA as it occurs in the cell nucleus, in contrast to cDNA.
- Germ-line** – the lineage of cells that are potentially transmissible to the next generation. In humans and other animals the germ-line separates from somatic cells very early in embryogenesis.

- Germinal mosaic** – a person who, owing to a mutation that occurred after they were conceived, has a mutant cell population in their germ-line, so that they can produce recurrent mutant gametes. A major pitfall in pedigree interpretation and risk estimation.
- GWAS** – genome-wide association study; a study in which SNPs spread across the genome are tested for association with a disease in a case-control study.
- Haploid** – of a cell or organism, having only a single genome (i.e. 23 chromosomes in humans).
- Haploinsufficiency** – the condition in which a single functional copy of a gene is not sufficient to produce a normal phenotype, so that loss of function mutations in the gene produce a dominant character.
- Haplotype** – a set of closely linked alleles on a chromosome that is normally inherited as a block.
- HapMap project** – an international collaborative project that aimed to catalog all the conserved ancestral chromosome segments in several different human populations (see *Section 13.2*).
- Hardy–Weinberg distribution** – the mathematical relationship between allele frequencies and genotype frequencies seen when no distorting factors are present. Seldom applies in humans to rare recessive conditions where many cases are due to consanguinity (see *Box 9.1*).
- Heritability** – the extent to which differences between people with respect to a character (in a particular population at a particular time) are due to the genetic differences between them. Heritabilities are correlation coefficients, symbolized as h^2 and taking values between 0 (no genetic influence) and 1 (wholly determined by genetic differences) (see *Section 13.2*).
- Heterochromatin** – chromatin that is highly condensed and genetically inactive. Found mainly at centromeres.
- Heteroduplex** – a DNA double helix containing mismatches (see *Section 5.2*).
- Heteroplasmy** – of a person or cell, having two or more genetically different types of mitochondria.
- Histone code** – the idea that combinations of different covalent modifications of histones in nucleosomes determine the structure and activity of chromatin.
- Homologous chromosomes** – the two No. 1 chromosomes, etc. in a person. Homologous chromosomes contain the same array of loci but, unlike sister chromatids, they are not copies of each other. They may differ in small ways (minor DNA sequence differences) or sometimes in large ways (because of translocations, etc.).
- Homozygous** – having both alleles at a locus the same. The criteria used to assess identity may be more or less stringent, depending on the question being addressed.
- Hybridize** – of complementary single strands of nucleic acid, to anneal forming a base-paired double helix.
- Incidental finding** – a test finding that may be clinically relevant, but is unrelated to the indication for which the test was performed (see *Section 12.4*).
- Induced pluripotent (iPS) cells** – pluripotent cells obtained by reprogramming differentiated somatic cells (see *Figure 14.6*).
- Informative meiosis** – in linkage analysis, a meiosis where the genotypes allow it to be scored as recombinant or non-recombinant.
- Intron** – a segment of a gene that is part of the primary transcript but is excised by the splicing machinery and is not included in the mature mRNA.

Inversion – a structural abnormality in which part of a chromosome is in the wrong orientation compared to the rest (see *Figure 2.19*).

Karyotype – a person's chromosome constitution – also used loosely to describe a display of a person's chromosomes (strictly, a karyogram), as in *Figure 2.8*, etc.

Linkage – the phenomenon whereby loci that are close together on a chromosome tend to segregate together in families. The extent of the tendency (between random segregation and invariable co-segregation) measures the genetic distance in centiMorgans (between 0 and 50 cM) between the loci.

Linkage disequilibrium – a population association of particular alleles at two or more loci. Seen when the loci are closely linked and the alleles are features of a shared ancestral chromosome segment.

Locus – the position of a gene on a chromosome; a type of gene (as distinct from alleles, which are variant forms of the same type of gene).

Locus heterogeneity – the situation where a clinical phenotype can be caused by mutations at any one of several different loci, cf. **allelic heterogeneity**.

Lod score – the statistical measure of the significance of evidence for or against linkage. Equals the logarithm (base 10) of the odds that the loci are linked, with a given recombination fraction, rather than not linked (see www.scionpublishing.com/NCG4 'Resources').

Loss of heterozygosity – in cancer research, the observation that tumor DNA is apparently homozygous for a DNA polymorphism for which the normal DNA of the same patient is heterozygous. Usually the result of loss of a chromosome. If seen repeatedly, implies that the chromosome in question carries a tumor suppressor gene.

Lyonization – X-inactivation (see *Figure 11.4*).

Lysosomal storage disease – an inborn error of metabolism where lysosomes are unable to degrade a certain type of material. As a result the material accumulates in lysosomes, leading to pathological effects.

Manifesting heterozygote – with an X-linked recessive condition, a female carrier who shows some clinical signs of the condition, most likely because by chance she has inactivated her good X in most cells of the affected tissue.

Mendelian – the manner in which genes and traits are passed from parents to their children. The four modes of mendelian inheritance are: autosomal dominant, autosomal recessive, X-linked dominant, and X-linked recessive. The term "mendelian" refers to Gregor Mendel (1822–84) who formulated the laws forming the foundation of classical genetics.

Meta-analysis – analysis of the combined data from a number of separate studies.

Metacentric – a chromosome that has its centromere in the middle (e.g. numbers 3 and 20 in humans).

Metaphase – the stage of mitosis or meiosis immediately before anaphase, when chromosomes are maximally contracted and aligned on the equatorial plane (metaphase plate) of the cell.

Methylation – attaching methyl (CH_3) groups to any molecule, but particularly used for converting cytosine in a CpG dinucleotide to 5-methyl cytosine as part of gene regulation (see *Section 11.4*).

Methylation-sensitive restriction enzyme – a restriction enzyme such as *HpaII* that will only cut unmethylated recognition sites (see *Section 11.4*).

Microarray – a solid support divided into a large number of cells in each of which a specific test sample or reagent has been anchored, allowing a large number of tests to be carried out in parallel. Microarrays of oligonucleotides, cDNAs, antibodies or tumor samples are widely used in genetic research.

Microdeletion – a chromosomal deletion that is too small (< 3–5 Mb) to be seen on a standard chromosome preparation; Detected by **fluorescence *in situ* hybridization**, **comparative genomic hybridization**, or **multiplex ligation-dependent probe amplification**.

MicroRNA (miRNA) – single-stranded RNA molecules 18–22 nucleotides long that regulate translation of mRNAs.

Microsatellites – short tandem DNA repeats where the repeat unit is 1–6 nucleotides (tandem repeats with longer repeat units are called mini-satellites). Polymorphic microsatellites are one of the main types of DNA marker for linkage analysis (see *Box 8.1*).

Microsatellite instability – a property of tumors that are deficient in repair of DNA mismatches caused by replication errors. Compared to the normal DNA of the patient, tumor DNA contains new alleles of many microsatellites from all across the genome.

Mismatch repair – a protein complex, including the MSH2 and MLH1 proteins, checks newly replicated DNA for wrongly incorporated nucleotides, cuts them out and re-synthesizes that stretch of the DNA.

Modifier gene – a gene that modifies the phenotype of a mendelian condition whose primary cause is a different gene.

Monosomy – having only one copy of one particular chromosome, but two of all the others (i.e. 45 in total for an autosomal monosomy).

Morpholino – nucleic acid analogs used to knock down expression of a gene.

Mosaic – having two or more genetically different cell lines. A person can be mosaic for a chromosomal variant or a single-gene change.

Multifactorial – a catch-all term to describe a character that is determined by many factors including several genes and environmental factors.

Multiplex ligation-dependent probe amplification (MLPA) – a method for simultaneously checking a large number (30–50) of short DNA sequences for copy number variations. Used especially for checking multi-exon genes for deletions or duplications of whole exons (see *Figures 4.6 and 4.11*).

Mutation – (1) the occurrence of a change in gene sequence, and (2) a sequence variant, the product of a mutation (see *Box 6.1*).

Next generation sequencing – a collective name for different technologies that conduct millions of sequencing reactions in parallel, thereby generating vastly more sequence data per run than Sanger sequencing (see *Figures 5.5–5.7*).

Non-allelic homologous recombination (NAHR) – recombination between mis-paired repeats, leading to a chromosomal deletion, duplication or inversion (see *Disease box 2*).

Noninvasive prenatal testing (NIPT) – genetic testing by next-generation sequencing of free fetal DNA present in the maternal circulation (see *Disease box 12*).

Non-parametric linkage analysis – linkage analysis that is based on the extent to which affected relatives share chromosomal segments, but does not depend on a specific genetic model for the cause of the phenotype (see *Section 13.2*).

Nonsense-mediated decay – a mechanism in cells that breaks down most mRNA molecules which contain a stop codon located more than 50 nucleotides upstream of the 3′-most splice site. Probably evolved to protect cells against dominant negative effects of truncated proteins (see *Figure 6.4*).

Nonsense mutation – a mutation that converts a codon for an amino acid into a stop codon (UAA, UAG or UGA in mRNA; TAA, TAG or TGA in the DNA).

Northern blotting – analyzing RNA by gel electrophoresis, transferring to a membrane and hybridizing to a labeled probe.

Nucleosome – the basic unit of chromatin, comprising 146 bp of DNA wrapped round a core consisting of two molecules each of histones H2A, H2B, H3 and H4 (see *Figure 2.18a*).

Nucleoside – a combination of a base and a sugar.

Nucleotide – a combination of a base, a sugar and a phosphate.

Odds ratio – in a case-control study, the odds ratio for a variant is the ratio of the odds, for people who do or do not have the variant, of being a case rather than a control (see *Box 13.2*).

Oligonucleotide ('oligo') – a short piece of single-stranded DNA or RNA.

Obligate carrier – a person whose pedigree shows that they must be a carrier of a recessive (autosomal or X-linked) condition. For X-linked conditions where new mutations are frequent, an obligate carrier must have affected or carrier relatives both in her own or a previous generation, and among her children or grandchildren. Having more than one affected child does not make a woman an obligate carrier because she might be a germinal mosaic.

Okazaki fragments – intermediates in DNA replication. As a replication fork moves along a double helix, one new strand growing in the 5′→3′ direction can grow continuously in the same direction as the movement of the fork, but the other is synthesized as a set of short (100–200 nucleotide) fragments that are later ligated together (see *Box 7.2*).

Oncogene – a gene that suffers gain of function mutations in cancer. Strictly applies only to the mutated version; the normal version is strictly called a **proto-oncogene**, but this distinction is often ignored.

One gene – one enzyme hypothesis – the hypothesis (defined by Beadle and Tatum in the 1940s) that the function of each gene is to direct synthesis of one specific enzyme. Now seen as only part of the story.

Passenger mutation – in tumorigenesis, a mutation that is the random result of the genetic instability of cancer cells, and which does not contribute to progression of the tumor.

Penetrance – the probability of a character being manifest, given a certain genotype. Penetrance is a property of a character or phenotype, not a gene or allele.

Pharmacodynamics – the study of the way a drug target responds to a drug.

Pharmacogenetics – the study of single gene effects on the metabolism or action of a drug.

Pharmacogenomics – the genome-wide study of drug targets or drug effects.

Pharmacokinetics – the study of genetic factors that influence the uptake, distribution, metabolism or elimination of a drug.

Phenocopy – a phenotype that resembles a genetic phenotype, but is produced by non-genetic means.

Phenotype – the observable characteristic of a person (including the result of tests).

Pleiotropic – of a mutation, having effects on many systems.

Polygene – an unfortunate term sometimes used to describe the genes responsible for a polygenic character. Polygenes are not a different type of gene, they are ordinary genes that have variants which have a minor effect on the character in question. The same genes may have major effects on a different character.

Polygenic – in mathematical theory a polygenic character is determined by the combined action of an infinite number of genes, each of which has an infinitesimally small effect. In reality, polygenic effects can be due to the combined effects of just a handful of genes.

Polygenic risk score – a measure of an individual's relative risk, based on a genome-wide analysis of all the variants the person carries. Polygenic risk scores are based on genome-wide association studies, and are specific to the particular population in which the initial GWAS was performed.

Polymorphism – a term used with varying meanings: (1) a variant present in a population at a frequency too high to be maintained by recurrent mutation alone (the correct usage in population genetics); (2) a variant that is relatively common in a population (used in mutation screening); (3) a non-pathogenic variant (a loose usage in the context of molecular pathology).

Population attributable risk (population attributable fraction) – of a cause of disease, the extent to which that particular cause is responsible for the overall incidence of the disease in the population (see *Box 12.3*).

Population stratification – the existence within one population of different subgroups that do not freely interbreed.

Positional candidate – in identifying disease genes, a gene located in a chromosomal region identified by linkage as containing a disease gene.

Positional cloning – identifying a disease gene through linkage analysis followed by testing positional candidates for mutations; as compared to identifying it through untargeted sequencing or investigation of the pathogenesis (see *Figure 8.3*).

Positive predictive value – of a test result, the proportion of cases positive on the test that actually have the condition being sought (see *Box 12.1*).

Predictive test – a test that shows whether a currently healthy person is likely subsequently to develop a late-onset disease.

Pre-mutation – in diseases caused by expanded nucleotide repeats, an expansion that is not long enough to cause the disease, but is long enough to destabilize the repeat, so that later generations are affected (see *Disease box 4*).

Primary transcript – the initial RNA product of transcribing a gene. Contains all the exons and introns of the gene. The introns are cut out when the primary transcript is processed to form the mature mRNA.

Primer – in DNA synthesis, a short (10–40 nt) oligonucleotide that hybridizes to a single strand of DNA and that is then extended by DNA polymerase adding nucleotides to its 3' end.

Prior probability – in Bayesian risk estimation, the initial estimate of how plausible each alternative hypothesis is (see *Box 14.7*).

Probe – a piece of single-stranded nucleic acid labeled, for example, with ^{32}P or a fluorescent dye, that is used in a hybridization assay to test for the presence of a complementary sequence.

- Prodrug** – a pharmacologically inactive substance that is converted within the body into an active drug.
- Promoter** – the DNA region immediately upstream of a gene on which the RNA polymerase complex is assembled to enable transcription of the gene.
- Prophase** – the first stage of mitosis or meiosis when the chromosomes are gradually condensing and becoming visible. Ends with the dissolution of the nuclear membrane.
- Prometaphase** – late prophase of cell division. Cytogeneticists normally karyotype mitotic cells in prometaphase because the chromosomes are more extended than at metaphase and show the banding better.
- Proteome** – the complete set of proteins in a cell or tissue.
- Proto-oncogene** – the normal, unactivated version of an **oncogene**.
- Pseudoautosomal region** – the regions at the tips of the X and Y chromosome short arms that contain 2.6 Mb of homologous DNA and recombine in meiosis. Genes in this region show an autosomal pattern of inheritance. There is another short pseudoautosomal region at the tips of the long arms.
- Pseudogene** – a non-functional copy of a gene at a separate locus from the normal functional copy (as distinct from a non-functional allele at the normal locus). Pseudogenes are very common in the human genome.
- QF-PCR** – quantitative fluorescence-based PCR; used for rapidly checking a DNA sample for numerical chromosome abnormalities (see *Figure 4.19*).
- Quantitative trait locus (QTL)** – a locus that contributes to the phenotype of a quantitative character.
- Quantitative character** – a character like blood pressure that everybody has, but with varying magnitude. Sometimes called a continuous character, cf. **dichotomous character**.
- Random mating** – a choice of mate unrelated to genotype; the opposite of **assortative mating**.
- Read depth** – in next-generation sequencing, the number of times a given sequence is read in one sequencing run (see *Figure 5.6*).
- Read length** – in next-generation sequencing, the average length of individual sequence reads.
- Real-time PCR** – various methods by which the accumulation of a PCR product can be followed as the reaction proceeds. The basis of most quantitative PCR assays.
- Recessive** – a character is recessive if it is not manifest in a heterozygote. Recessiveness and dominance are properties of characters, not of genes or alleles.
- Recombinant** – a gamete produced by a person is recombinant for two loci if the two alleles that it carries came from different parents of the person (see *Figure 8.4*).
- Recombinant DNA** – DNA produced by ligating together sequences derived from different sources – typically a human sequence of interest ligated into a vector.
- Recombination** – the exchange of chromosome segments during meiosis.
- Relative risk** – the risk of a disease for a person with a specific genotype, or a specific relationship to an affected person, compared to the risk in the general population. Note that relative risks are quite distinct from absolute risks. A relative risk of 10 may be of no clinical significance if it only raises the absolute risk from 1 in 10000 to 1 in 1000.
- Restriction endonuclease** – an enzyme that cuts double-stranded DNA at a specific sequence, usually a 4- or 6-nucleotide palindrome (see *Box 4.2*).

- Restriction fragment length polymorphism (RFLP)** – a DNA polymorphism due to a nucleotide change that creates or abolishes the recognition site for a restriction enzyme (see for example, *Figures 5.8 and 5.10*).
- Robertsonian translocation** – a special type of translocation in which two acrocentric chromosomes are joined close to their centromeres (see *Figure 2.19*).
- ROC curve** – receiver-operating characteristic curve; a graph of sensitivity vs. (1–specificity) of a test; the area under the curve is a measure of the discriminatory power of a test (see *Box 13.3*).
- RT–PCR** – reverse transcriptase – polymerase chain reaction; a technique for making many DNA copies of an RNA. A common method for studying messenger RNA. Not to be confused with **real-time PCR**.
- Sanger sequencing** – DNA sequencing by the technique illustrated in *Figure 5.4*; the traditional method, as opposed to **next-generation sequencing**.
- Screening test** – a test used to select people at high risk from a population. Normally followed by a **diagnostic test** (see *Section 12.2*).
- Second cousins** – two people are second cousins if their parents are first cousins.
- Sense strand** – in a gene the strand of the double helix whose sequence corresponds to the sequence of the messenger RNA (the opposite of the template strand).
- Sensitivity of a test** – in a test for a disease, etc., the proportion of affected individuals that the test picks up (see *Box 12.1*).
- Sex-limited** – a character that is seen in only one sex for anatomical or physiological reasons.
- Short interfering (si)RNA** – short double-stranded RNA molecules that inhibit expression of a matching gene.
- Sibs (siblings)** – brothers or sisters.
- Signal peptide** – the N-terminal dozen or so amino acid residues of a nascent protein that determine where it will be transported. Signal peptides are cleaved off once they have performed their function.
- Single nucleotide polymorphism (SNP)** – any polymorphic variation at a single nucleotide (see *Box 8.1*). Alternatively called single nucleotide variants (SNVs).
- Single strand conformation polymorphism (SSCP)** – a quick but fallible method for scanning a DNA fragment (up to 300 nt) for variants.
- Sister chromatids** – the two chromatids of a chromosome as seen in a dividing cell. Sister chromatids are copies of each other, made during the preceding round of DNA replication.
- Slipped-strand mis-pairing** – a mistake in replication of a tandemly repeated sequence that results in the newly synthesized strand having extra or missing repeat units compared to the template.
- Small RNA molecules** – RNA molecules less than 200 nt long. Small RNAs have many different functions in cells, particularly in gene regulation.
- SNP chip** – a microarray of allele-specific oligonucleotides, used for genotyping many SNPs in a single operation. Can also be used to check for copy number variants.
- Somatic mutation** – a mutation in a cell of the body that will not be transmitted to offspring.
- Southern blotting** – analyzing DNA by restriction digestion, gel electrophoresis, transfer to a membrane and hybridizing to a labeled probe (see *Figure 4.4*).
- Splice isoforms** – alternative forms of a protein produced by alternative splicing of exons.

Spliceosome – the large protein–RNA machine that splices introns out of **primary transcripts** (see *Disease box 10*).

Stem cell – a cell that is capable both of self-replication and of giving rise to a variety of differentiated cell lineages (see *Figure 14.5*).

Stop codon – in a messenger RNA, a UAG, UGA or UAA codon that signals the ribosome to dissociate and cease extending the polypeptide chain; the corresponding sequence in a gene.

Stratified medicine – treatment of a patient guided by his/her genotype.

Submetacentric – of a chromosome, having a long arm and a short arm, e.g. most human chromosomes (the others are metacentric or acrocentric).

Tandem repeat – direct DNA sequence repeats that are adjacent to each other. Other types of repeats are inverted repeats (or palindromes) and dispersed repeats.

Telomerase – a ribonucleoprotein that can add TTAGGG units to the telomeres of chromosomes.

Telomere – the special structure that stabilizes the ends of chromosomes, comprising specific proteins complexed to tandemly repeated TTAGGG DNA sequences (see *Box 7.2*).

Template strand – in a gene, the strand of the double helix that base-pairs with the nascent RNA during transcription.

Teratogen – any agent that interferes with normal embryonic or fetal development.

Trans-acting – of a genetic regulatory element, regulating a gene or genes that lie elsewhere in the genome (normally by making a diffusible regulatory products); cf **cis-acting**.

Transcription factor – a protein whose action is to facilitate transcription of a gene or genes by helping bring the RNA polymerase to the promoter.

Transfer RNA – small RNA molecules that transport amino acids to ribosomes that are synthesizing proteins.

Translocation – a structural abnormality in which two chromosomes swap non-homologous segments (see, for example, *Figure 2.16*).

Transposon – a ‘jumping gene’: a mobile genetic element that can move from one chromosomal location to another, either via excision or by making a mobile copy. They can be seen as a sort of intracellular virus. Transposons can be recognized by certain sequence features. About 50% of the human genome consists of transposons, but the great majority have accumulated mutations that destroy their ability to transpose.

Triploid – a cell or organism having three copies of the genome (in humans, 69 chromosomes). Normally lethal in animals including humans.

Trisomy – having three copies of one particular chromosome, but two of all the others, i.e. 47 in total. (see for example, *Figure 2.10*).

Trisomy rescue – the main mechanism producing uniparental disomy. A chance mitotic non-disjunction in a trisomic early embryo produces one cell with the correct chromosomal number, from which the whole baby develops (see *Figure 11.11*).

Tumor suppressor gene – a gene that suffers loss of function mutations in cancers.

Twin concordance – the likelihood that when one of a pair of twins has a certain condition, the co-twin will also have it.

Unbalanced – of a chromosomal abnormality, having extra or missing material, rather than just the correct material rearranged.

Uniparental disomy – having both copies of one chromosome pair inherited from the same parent. In isodisomy the two are copies of the same parental chromosome, in heterodisomy both chromosomes of that parent are present.

Unrelated – everybody is related if you go back far enough. In genetics, people are usually described as unrelated if they do not share any common great-grandparent.

Upstream – on a nucleic acid strand, in the 5' direction (of the sense strand in a gene).

Variable expression – of a DNA variant, producing different phenotypes in different people.

Vector – a DNA sequence into which a piece of DNA of interest can be ligated, allowing it to be introduced into cells and manipulated. Most vectors are engineered versions of natural viruses (see *Box 14.8*).

X-inactivation – the mechanism whereby in each cell of a person, genes on all but one of the X chromosomes are switched off, so that only one gene copy is active regardless of the number of X chromosomes present (see *Figure 11.4*).

Index

b after a page number indicates that the entry appears in a box, **f** in a figure, **ff** on this and following pages, **g** in the Glossary, **t** in a table. For syndromes and diseases see the *Disease index* for a full listing.

- 100 000 Genomes Project, 136, 137t
- 12p tetrasomy, mosaic in Pallister–Killian syndrome, 170
- 21-hydroxylase, gene conversion, 260
- 23andMe company, 329
- 3' untranslated sequence (3'UT), 63f, 64, 415g
- 45,X chromosome constitution, *see* Turner syndrome
- 5' untranslated sequence (5'UT), 63f, 64, 415g

- Abnormalities, constitutional, 52
- Acadians of Louisiana, 237t
- ACCE framework, 320–1
- Acetylators, fast and slow, 258
- Achondroplasia
 - case 14 (Jenkins family), 143
 - genotype–phenotype correlation, 168t
 - high frequency of new mutations, 160, 242
 - mutation–selection balance, 160
 - paternal age effect, 160
 - pregnancy risks, 145
- Acrocentric chromosomes, 30b, 415g
- Acute transforming retroviruses, 181, 182t, 415g
- Adapters (for PCR), 95
- Adenine, chemical formula, 71b
- Adenoviruses, as vectors for gene therapy, 389, 389f
- Adoption studies, for estimating heritability, 336–7, 337f
- Adverse drug reactions, 256
 - Types A and B, 256
 - with gentamicin, 393
- Affected sib pair method, 415g
- Aging, theories, 168
- Alanine (Ala, A), chemical formula, 72b
- Alcohol
 - reproductive risks, 370
 - risk to fetus, 370
- Aligning reads (NGS), 126f
- Allele frequencies, 232, 234–5, 415g
 - factors changing them, 234–5
- Allele-specific oligonucleotides (ASOs), 87, 93, 119, 415g
- Allele-specific PCR (ARMS test), 119, 119f, 133f
- Allelic association, *see* Linkage disequilibrium
- Allelic heterogeneity, 162f, 259, 266, 308, 314t, 415g
- Allelic homogeneity, 118, 160, 162f
- Alpha-fetoprotein (AFP), 307f, 311
- Alpha-thalassemia, 33, 33b
- Alternative promoters, 73f
- Alternative splicing, 73f, 150, 415g
- Alzheimer disease
 - ApoE4 susceptibility factor, 333, 346, 348
 - brain pathology, 345–6, 345f
 - case 25 (Yamamoto family), 333
 - mendelian subset, 345, 345t, 346
 - molecular pathology, 347f
 - risk, 335t
- American College of Medical Genetics and Genomics, proposed mandatory checks, 325, 326t
- Amino acids, formulae and abbreviated names, 72b
- Amish, Old order, 238
- Amniocentesis, 28b, 378, 379, 380f, 415g
- Amniotic bands, 374f
- Amniotic fluid, 28b
- Amyotrophic lateral sclerosis, details of
 - expanded repeats, 112t
- Analytical validity (of a test), 320, 327, 328, 415g
- Anaphase, 35, 36f, 415g
- Anaphase lag, 42
- Ancestors, 2^N after *N* generations, 340f
- Ancestral chromosome segments, 339
- Androgen insensitivity syndrome, 368
- Aneuploidy, *see* Chromosome abnormalities, trisomy
- Angelman syndrome, 33b, 372
 - case 22 (Qian family), 277
 - causes, 287, 287t
- Annotation, of Reference Sequence, 74, 415g
- Antenatal screening, *see* Screening, prenatal
- Antibodies, *see* Immunoglobulin
- Anticipation, 415g
 - in myotonic dystrophy, 113f
 - in pedigrees, 113
- Anticonvulsants, reproductive risks, 370
- APC protein, *see* Familial adenomatous polyposis
- ApoE protein, E2, E3, E4 variants, 346, 346t
- Apoptosis, 178, 187, 188, 416g
- Arginine (Arg, R), chemical formula, 72b
- Array-CGH, *see* Comparative genomic hybridization
- Ascorbic acid
 - inability of humans to make, 268b
 - in collagen biosynthesis, 65
- Ashkenazi Jews, founder effects among, 236
- Asparagine (Asn, N), chemical formula, 72b
- Aspartic acid (Asp, D), chemical formula, 72b
- Association versus linkage, 338b, 416g
- Assortative mating, 233, 235, 416g
- Audiogram, 208f
- Autism, 358–60b
 - diagnostic criteria, 358–9
 - epidemiology, 359
 - genetic investigations, 101, 359
 - overlap of susceptibility factors with other conditions, 360f
- Autosomal dominant pedigree pattern, 15f, 16b
- Autosomal monosomy, 32b
- Autosomal recessive pedigree pattern, 16b
- Autozygosity mapping, 213ff, 218f, 416g
- Azathioprine
 - metabolism, 261, 261f
 - risk of adverse reaction, 257t

- B lymphocytes, 252b, 259, 265, 286
- Back mutation, 234
- Balanced abnormalities
 - detection methods, 93, 100, 110t, 416g
 - reasons they may be pathogenic, 52
- Balanced versus unbalanced abnormalities, 50
- Banding (of chromosomes), 27, 27f, 28
- Barker hypothesis (metabolic programming), 298
- Barr body, 280
- Base pairs, 59
- Bayesian calculation of risk, 383, 384b, 416g
- Beadle and Tatum, 'one gene – one enzyme' hypothesis, 255
- Bedouin, 237t
- Beta-amyloid protein, role in Alzheimer disease, 345f, 345–6
- Beta-catenin, 190, 195, 196f
- Beta-globin gene, 119, 130f, 150
 - activating a cryptic splice site, 159
- Beta-thalassemia
 - allelic heterogeneity among Jews in Kurdistan, 238
 - case 13 (Nicolaidis family), 117
 - due to p.Glu26Lys variant, 150
 - genotype–phenotype correlation, 168t
 - management, 129
 - mutation testing, 129, 130f
 - spectrum of mutations in Greek Cypriots, 118, 129t
- Bias of ascertainment, 113, 416g
- Biobanks, privacy concerns, 245
- Biochemical genetics, 255
- Biological age, predicted from CpG methylation, maybe, 293
- Biotin, biotin–streptavidin techniques, 87
- Bisulfite sequencing, 294, 294f, 416g
- Bivalents (chromosomes), 37f
- Blaschko's lines, 169, 169f, 282
- Bone marrow transplantation, 263, 264, 265, 393
- Bonferroni correction, 339, 416g
- Bottleneck, 236
- BRCA1/2, exon structures, 191, 192f
- Breast cancer
 - BRCA1/2 founder mutations, 191
 - case 16 (Wilson family), 176
 - classification of tumors, 200

- Breast cancer – *continued*
 management options for *BRCA1/2* carriers, 194
 mutation testing example, 194
 pedigree scoring system, 192b
 susceptibility factors, 191, 193t
 targeted drugs, 271t
 Bromodomain, reader of acetylated histones, 296
 Brugada syndrome, 326t
 Bulgarian gypsies, 237t
- Caenorhabditis elegans* (nematode worm), 279, 279f
 gene count, 76
 lack of DNA methylation, 292
- Cancer
 breast, *see* Breast cancer
 chromosomal rearrangements in, 183, 184t
 classification of tumors, 200
 colon, *see* Colon cancer
 deletions in, 179b, 185
 driver versus passenger mutations, 178
 familial, 186t; *see also individual cancers*
 genomic instability, 178, 179b
 immunotherapy, 399b
 maybe just 12 signaling pathways involved, 200, 201f
 multi-omics studies, 197b
 multistage development, 178, 189, 189f
 mutation signatures, 180
 oncogenes versus tumor suppressor genes, 178
 result of natural selection, 177
 retinoblastoma, *see* Retinoblastoma
 stem cells, 178
 targeted drugs, 270–1
 the six capabilities, 200
- Cancer Genome Anatomy Project, 197
 Carbamazepine, risk of adverse reaction, 257t
 Carbimazole, reproductive risks, 370
 Carrier frequencies, using Hardy–Weinberg to calculate, 234
 Carriers (of X-linked condition), obligate, 11f
 CAR-T cells, for cancer therapy, 270, 399
 Cascade testing, 306, 309, 310, 313, 318–19, 376, 416g
 for familial hypercholesterolemia, 306
 Caspases, 188
 Causative variants, difficulty of identifying from GWAS data, 344
 cDNA, 111, 416g
 Cell cycle, checkpoints, 186, 186f, 416g
 Cell division, 34ff
 Cell-free DNA (in circulation), 195b, 329–30, 379
 Central Dogma, 58
 Centromere, 28, 34, 416g
CFTR gene, 62t, 67, 67f, 121
 5T / 7T / 9T variants, 149
 activation of cryptic splice site, 150
 p.F508del variant, 154
 view in ENSEMBL, 75b
 CGH, *see* Comparative genomic hybridization
 Chaperone molecules, 64
 Chiasma, *see* Crossovers
 Chimera, 416g
 Chimeric genes, 52, 147, 162, 183, 416g
 Cholesterol, normal ranges in serum, 318
- Chorion villus biopsy, 28b, 100, 378, 379, 379f, 416g
 problem interpreting mosaicism, 171
 Chromatids, 30b, 416g
 Chromatin, 30b, 47, 279, 417g
 conformation, 292, 292f, 296ff, 297b
 domains, types A and B, role in regulation of transcription, 297
 remodeling complexes, 155, 292, 297
 Chromodomain, reader of methylated histones, 296
 Chromosome abnormalities, 31ff
 deletions, 49, 49f
 detection methods, 110t
 inversions, 49, 49f, 51f
 mosaic, 375
 nomenclature, 30b
 numerical, 48
 recurrence risk, 45, 46
 structural, 31, 48ff, 49f
 trisomy, 30b, 32b, 39, 40f, 41f, 48
 Chromosome bands, nomenclature, 29f, 30b
 Chromosome breakage syndromes, 375
 Chromosome conformation capture (3C), 297
 Chromosome painting, 110
 Chromosome rearrangement
 activating oncogenes, 183, 184t, 271t
BCR–ABL1, 202, 202f, 203f
EML4–ALK gene fusion, 271t
TEL–AML1, 190, 191f
 Chromothripsis, 178, 417g
 Chronic myeloid leukemia, *see* Leukemia, chronic myeloid
- Ciliopathies, gene panel, 137t
 Circos plot, 199f
 Circulating tumor cells, 195b
 Clinical geneticist
 a bad measure of output, 101
 roles of, 365, 401
 Clinical utility (of a test), 321, 327, 328
 Clinical validity (of a test), 320–1, 327, 328, 417g
 ClinVar database, 221
 Clonal deletion, in maturing immune system, 259
 Cloning, in *E. coli*, etc., 93
 Cocaine, reproductive risks, 370
 Cochlear implants, 219
 Codeine, metabolized by *CYP2D6*, 256
 Codon, initiation, 151t
 premature termination, 152, 152f, 153f, 159f
 stop, 151t
 Codons, table of, 151t
 Coefficient of inbreeding
 definition, 240, 417g
 using Sewall Wright's path coefficient method to calculate, 241b
 Coefficient of relationship, definition, 240, 417g
COL2A1 gene, 70, 121, 134, 158
 molecular pathology of variants, 158, 167
 Colcemid, 27f
 Collagen
 biosynthesis, 65b
 type I, 65
 type II, 70, 158
 type XI, 70
 Colon cancer
 classification of tumors, 201
 familial adenomatous polyposis, *see* Familial adenomatous polyposis
- Lynch syndrome (hereditary non-polyposis), 186t
MYH-associated polyposis, 326t
 targeted drugs, 271t
 Companion diagnostic, 258, 270, 271
 Comparative genomic hybridization, 87, 91, 92f, 100f, 108t, 375, 417g
 abnormalities detectable, 110t
 Compound heterozygotes, 65, 215, 417g
 Confined placental mosaicism, 171, 330
 Congenital, definition, 20, 417g
 Congenital adrenal hyperplasia, 322t, 368
 Connexin 26, 153, 153f, 217
 Consanguinity
 reproductive risks, 371
 risks and advantages of consanguineous marriage, 216, 240
 Constitutional abnormalities, 52
 Controls, problems of matching, 339
 Copy number variants (CNVs), 33, 49, 51f, 417g
 and neurodevelopmental vulnerability, 359
 pathogenic versus non-pathogenic, 92, 146
 Cord blood, source of stem cells, 393–4, 396
 COSMIC database, 197
 Cost–benefit ratio (of a program), using discounted cash flow, 324
 Cousins
 definitions of first, double first, second, 6b
 gene sharing by, 241b
 CpG islands, 295, 417g
 CpG sequences, 236, 292, 295, 295b, 417g
 Creatine kinase, 98, 285
 CRISPR–Cas technology, 222, 390
 Crossovers, 37f, 38, 51
 Cryptic splice sites, 149–50, 159, 159f, 224, 417g
CYP2C9, variants affecting drug metabolism, 256, 256f, 257t, 270
CYP2C19, variants affecting drug metabolism, 256, 256f
CYP2D6, variants affecting drug metabolism, 256, 256f, 257b
 Cysteine (Cys, C), chemical formula, 72b
 Cystic fibrosis
 calculating carrier frequencies, 234
 carrier screening, 314–15
 case 2 (Brown family), 2
 clinical features, 3t, 368
 frequencies of *CFTR* variants, 118, 314t
 genotype–phenotype correlation, 167t
 multiplex allele-specific PCR test, 133f
 newborn screening, 313, 313f
 on US Core Disorder screening list, 322t
 presumed heterozygote advantage, 237, 243
 sweat test, 2
 treatment, 132
 Cytochrome c, role in apoptosis, 188
 Cytomegalovirus, reproductive risks, 370
 Cytosine
 chemical formula, 71b
 deamination, 168
- Database
 ClinVar, 221
 ExAC, 135, 164
 Genecards, 221
 GnomAD, 119, 135, 164, 221, 402
 OMIM, 22
De novo deletion, 101

- De novo* mutations
 average number per person, 169
 in autism, 359
 in gene identification, 210, 215
 major cause of severe developmental disorders, 225
- Deafness
 case 18 (Choudhary family), 207
 example of gene identification, 216ff
 gene panel, 136, 137t
- Deamination (of 5-MeC), 236
- Death receptors, 188
- Debrisoquine, adverse reaction to, 257b
- Deciphering Developmental Disorders (DDD)
 study, 225b
- Deformation, definition in dysmorphology, 374b
- Deletion
 15q11–q13 in Angelman and Prader–Willi syndromes, 287t
 chromosomal, *see* Chromosomal abnormalities, deletions
 detection by CGH, 91–3, 92f
 detection by FISH, 89, 97f
 detection by MLPA, 90
 detection by paired-end sequencing, 127f
 detection by SNP chip, 93
 generation by non-allelic homologous recombination, *see* Non-allelic homologous recombination
 in Duchenne muscular dystrophy, 99f; *see also* Muscular dystrophy, Duchenne
 of whole exons, 68f, 158
- Denaturation, 86f, 94f, 95f
- Denaturing high performance liquid chromatography (dHPLC), 121, 418g
- Dental genetics specialty, 219
- 'Designer babies', 400, 401
- Development, an epigenetic process, 283
- Diabetes
 reproductive risks, 369
 transient neonatal, imprinting effects, 291t
 treatment, 386t
- Diabetes Type 2, 349–52
 case 26 (Zuabi family), 334
 mendelian subset (MODY), 349
 molecular pathology, 350f
 prospective predictive studies, 351, 352t
 risk, 335t
 susceptibility factors, 349–51, 351f
 worldwide epidemic, 349
- Diagnosis, importance of making, 371
- Diagnostic odyssey, 371
- Diastrophic dysplasia, 266
 founder effects among Finns, 237t
- Dideoxy nucleotides, chemical formulae, 123f, 418g
- Dideoxy sequencing, *see* Sanger sequencing
- Differentially methylated regions (DMR), near imprinted loci, 290
- Di George–VCFs syndrome, case 7 (Green family), 25
- Digital droplet PCR, 109, 109f, 418g
- Direct-to-consumer (DTC) genetic testing, 326–9, 367, 377
 ethics, 327, 328
 validity and utility of results, 327, 328
- Discounted cash flow, 324
- Disease, risk for consanguineous couple, 240, 244t
- Disease genes, strategies for identifying, 209t, 215f
- Disruption (in dysmorphology), definition, 374b
- DNA
 5' and 3' ends, 60b
 accessibility, 73
 cell-free in circulation, 195b, 329–30, 379
 chemical formulae, 60b, 71b
 junk?, 77
 markers, 105
 non-coding, 77
 repetitive, 76f
 replication (principle), 60f
 sense and template strands, 60b, 62f
 sequencing, 121ff
 denaturing and hybridization, 86ff, 86f
- DNA forensics, 245ff
- DNA ligase, 90, 91, 95
- DNA markers, 211, 212b
- DNA methylation, 292, 292f, 293f
 and Fragile X full expansion, 148, 148f
 and gene silencing, 148
 changeable over lifetime, 293
 how to study, 294, 294f
 in cancer, 198
- DNA polymerase, 60b
- DNA profiling, for criminal investigations, 245
- DNA repair, 168
- Dominant
 effects due to haploinsufficiency, 163, 163f
 negative effect, 158, 158f, 163, 418g
 pedigree patterns, 16, 418g
 X-linked, 16, 18, 18f
- Dominant versus recessive, 163, 163f, 184
- Dosage-sensitive genes, 146, 418g
- Double-strand breaks, 168
- Down syndrome, 32b, 40f, 372
 age-related risk, 42f, 311
 case 8 (Howard family), 26
 due to translocation, 48
 genotype–phenotype correlation, 168t
 prenatal diagnosis, 70, 312b
 prenatal screening, 308, 310–12, 315
- Drift, genetic, 235
- Driver mutation, in cancer, 178, 418g
- Drosophila melanogaster* flies
 as model organisms, 223
 lack of DNA methylation, 292
- Drug responses, variable, 256
- Duchenne muscular dystrophy, *see* Muscular dystrophy, Duchenne
- Duplication
 detection by CGH, 91–2, 92f
 detection by MLPA, 90
 generation by non-allelic homologous recombination, *see* Non-allelic homologous recombination
- Dwarf Sports Association, 144
- Dynamic mutations, 112b
- DYS* (dystrophin) gene, 62t, 68, 68f, 98
 how it was identified, 210
 frameshifting versus frame-neutral variants, 156, 156t
- Dysmorphology, 372ff, 418g
- Dysplasia, definition, 374b
- Dystroglycans, function in muscle, 157f
- Dystrophin, 189
 function in muscle, 157f
 immunostaining, 4f
 molecular pathology, 164
- Elastin, 53
- Electrophoresis, 89b
- Electroporation, 389, 389f
- Embryonic stem (ES) cells, 394, 394f, 418g
- Empiric risks, 335, 383, 418g
- ENCODE project, 77, 150, 298, 418g
- Enhancers, 74, 77, 418g
 action limited by TAD boundaries, 297
 toleration of sequence changes, 148
- ENSEMBL genome browser, 75b
- Epigenetic effects, transgenerational?, 298, 298b
- Epigenetic marks, 283–4
- Epigenetics, two contrasting definitions, 279b, 418g
- Epimutations, 290, 418g
- Error rate, in NGS, 125
- Ethics, of screening, 306, 309–10, 315, 317, 319, 321, 323t, 324, 325, 328
- Euchromatin, 30b, 296, 418g
- Eugenics, scientific flaws, 245
- European Society of Human Genetics (ESHG), recommendations for testing, 325
- ExAC database, 135, 164
- Exome, 76, 419g
- Exome sequencing, 136, 219
 example of filtering list of variants, 155
 example of Karol Kowalski (case 3), 134
 reasons for failure to identify a gene, 223ff
- Exon junction complex, 152
- Exon shuffling, in evolution, 162
- Exons, 60, 62f, 62t, 419g
 numbers in different genes, 62t
- Expanded repeats
 abnormal (non-ATG) translation, 154
 table of examples, 112t
- Expression, variable, 18
- Fabry disease, treatment, 386t
- Factor V Leiden, 320
- False positive and negative rates of a test, definition, 308b
- Familial, definition, 20, 419g
- Familial adenomatous polyposis (FAP), 177, 177f, 184t, 326t
 case 17 (Xenakis family), 176
 multistage development, 189f
 questions about testing children, 196
 role of APC protein, 195–6, 196f
 screening, 196
- Familial hypercholesterolemia (FH)
 cascade screening, 318–19
 case 24 (Smit family), 305
 molecular mechanism, 318
- Familial searching, in criminal investigations, 246
- Family history, how to take, 6b
- Family studies, for estimating heritability, 336–7, 337f
- Fetal anomaly scan, 378
- Fetal exclusion test, 105, 105f
- Fetal sexing, 99, 265, 381
- FGFR3* mutation, positive selection in male germ-line, 160
- Finns, founder effects among, 236
- FISH, *see* Fluorescence *in situ* hybridization
- Fluorescence *in situ* hybridization (FISH), 89, 90f, 97, 97f, 108t, 375, 381, 419g
 abnormalities detectable, 110t
 chromosome painting, 110
 on interphase cells, 90, 108t, 191f
 to detect rearrangement in cancer, 190, 202f

- Fluorouracil, risk of adverse reaction, 257t
- Founder effects
examples, 237t
on allele frequencies, 236f, 419g
versus heterozygote advantage, 238
- Fragile X tremor/ataxia syndrome (FXTAS), 106
- Fragile X syndrome (FRAXA)
case 11 (Lipton family), 83
detecting full expansions, 106
genotype–phenotype correlation, 168t
manifesting as autism, 359
molecular pathology, 164, 112t, 148
- Frameshifts, 150, 152, 153f, 161t, 419g
- Free fetal DNA (in maternal circulation), 329b
- Fugu pufferfish, 77
- Functional genomics, 76, 419g
- G-banding, 27f, 28, 29f, 30b, 419g
- G1–S checkpoint, 186f, 187, 187b
- Gap junctions, 163
- Garrod, Archibald, 255b
- Gas chromatography–mass spectrometry (GC–MS), for checking metabolism, 255
- Gene
ABLI, 182t, 184t
ABO, 37
ACE, 347f
ACTA2, 326t
ACTC1, 326t
ADAMTS9, 351f
ADNP, 302t
AF9, 184t
AFF, 184t
AKT, 199f
AKT1, 170
alpha-globin, 33
AMY1 (salivary amylase), 33, 146
ANKRD26, 203
APC, 186t, 189, 189f, 326t
APOB, 318, 326t
APP, 345t, 347f
ARID1B, 155, 163, 221
ATM, 162f, 186t, 187, 187f, 189, 193t
ATP10A, 290f
ATRX, 302t
AXIN, 190
BCL2, 184t
BCR, 184t
BRAF, 78b, 189f, 190, 271t
BRCA1, 186t, 187, 187b, 189, 191, 326t
BRCA2, 186t, 187, 189, 191, 326t
BRIP1, 193t
C4A / B (complement component 4), 259f
CACNA1S, 326t
CAMK1D, 351f
CBFB, 184t
CCND1, 184t
CDKALI, 351f
CDKN2A (cyclin-dependent kinase), 186t, 189, 189f, 351f
CFTR (cystic fibrosis), see CFTR gene
CH25H, 347f
CHD7, 302t
CHEK2, 193t
CHRNA2, 347f
CLU, 348
COL2A1 (collagen II), see COL2A1 gene
COL3A1, 326t
COL7A1 (collagen VII), 62t
COL11A1 (collagen XI), 70
CREBBP, 300, 301t
CTLA4, 399
CTNNA1, 184t
CYP21A (steroid 21-hydroxylase), 259f, 260
CYP2C9, 256, 256f
CYP2C19, 256, 256f
CYP2D6, 33, 256, 256f
DDIT3, 184t
DDX41, 203
DNMT1, 293
DNMT3A (DNA methyltransferase), 301t
DNMT3B (DNA methyltransferase), 299, 301t
DSC2, 326t
DSG2, 326t
DSP, 326t
DTDST (diastrophic dysplasia sulfate transporter), 266, 266f
DYS (dystrophin), see DYS (dystrophin) gene
EGFR, 199f, 271t
EHMT1, 301t
EP300, 300, 301t
ERBB2, 182t, 199f, 271
ERCC6, 302t
ERK1 / 2, 78b
ETO, 184t
ETV6 (TEL), 184t
EVII, 184t
EZH2, 296, 301t
F8 (blood clotting factor VIII), 147, 147f
FBN1 (fibrillin), 326t
FES, 182t
FGFR2 (fibroblast growth factor receptor 2), 193t
FGFR3 (fibroblast growth factor receptor 3), 160, 162f
FKHR, 184t
FMR1, 88, 106
FMS, 182t
FOXO, 199f
FTO, 351f
FUS, 184t
GAB2, 347f
GJB2 (connexin 26), 153f, 163, 217
GLA, 326t
GNAS1, 73f, 368
GOLGA6L2, 290f
GRB2, 78b
GTNAB, 267
GULO (L-gulonono-gammalactone oxidase), 268
HBB (beta-globin), see Beta-globin gene
HDAC4 / 8, 301t
HEXA (hexosaminidase A), 239
HFE (hemochromatosis), 320
HHEX, 351f
HLA-A / B / DR, 62t, 120, 259, 259f
HNF1B, 351f
HOXA9 / 11 / 13, 184t
HRAS, 78b, 170, 182t
HTT (Huntingtin), 67, 67f
IFNA6 (interferon A6), 62t
IGF2BP2, 351f
IGH (immunoglobulin heavy chain cluster), 184t
IL2RG (interferon receptor subunit), 265, 265f, 286
INS (insulin), 62t
JAZF1, 351f
KAT6B, 301t
KCNE1, 139
KCNE2, 140
KCNH2, 140, 326t
KCNJ11, 351f
KCNQ1, 139, 326t
KDM1A, 296
KDM5C, 301t
KDM6A, 301t
KIT, 271t
KLK4 (kallikrein 4), 220
KMT2A (MLL), 183, 184t
KMT2D (lysine methyltransferase), 300, 301t
KRAS, 78b, 170, 182t, 189f, 271t
LCT (lactase), 268
LDLR (low density lipoprotein receptor), 305, 326t
LMNA, 326t
LSP1, 193t
MAGEL2, 290f
MAP3K1, 193t
MAPT, 347f
MDM2, 187f, 188
MECP2, 297, 300
MEK1 / 2, 78b
MEN1 (multiple endocrine neoplasia), 186t, 326t
MET, 199f
MKRN3, 290f
MLH1, 186t, 188, 326t
MLH3, 188
MLK1, 184t
MLLT1 / 4, 184t
MRE11, 187f
MSH2, 186t, 188, 326t
MSH6, 188, 326t
MTNR1B, 351f
MUTYH, 326t
MYB, 182t
MYBPC3, 326t
MYC, 182t
MYH11 (myosin heavy chain), 184t, 326t
MYH7, 326t
MYL2 / 3, 326t
MYLK, 326t
NBS (Nijmegen breakage syndrome), 187f
NDN, 290f
NF1 (neurofibromatosis 1), 78b, 186t, 189, 199f
NF2 (neurofibromatosis 2), 186t, 326t
NOTCH2, 351f
NPAP1, 290f
NRAS, 78b, 170
NSD1, 210, 301t
NUP98, 184t
OP1MW (green color vision pigment), 33, 146
PAH (phenylalanine hydroxylase), 62t
PALB2, 193t
PAX3, 184t
PCSK9, 126f, 318, 326t
PD1, 399
PDGFRA, 199f
PD-L1, 399
PHF8, 301t
PI3K, 199f
PI3KCA, 170
PICALM, 348
PKP2, 326t
PLAG1, 184t
PML, 184t
PMS1, 188
PMS2, 188, 326t

- POLG1*, 69
PPARG, 351f
PRKAG2, 326t
PRPF3 / 8 / 31 (spliceosome components), 273
PSEN1 / 2, 345t, 347f
PTC (patched), 186t
PTEN, 199f, 326t
PTPN11, 78b
RAD50, 187f
RAF1, 78b
RARA (retinoic acid receptor), 184t
RAS, 190, 199f
RB1, 180, 186t, 187f, 326t
RBM15, 184t
RET, 186t, 326t
RIT1, 78b
RPN1, 184t
RPS6KA3, 301t
RUNX1 (*AML1*), 184t
RYR1 / 2, 326t
SCN5A, 140, 326t
SDHAF2, 326t
SDHB / C, 326t
SDHD, 284, 326t
SETBP1, 215, 301t
SETD2, 296, 301t
SHC, 78b
SHOC2, 78b
SIS, 182t
SLC30A8, 351f
SMAD2 / 4, 189f, 190
SMAD3, 326t
SMARCA2, 302t
SMARCA1, 302t
SMN1, 149, 391
SMN2, 149, 155, 391
SNRPN, 290f
SORL1, 347f
SOS1, 78b
SOX10, 210
SPRED1, 78b
SRC, 182t
SRCAP, 302t
SRY, 32b
SS18, 184t
SSX1 / 2 / 4, 184t
STK11, 326t
SUMF1, 267
SWI/SNF, 302t
TCF7L2, 351f
TGFBR1 (TGF- β receptor), 326t
TGFBR2 (TGF- β receptor), 188, 188f, 190, 326t
THADA, 351f
TMEM43, 326t
TMIE, 217
TNNI3, 326t
TNNT2, 326t
TNRC9, 193t
TP53, 180, 186t, 187, 187f, 189f, 326t
TPM1, 326t
TPMT, 120
TSC1 / 2, 326t
TSPAN8, 351f
TTN (titin), 76, 189
TUBGCP5, 290f
TUPLE1, 97
UBE3A, 287t, 289, 290f
UGT1A1, 257t
VHL (von Hippel–Lindau), 186t, 326t
VKORC1, 257t, 270, 270f
WFS1, 351f
WT1, 326t
 Gene conversion, alternative to recombination, 259, 260b, 419g
 Gene editing, 390
 Gene frequencies, *see* Allele frequencies
 Gene identification, strategies, 209t, 215f
 Gene panels
 for NGS, 136, 137t
 virtual, 136, 137t
 Gene pool, 232, 419g
 Gene silencing, therapy, 390
 Gene supplementation, 390
 Gene therapy, 388ff
 ex vivo, 398f
 for mitochondrial disease, 390, 391f
 germline versus somatic, 390, 390f
 Gene tracking, 105, 105f, 212, 265, 419g
 for pre-implantation diagnosis, 381, 381f
 GeneClinics website, 385
 GeneReviews website, 385
 Genes
 chimeric, *see* Chimeric genes
 disruption by rearrangements, 147, 147f
 FGFR (fibroblast growth factor receptor), 165b
 for noncoding RNA, 59
 tumor suppressor, *see* Tumor suppressor genes
 variable numbers of, 33, 49, 51f
 variable numbers of copies, 146
 whole gene deletions, 146
 Genetic code, 64, 151t, 209
 Genetic drift, 235
 Genetic Information Nondiscrimination Act (USA), 104
 Genetic mapping, *see* Chapter 15
 Genetic markers, 212b, *see also* Chapter 15
 defining ancestral chromosome segments, 339–41
 history of development, 212
 requirements, 212
 Genetic services, 365ff
 Genetic susceptibility factors, history of attempts to identify, 338–9
 Genome-wide association studies (GWAS), 338–44, 342f
 difficulty in identifying causative variants, 344
 missing heritability problem, 353–4
 threshold of significance, 339
 Genotype–phenotype correlation, 18f, 160, 164ff
 in DTDST deficiency, 266f
 in *FGFR* genes, 165b
 in phenylketonuria, 267f
 often looser than previously thought, 167, 225
 Gentamicin, risk of adverse effect, 393, 401
 Germ-line mosaicism, 21, 21f, 420g
 Gleevec, 203
 Glutamic acid (Glu, E), chemical formula, 72b
 Glutamine (Gln, Q), chemical formula, 72b
 Glybera, a cautionary tale, 388
 Glycine (Gly, G), chemical formula, 72b
 GnomAD database, 119, 135, 164, 221, 402
 Golden State killer, how DNA was used in investigation, 245
 GTPases, 79
 Guanine, chemical formula, 71b
 Guthrie card, 262
 GWAS, *see* Genome-wide association studies
 Haploid, 48, 420g
 Haploinsufficiency, 163, 163f, 164, 420g
 Haplotype, 264f, 420g
 Haplotype blocks, 341
 HapMap project, 341, 420g
 Hardy–Weinberg distribution, 232ff, 233b, 420g
 HeLa cells, 180
 Hemochromatosis
 low penetrance of variants, 320
 treatment, 386t
 Hemoglobin S, identification of amino acid change, 255
 Hemophilia
 effect of X-inactivation in a carrier, 281
 in Queen Victoria's descendants, 407
 treatment, 386t
 Herceptin (trastuzumab), 271
 Hereditary breast and ovarian cancer, ACMGG list for mandatory checking, 326t
 Hereditary Hearing Loss Homepage, 216
 Hereditary persistence of fetal hemoglobin, 392
 Heritability, 336–7, 337f, 420g
 Herpes viruses, as vectors for gene therapy, 389, 389f
 Heterochromatin, 30b, 296, 420g
 Heterodisomy (in UPD), 289
 Heteroduplex, detection, 121, 420g
 Heterogeneity
 allelic, 162f, 259, 266, 308, 314t
 locus, 214
 Heteroplasmy (for a mitochondrial variant), 20, 52, 420g
 Heterozygote, manifesting, 285
 Heterozygote advantage, 237
 in cystic fibrosis, 237
 in sickle cell disease, 237, 238f
 versus founder effect, 238
 Hexosaminidase A, 231, 239
 Histidine (His, H), chemical formula, 72b
 Histone code, 296, 420g
 Histone deacetylases (HDACs), 296
 Histone modification, 47, 292, 292f, 295ff, 296f
 writers, readers and editors, 295–6
 Histones, 47, 279
 HLA alleles
 diversity, 259, 259f
 linkage disequilibrium, 264t
 problem of matching for transplant, 264
 Homogentisic acid, in alkaptonuria, 254f, 255
 Homologous chromosomes
 behavior in mitosis versus meiosis, 35ff, 36f, 40f
 definition, 30b, 420g
 Homopolymer runs, 153, 188
 Hotspots, for mutation, 295
 Human Genome Project, 74, 92, 121
 Human Genome Reference Sequence, 58, 74, 122
 Human Genome Variation Society, 120
 Hunter syndrome, 253
 Huntington disease
 case 1 (Ashton family), 1
 diagnostic test, 7b, 103, 104f
 differential diagnoses, 7b
 genotype–phenotype correlation, 167t
 management and treatment, 9, 104, 385, 392, 398
 molecular pathology, 112t, 153
 pedigree, 9f, 15f
 predictive test, 7b, 103–4, 367b
 questions around insurance, 103–4

- Huntington disease – *continued*
 symptoms, 2f
 Hybridization, 86ff, 420g
- Imatinib (Gleevec), 203, 203f, 270
- Immune checkpoint blockade, treatment for cancer, 399
- Immunodeficiency
 gene panel, 137t
 gene therapy, 389
- Immunodeficiency, severe combined (SCID), [case 21 \(Portillo family\)](#), 252
- Immunogenetics
 distinguishing self from non-self, 258ff
 generating antibody diversity, 271ff
- Immunoglobulin
 classes, 272
 genes on chromosomes 2, 14, 22, 286
 mechanisms generating diversity, 272–3
- Immunoreactive trypsin (screening test for cystic fibrosis), 313, 313f
- Immunotherapy, of cancer, 399b
- Imprinted loci, Otago database, 291
- Imprinting, 283ff
 a ‘selfish gene’ response?, 291
 an epigenetic process, 283, 284f
- Imprinting disturbance, multilocus, 291
- Inborn errors of metabolism, 253ff, 255b, 266ff
- Inbreeding, coefficient, *see* Coefficient of inbreeding
 risk of recessive disease, 244t
- Incidental findings, 325–6, 420g
- Induced pluripotent stem cells (iPSC), 222, 393–4, 394f, 420g
- Infections in pregnancy, risks, 370
- Insulin resistance, 334t, 349
- Insurance, questions around genetic testing, 103–4
- Intellectual disability
[case 3 \(Kowalski family\)](#), 3
 example of exome sequencing, 134, 155
 gene panel, 137t
 genotype–phenotype correlation, 167t
 possible causes, 11
 with dysmorphism, 31, 101
- International Cancer Genome Consortium, 198
- International Human Epigenome Consortium, 298
- International Mouse Knockout Consortium, 220, 222
- Introns, 60, 62f, 420g
 containing snoRNA genes, 290
- Inversion (chromosomal), 30b, 421g
- IONIS-HTT (potential treatment for Huntington disease), 9
- iPSC, *see* Induced pluripotent stem cells
- Irinotecan, risk of adverse reaction, 257t
- Isodisomy (in UPD), 289
- Isoleucine (Ile, I), chemical formula, 72b
- Ivacaftor, treatment for p.F508del cystic fibrosis, 154
- Karyotype, normal male, 39f, 421g
- Karyotyping
 abnormalities detectable, 110t
 obsolete?, 38
 technique, 27, 27f, 28b
- Kniest dysplasia, 158
- Knudson, Alfred, two-hit hypothesis, 184, 185f
- Kozak sequence, 64
- Lambda value (risk to relative), 335
- Leber hereditary optic neuropathy (LHON) and mitochondrial DNA variants, 69, 130, 157
[case 6 \(Fletcher family\)](#), 5
 genotype–phenotype correlation, 167t
 pedigree, 13f
 retinal appearance, 5f
- Lentiviruses, as vectors for gene therapy, 389, 389f
- Leucine (Leu, L), chemical formula, 72b
- Leukemia, acute lymphocytic (ALL), 175f, 184t, 190
[case 15 \(Tierney family\)](#), 175
 chromosomal rearrangements in, 184t
 factors affecting treatment, 396
 treatment with azathioprine, 261
- Leukemia, chronic myeloid (CML), 184t, 202b
 treatment with imatinib, 203
- Lifestyle genetic testing, *see* Direct-to-consumer genetic testing
- LINEs (long interspersed nuclear elements), 77
- Linkage analysis, 211ff, 211f, *see also* Chapter 15
 in familial cancers, 185, 186t, 191, 195
- Linkage disequilibrium, 341, 421g
 example of HLA haplotypes, 264t
- Linkage versus association, 338b, 421g
- Liposomes, 389, 389f
- Liquid biopsies, 194, 195b, 402
- Lithium, reproductive risks, 370
- Locus heterogeneity, 214, 421g
- Lod score, 211, 421g, *see also* Chapter 15
- Long-read NGS technologies, 127
- Loss of function, not necessarily pathogenic, 160
- Loss of heterozygosity, in cancer, 189f, 421g
- Loss versus gain of function, 161
 spectrum of variants, 162, 162f
- Low density lipoprotein receptor, 318
- Lyonization, *see* X-inactivation
- Lysine, methylation and acetylation in histones, 296f
- Lysine (Lys, K), chemical formula, 72b
- Macrophages, 259
- Major histocompatibility complex (MHC), 258–9, 259f
- Malaria, 237, 316
- Male-lethal condition, 19
- Malformation, definition, 374b
- Malformation sequence, definition, 374b
- Malformation syndrome, definition, 374b
- Malformations, multiple, 366
- Management and treatment, 385ff
- Manhattan plot, 339f
- Manifesting carriers, of an X-linked condition, 285, 421g
- Maple syrup urine disease, treatment, 386t
- Marfan syndrome, 326t, 367, 373
 treatment with losartan or atenolol, 388
- Mass spectrometry, 111
- Massively parallel sequencing, *see* Next-generation sequencing
- Matchmaker Exchange, 215, 220, 402
- Mating, random versus assortative, 232–3
- Matrilineal inheritance, 16, 20
- Meiosis, recombinant versus non-recombinant, 211
- Meiosis, 35ff, 36f
 consequences of non-disjunction, 41f
- Melanoma, familial, 186t
- Melting curve analysis, 121
- Mendelian characters, are exceptional, 165, 421g
- Messenger RNA, *see* mRNA
- Meta-analysis, 346, 421g
- Metabolic block, 253f, 254f
- Metabolic pathway, of phenylalanine and tyrosine, 254f
- Metabolic syndrome, 334t, 352
- Metabolizers (of drugs), poor, intermediate, extensive, ultra, 256, 256f
- Metacentric chromosomes, 31b, 421g
- Metaphase, 35, 36f, 421g
- Methionine (Met, M), chemical formula, 72b
- Methylation, of promoter, 106
- Methylation of DNA, *see* DNA methylation
- Methyltransferase, maintenance (DNMT1), 293, 293f
- M-FISH, 110, 179b
- MHC, *see* Major histocompatibility complex
- Microarrays, 87, 91, 422g
 expression arrays, 197
- Microdeletions, 31, 33b, 101, 422g
 Y chromosome, 368
- Microduplications, 31, 33b
- Micro-exons, 223
- MicroRNA, in cancer, 198, 422g
- Microsatellite instability, 188, 422g
- Microsatellites, 212, 422g
- Migration, effect on allele frequencies, 235
- Minor allele frequency (MAF), 343
- Mismatch repair, 188, 422g
- Mis-sense changes, 150, 151, 161t
 programs to predict effect, 151
- Missing heritability, problem with GWAS, 353–4
- Mitelman catalog of chromosomal rearrangements in cancer, 184t
- Mitochondrial genome, 69, 69f
- Mitochondrial inheritance, 16b, 20
- Mitochondrial replacement therapy, 390, 391f
- Mitosis, 35ff, 36f
- MLPA, *see* Multiplex ligation-dependent probe amplification
- Monoclonal antibodies, for treatment of cancer, 270–1
- Monosomy, autosomal, 32b, 422g
- Mosaicism, 20ff, 52ff, 169b, 422g
 confined placental, 171b, 330b
 detection by FISH, 90
 germ-line, 21, 21f
 in Turner syndrome, 42
 when it is clinically significant, 21
- mRNA
 nonsense-mediated decay, *see* Nonsense-mediated RNA decay
 stability, 64
 testing, 111
- Multidisciplinary teams (MDT), 366, 402
- Multifactorial conditions, basic epidemiology, 334, 422g
- Multiple enzyme deficiencies, examples, 267
- Multiple malformations, 366
- Multiplex ligation-dependent probe amplification (MLPA), 90, 91f, 99f, 108t, 375, 422g
- Muscular dystrophy, Becker versus Duchenne, genotype–phenotype correlation, 167t
- Muscular dystrophy, Duchenne, 98, 156, 156t
[case 4 \(Davies family\)](#), 4
 dystrophin gene deletions, 68, 68f, 98, 99f

- muscle histology, 4f, 285, 285f
- obligate carriers, 11f
- options for testing, 381–2, 382f
- pedigree, 11f
- proportion of new mutations, 243
- proposal for screening newborns, 315
- prospects for gene therapy, 397
- Mutagenesis, insertional, 389, 397
- Mutation, different definitions of the word, 145b
- Mutation–selection dynamic, 242, 243, 243f
- Mutation hotspots, 153
- Mutation rate, 178
- Mutation signatures, in cancer, 180
- Mutations
 - activating oncogenes, 182
 - de novo*, 169, 215
 - driver versus passenger in cancer, 178
 - nomenclature, 120b
 - somatic, 169
- Myotonic dystrophy
 - anticipation in, 113f
 - causative change is in untranslated region of mRNA, 224
 - details of expanded repeats, 112t
 - founder effects in Quebec–Sanguenay, 237t
 - reproductive risks, 369
- N-acetyltransferase, fast and slow acetylators, 258
- NAHR, *see* Non-allelic homologous recombination
- National Screening Committee (UK), 315, 321, 323b
 - key criteria for a screening program, 323t
- Natural killer (NK) cells, 252b, 265
- Natural selection
 - effect on allele frequencies, 235, 236
 - for lactose tolerance, 268b
- Neurofibromatosis 1 (NF1), 21b, 78b, 186t, 367
 - high frequency of new mutations, 242
- Newborn screening, 309, 313, 313t, 315, 321
- Next-generation sequencing (NGS), 122, 125ff, 422g
 - aligning reads, 126f, 134
 - error rate, 125
 - filtering the list of variants, 134
 - long-read technologies, 127
 - paired-end, 127, 127f
 - read length and depth, 125, 126f
 - strategies, 136, 375
 - use of virtual gene panels, 136, 137t
 - variant-calling programs, 127, 134
- NGS, *see* Next-generation sequencing
- Nitisinone, treatment for Type I tyrosinemia, 255
- Nomenclature
 - for chromosomal abnormalities, 30b
 - of DNA sequence variants, 120b
- Non-allelic homologous recombination (NAHR), 33, 53f, 97, 101, 422g
- Non-disjunction, 41f
- Non-invasive prenatal testing (NIPT), 315, 329b, 378, 379, 402
 - for single gene disorders, 330
- Non-parametric linkage, 338, 422g
- Non-penetrance, example, 17, 17f
- Nonsense changes, 150, 152, 161t, 265f, 423g
- Nonsense-mediated RNA decay, 152, 152f, 156, 196, 423g
- Normal transmitting male (in Fragile X syndrome), 107
- Nuchal translucency, test for Down syndrome, 311, 311t
- Nucleosomes, 47, 47f, 279, 423g
- Nucleotides, 59, 423g
- Nusinersen, treatment for SMA, 392
- Obligate carrier (of an X-linked condition), 285, 423g
- Odds ratio, 343b, 423g
 - definition, 308b
 - for breast cancer susceptibility factors, 193t
 - normally small for GWAS variants, 343
- Okazaki fragments, 180, 423g
- Oligonucleotide probes, 86, 423g
- Oligonucleotides, allele-specific, *see* Allele-specific oligonucleotides
- Oncogenes, 178, 181ff, 423g
 - activation processes, 182–3
- One gene – multiple proteins, 73f
- One gene – one enzyme hypothesis, 58, 71, 255, 423g
- Opportunistic screening, 308
- Organoids, 222
- Orphanet, 385
- Otoacoustic emissions, 207, 208f
- Oxford Nanopore NGS technology, 127
- Oxidative phosphorylation, effect of mitochondrial defects, 157
- p14, p16 proteins, products of *CDKN2A* gene, 187f
- p53 protein, product of *TP53* gene, *see* Gene, *TP53*
- P450 cytochromes, 256, 269
- P values, correction for multiple testing, 339
- PacBio NGS technology, 127
- Paired-end sequencing, 127, 127f
- Paracentric inversion, 49, 51f
- Para-hydroxy phenylalanine dioxygenase (PHPA), deficiency, 254f
- Parent–child trios, for identifying *de novo* variants, 215
- Paris Convention, 30
- PARP inhibitors, treatment for ovarian cancer, 271t
- Passenger mutation (in cancer), 178, 423g
- Paternal age effects
 - due to dynamics of spermatogenesis, 160
 - with *FGFR3*, *HRAS*, *RET*, *PTPN11* genes, 160, 244
- PCR, 93ff, 94f, 95f, 95b, 108t
 - allele-specific (ARMS test), 119, 119f, 130f
 - annealing temperature, 96
 - digital droplet, 109, 109f
 - multiplex, 95
 - primers, 94
 - problems with GC-rich sequences, 87, 106
 - QF (quantitative fluorescence), 109, 109f, 312
 - real-time, 109
 - RT–PCR, 111
 - single cell, 381
 - triplet-primed, 106, 107f
- Pedigree
 - how to draw, 6b
 - symbols, 6b
- Pedigree pattern
 - autosomal dominant, 15f, 16b
 - autosomal recessive, 16b
 - mitochondrial, 16b
 - X-linked, 11f, 16b
 - Y-linked, 16b
- Pedigree patterns
 - imprinted, 284
 - mendelian are exceptional, 165
- Pedigrees, how to interpret, 14
- Penetrance
 - definition, 17, 423g
 - reduced, 17
- Pericentric inversion, 49
- Personalized prescribing, *see* Stratified medicine
- Pharmacodynamics, definition, 256, 423g
- Pharmacogenetics, 256ff, 423g
- Pharmacokinetics, definition, 256, 423g
- Phenotypic spectrum of a disorder, often broader than previously thought, 225
- Phenylalanine (Phe, F), chemical formula, 72b
- Phenylalanine hydroxylase
 - deficiency in phenylketonuria, 254f
 - how the gene was identified, 210
- Phenylketonuria (PKU)
 - [case 20 \(Vlasi family\)](#), 251
 - metabolic pathway, 254f
 - newborn screening, 317
 - poor correlation between enzyme activity and IQ, 165, 267f
 - risk to baby of affected mother, 263f, 369
 - treatment, 262, 386t, 396
- Philadelphia chromosome, in chronic myeloid leukemia, 202
- Phosphorus-32 (³²P) labeling, 87
- Phytohemagglutinin, 27f
- PI3k–Akt–mTOR pathway, 170
- Pleiotropy, 3b, 424g
- pLI statistic, in ExAC database, 164
- Polygenic model, of non-mendelian conditions, 354, 424g
- Polygenic risk scores, 348, 354–5, 355t, 383, 402, 424g
 - fail to identify most cases, 356, 358
 - for breast cancer, 193
 - population-specific, 355
 - promising tool for identifying high-risk individuals, 356, 358
- Polymorphic, definition, 212, 424g
- POLYPHEN-2 program, for predicting effect of an amino acid substitution, 151, 155, 223
- Population attributable risk, 320, 321b, 424g
- Population stratification, 338, 424g
 - invalidates Hardy–Weinberg expectations, 235
- Positional cloning, 210, 210f, 424g
- Positive predictive value of a test
 - definition, 308b, 424g
 - depends on frequency of condition, 320
- Post-translational modification (of a protein), 65
- Prader–Willi syndrome, 33b, 372
 - [case 23 \(Rogers family\)](#), 278
 - causes, 287, 287t
 - role of *SNHG14* RNA, 290, 290f
 - test for UPD, 288f
- Predictive test, for Huntington disease, 7b, 103, 424g
- Pre-implantation genetic diagnosis, 212, 381b
- Premutation alleles (of unstable expansions), 106, 424g
- Prenatal diagnosis, 378ff
 - factors to consider before decision, 378
- Prenatal screening, 308, 309f
 - differing attitudes, 324

- Primary transcript, 62, 74, 424g
 Primer, in DNA replication, 180
 Primers, for sequencing, 122, 125, 424g
 Prior probability, 384, 424g
 Privacy, and DNA databases, 245
 Probes, 86, 89, 424g
 Prodrugs, 256, 425g
 Proline (Pro, P), chemical formula, 72b
 Promoter, problems assessing variants in, 224, 425g
 Proof-reading, in DNA replication, 168
 Prophase, 34, 36f, 425g
 Protein functional domains, 76, 76f, 162
 Proteome, 111, 425g
 Proteomic testing, scope and limitations, 111
 Proto-oncogenes, *see* Oncogenes
 Pseudoautosomal region, 37, 280, 425g
 Pseudogenes, 22b, 77, 425g
 PWS / AS critical region (15q11–13), genes and transcripts, 290f
 Pyrosequencing, 131f, 132
- Quadrivalent, 46f, 51
 Quantitative fluorescence (QF) PCR, 109, 109f, 312, 425g
- Random mating, 232, 425g
 RAS–MAPK pathway, 78b, 167, 182
 R-banding (of chromosomes), 28
 Reactive oxygen species, 168
 Read length and depth, in NGS, 125, 126f, 425g
 Reading frame, 64
 Real-time PCR, 109, 425g
 Reasons for referral to genetic services, 366–8, 367b
 Receptor-mediated endocytosis, 389, 389f
 Recessive disease, relative risk for consanguineous couple, 244t
 Reciprocal translocation
 balanced, 43, 44f
 case 5 (Elliot family), 4
 consequences for family, 12, 12f
 origin, 45f
 segregation in meiosis, 46f
 unbalanced segregation product, 44f, 100f
 Reciprocal translocations, detection by interphase FISH, 90
 Recombinants, in linkage analysis, 211, 211f, *see also Chapter 15*
 Recombination, 38, 425g
 molecular mechanism, 260b
 Recombination hotspots, 340
 Reference Human Genome Sequence, 213
 Relationship
 coefficient of, 240–1, 242f
 definitions, 6–7
 extent of gene sharing, 241b
 Repetitive DNA, 76f
 Restriction enzymes, 88b, 425g
 Restriction enzymes *HpaII*, *MspI*, use to study DNA methylation, 294
 Restriction site, testing for, 119, 130f, 131f
 Retinitis pigmentosa, 367
 Retinoblastoma, 184, 186t, 326t
 sporadic versus familial, 185f
 Retinoids, reproductive risks, 370
 Retroviruses, acute transforming, *see* Acute transforming retroviruses
 as vectors for gene therapy, 389, 389f
- Reverse-phase protein arrays, 198
 Reverse transcriptase, 59, 111
 Reverse transcriptase (RT)–PCR, 111, 426g
 Ribosomes, 63, 63f
 Right not to know (of a genetic risk), 104
 Ring chromosomes, 49f
 Risk, relative versus absolute, 320, 343, 425g
 Risk assessment, 383
 Risk from inbreeding, depending on allele frequency, 244t
- RNA
 long non-coding, 77
 messenger, 63
 polymerase, 62f
 primary transcript, 62, 62f
 pros and cons of testing, 111
 ribosomal, 63
 snoRNA, 290
 toxic, 113
 transfer, 64
 RNA-seq, 111
 Roadmap Epigenomics Consortium, 298
 Robertsonian translocation, 31b, 50f, 426g
 ROC curve
 definition, 347, 348b, 426g
 statistics for Alzheimer susceptibility factors, 347
 statistics for Type 2 diabetes susceptibility factors, 352t
- Sanger sequencing, 122ff, 124f, 426g
 advantages over NGS, 124
 SANTA CRUZ genome browser, 75b
 Sarcoglycans, function in muscle, 157f
 Savior siblings, 400
 Screening
 adults, 309
 carrier, 309, 314, 316–17
 cascade, 306, 309, 310, 313
 economics of, 324
 ethical issues in, 306, 309, 315, 317, 321, 324, 325, 328
 for cystic fibrosis, 313, 313f, 314–15
 for Down syndrome, 308, 310–12, 315
 for Tay–Sachs carriers, 310, 316–17
 newborn, 309, 313, 313t, 315, 317, 321
 opportunistic, 308
 prenatal, 308, 309f
 Screening test, measures of performance, 308b
 Screening versus diagnostic tests, 306, 307f, 426g
 Selection in spermatogonia, effects with certain genes, 244
 Self versus non-self, distinguishing, 259
 Selfish gene theory, applied to imprinting, 291
 Senescence, of cells in culture, 180
 Sensitivity of a test, definition, 308b, 426g
 Sequence variants
 nomenclature, 120b
 summary of types affecting genes, 146
 typical numbers in a genome, 127
 Sequencing
 massively parallel, *see* Next-generation sequencing
 Sanger (dideoxy), *see* Sanger sequencing
 Serine (Ser, S), chemical formula, 72b
 Sewall Wright's path coefficient method, example calculations, 241b, 242f
 Sex-limited conditions, 19, 426g
- Sibs, gene sharing by, 241b, 426g
 Sickle cell disease
 detecting the pathogenic variant, 119f
 molecular pathology, 164
 on US Core Disorder screening list, 322t
 SIFT program, for predicting effect of an amino acid substitution, 151, 155, 223
 Signal peptide, 65, 426g
 SINEs (short interspersed nuclear elements), 77
 Single nucleotide polymorphisms (SNPs or SNVs), 212, 426g
 Single strand conformation polymorphism (SSCP) analysis, 121, 426g
 Single-cell genomics, 200, 381
 Sister chromatids, 28, 31b, 34, 35, 90, 426g
 Skin biopsy, 28b
 SKY karyotyping, 110, 179b
 Small nucleolar RNA (snoRNA), coded in introns of *SNHG14* RNA, 290
 SNP chips, 87, 91, 93, 102f, 108t, 119, 339, 375, 426g
 low analytical validity for rare variants, 328, 377
 SNPs, *see* Single nucleotide polymorphisms
 SNVs, *see* Single nucleotide polymorphisms
 Somatic mutations, 169, 426g
 Southern blotting, 87ff, 87f, 88b, 106, 108t, 426g
 Specificity of a test, definition, 308b
 Spinal muscular atrophy, 149
 on US Core Disorder screening list, 322t
 role of SMN protein in spliceosome, 274
 treatment, 391b
 Splice isoforms, 150, 426g
 Splice sites
 cryptic, 149–50
 donor and acceptor, 63
 effect of variants, 149–50, 159
 Spliceosome, 62, 149, 273b, 427g
 pathogenic variants of spliceosomal proteins, 273–4
 Splicing (of primary transcript), 62, 62f
 alternative, 71, 150
 Statins, for treatment of FH, 305, 319
 Stem cells, 393, 393f, 427g
 Stickler syndrome, 158
 case 10 (O'Reilly family), 57
 COL2A1 mutation, 134
 genotype–phenotype correlation, 168t
 Stop codon, 64, 427g
 Stratification (of population), invalidates Hardy–Weinberg expectations, 235
 Stratified medicine, 258, 269ff, 402, 427g
 in cancer, 270–1, 271t
 Structural variants, detection by paired-end sequencing, 127f
 Submetacentric chromosomes, 31b, 427g
 Succinylacetone, toxic by-product in Type I tyrosinemia, 254
 Succinylcholine, risk of adverse reaction, 257t, 258
 Support groups, 144, 225, 385
 Susceptibility factors, for complex disease, *see* Genetic susceptibility factors
 SYBR Green, 121
 Syndromes, *see Disease index*
 Synonymous changes, 151, 161t
- T cell receptors, mechanisms generating diversity, 273
 T lymphocytes, 252b, 259, 265, 286, 399
 Tag-SNPs, 341

- Tamoxifen, 194
- Tay–Sachs disease
 carrier screening, 310, 316–17
[case 19 \(Ulmer family\)](#), 231, 232f
 high frequency among Ashkenazi Jews, 239
 pathogenic variants in Jews and non-Jews, 240t
- Telomerase, 34, 181, 427g
- Telomeres, 31b, 34, 427g
 shortening, 180b
- Testicular biopsy, 28b, 35
- Testing
 carrier, 376
 cascade, 376
 children – pros and cons, 197, 239
 direct-to-consumer, *see* Direct-to-consumer genetic testing
 lifestyle genetic, *see* Direct-to-consumer genetic testing
 metabolic, 375
 predictive, 376
 pre-implantation, *see* Pre-implantation genetic diagnosis
 prenatal, *see* Prenatal diagnosis
 single-cell, 381b
- Tests
 for a specific sequence variant, 118–19, 130f, 133f, 135
 for any variant in a gene, 120–1
 for disease susceptibility, 338–44
 for DNA methylation, 294, 294f
 for microdeletions or microduplications, 91, 93, 97, 97f, 101
 gene panel, 136–7
 liquid biopsies, 194, 195b, 402
 screening versus diagnostic, 306, 307f
- Tetrahydrobiopterin (BH4), deficiency rare cause of PKU, 262, 317
- Thalassemia, *see* Alpha-thalassemia; Beta-thalassemia
 benefits of prenatal testing, 380
- The Cancer Genome Atlas (TCGA) project, 197b
- Thiopurine methyltransferase (TPMT)
 low-activity alleles, 262, 262f
 role in metabolism of azathioprine, 261, 261f
- Threonine (Thr, T), chemical formula, 72b
- Thymine, chemical formula, 71b
- Thyroid cancer, targeted drugs, 271t
- Tissue types, *see* HLA alleles
- Topologically associating domains (TADs), 297
- Transcription, 58f, 62, 62f
 inhibited by promoter methylation, 106
 regulation, 72ff
- Transcription factors, 74, 427g
 primary role in regulation of transcription, 292, 298
- Transcriptome, 111
- Transfer RNA, *see* tRNA
- Transgenerational epigenetic effects, controversy, 298, 298b
- Translation (protein synthesis), 58f, 63f
- Translocation, reciprocal, *see* Reciprocal translocation
- Translocation, Robertsonian, *see* Robertsonian translocation
- Translocations, X;autosome, *see* X;autosome translocations
- Transplantation, tissue matching, 258
- Transposons, 77, 427g
- Trios (parent–child), for identifying *de novo* variants, 215
- Triploidy, detection methods, 110t, 427g
- Trisomy, 31b, 427g
 detection methods, 110t
 risk versus maternal age, 42f
- Trisomy rescue, cause of UPD, 288, 289f, 427g
- tRNA, 64, 427g
- Tryptophan (Trp, W), chemical formula, 72b
- Tuberous sclerosis, treatment with Sirolimus, 388
- Tumor biopsy, 28b
- Tumor suppressor (TS) genes, 178, 184ff, 427g
- Turner syndrome (45,X), 368
 a puzzling question, 285
 a suggestion of imprinting, 286
[case 9 \(Ingram family\)](#), 26
 genotype–phenotype correlation, 168t
 karyotype, 43f
 management and treatment, 43
 mechanism and risk, 42
 mosaicism, 42
 need to check for Y chromosome mosaicism, 70, 103
- Twin concordance, 337, 427g
- Twin studies, for estimating heritability, 336–7, 337f
- Twins
 monozygotic, 38
 monozygotic versus dizygotic, 337
- Two-hit hypothesis, in cancer, 184, 185f
- Tyrosinase, deficiency in albinism, 253, 254f
- Tyrosine (Tyr, Y), chemical formula, 72b
- Tyrosine aminotransferase, deficiency in tyrosinemia Type II, 254f
- Tyrosine kinases, 182f, 182t
BCR–ABL1 fusion product, 202
- Tyrosinemia Type I, 254f
 treatment with Nitisinone, 255, 386t, 387
- Uniparental disomy (UPD), 283, 428g
 heterodisomy versus isodisomy, 289
 origin through trisomy rescue, 288, 289f
 using DNA markers to check, 288f
- Unique (support group), 385
- Unstable repeats, *see* Dynamic mutations
- Uracil, chemical formula, 71b
- Valine (Val, V), chemical formula, 72b
- Variable expression, 18, 428g
- Variant, pathogenicity wrongly assessed, 223
- Vectors, for gene therapy, 389, 389f, 428g
- Victoria, Queen, 407
- Virtual gene panels, for analyzing exome sequence, 136, 137t
- Viruses, retroviruses, *see* Retroviruses
- Vitamin C, inability of humans to synthesize, 268b
- Vitamin K, role in blood clotting, 270, 270f
- Waardenburg syndrome, 321
- Warfarin
 difficulty of optimising dosage, 270
 mechanism of action, 270f
 metabolism, 270f
 risk of adverse reaction, 257t
- Wellcome Trust Case–Control Consortium, 338, 339f
- WGS, *see* Whole genome sequencing
- Whole genome sequencing (WGS), 137
- World Dwarf Games, 144f
- Wrong assessments, of pathogenicity, 223
- X;autosome translocation
 causing Duchenne muscular dystrophy, 282
 skewed X-inactivation in a carrier, 282, 282f
- X;autosome translocations, 52
- X-inactivation, 280ff, 281f, 428g
 an epigenetic process, 280
 genes escaping inactivation, 285
 in a carrier of an X-linked condition, 281
 in a carrier of X-linked immunodeficiency, 282, 286
- X-inactivation center, 282
- X-linked dominant inheritance, 16b, 18, 19f
- X-linked pedigree pattern, 16b
- X–Y homologous genes, 286
- Y-linked inheritance, 16b, 19
- Zebrafish, as model organisms, 223
- Zika virus, reproductive risks, 370

Disease index

- 3-hydroxy-3-methylglutaric aciduria, 322t
 3-methylcrotonyl-CoA carboxylase deficiency, 322t
 46,XX males, 368b
 47,XXX females, 32b, 280
 47,XXY males (Klinefelter syndrome), 32b, 368b
 47,XYY males, 32b, 280
 Achondrogenesis, 158, 266
 Achondroplasia, *see* Achondroplasia in *main index*
 Albinism, 253
 Alpers syndrome, 69
 Alpha-thalassemia / mental retardation syndrome (ATRX), 300, 301f, 302t, 375
 Alport syndrome, 385
 Alzheimer disease, *see* Alzheimer disease in *main index*
 Amelogenesis imperfecta, 219, 219f
 Amenorrhea, 367
 Androgen insensitivity syndrome, 368b
 Anhidrotic ectodermal dysplasia, 281
 Anophthalmia / microphthalmia, 137t
 Apert syndrome, 166f, 166t
 Argininosuccinic aciduria, 322t
 Arrhythmogenic right ventricular cardiomyopathy (ARVC), 138, 326t
 Aspartylglucosaminuria, 237t
 Ataxia-telangiectasia, 162f, 186t
 Atelosteogenesis type II, 267
 Autism, *see* Autism in *main index*
- Baratella–Scott syndrome, 224
 Bardet–Biedl syndrome, 237t
 Beare–Stevenson cutis gyrata, 166t
 Beckwith–Wiedemann syndrome, 290, 291t
 Beta-ketothiolase deficiency, 322t
 Biotinidase deficiency, 322t
 Blindness, 136
 Brachydactyly with intellectual disability, 301t
 Brugada syndrome, 138, 140
 Burn–McKeown syndrome, 274
 Butyrylcholinesterase deficiency, 237t
- Cancer
 breast, *see* Breast cancer in *main index*
 Burkitt's lymphoma, 183, 183f, 397
 colon, *see* Colon cancer in *main index*
 familial adenomatous polyposis (FAP), *see* Familial adenomatous polyposis in *main index*
 gastrointestinal stromal tumors, 271t
 glioblastoma multiforme, 199f
 Gorlin syndrome, 186t
 leukemia, acute lymphocytic (ALL), *see* Leukemia, acute lymphocytic in *main index*
 leukemia, chronic myeloid (CML), *see* Leukemia, chronic myeloid in *main index*
 Li–Fraumeni syndrome, 326t
 liposarcoma, 184t
 lung, 271t
 lymphoma, follicular, 184t
 lymphoma, Hodgkin, 203
 Lynch syndrome, 186t, 326t
 melanoma, 186t, 271t
 multiple endocrine neoplasia, 186t, 326t
 myelodysplastic syndrome, 203
 ovarian, 271t, 326t
 paragangliomas (glomus body tumors), 284f, 326t
 prostate, 194
 PTEN hamartoma tumor syndrome, 326t
 renal (von Hippel–Lindau disease), 186t, 326t, 367
 retinoblastoma, *see* Retinoblastoma in *main index*
 rhabdomyosarcoma, 184t
 synovial sarcoma, 184t
 thyroid, 271t, 326t
 Wilms tumor, 326t
 Catecholaminergic polymorphic ventricular tachycardia (CPVT), 138, 326t
 Celiac disease, 335t
 Cerebrocostomandibular syndrome, 274b
 CFC syndrome, 78b
 Channelopathies, 138, 139
 Charcot–Marie–Tooth disease, 237t
 CHARGE syndrome, 302t
 Chickenpox, 370
 Chondrodysplasias, 158
 Chromosome breakage syndromes, 375
 Ciliopathies, 137t
 Citrullinemia, 322t
 Claes–Jensen syndrome, 301t
 Cleft lip, 374f
 CLOVES syndrome, 170, 388
 Cockayne syndrome B, 302t
 Coffin–Lowry syndrome, 301t
 Coffin–Siris syndrome, 302t
 Cohen syndrome, 375
 Congenital adrenal hyperplasia, 322t, 368b
 Congenital hypothyroidism, 137t
 Cornelia de Lange syndrome, 301t
 Costello syndrome, 78b
 Creatine synthesis deficiency, 386t
 Cri du chat syndrome, 31, 33b
 Crouzon syndrome, 166f, 166t
 Cystic fibrosis, *see* Cystic fibrosis in *main index*
 Cystinosis, 154, 386t, 387, 387f
- Deafness, *see* Deafness in *main index*
 Dentatorubral–pallidolusian atrophy, 112t
- Diabetes Type 1, 335t, 349
 Diabetes Type 2, *see* Diabetes Type 2 in *main index*
 Diastrophic dysplasia, 237t, 266
 Di George–VCFS syndrome, 25, 33b, 40, 97
 Down syndrome, *see* Down syndrome in *main index*
 Duchenne muscular dystrophy, *see* Muscular dystrophy, Duchenne in *main index*
 Dwarfism, pituitary, 386t
- Edwards syndrome, 32b, 372
 Ehlers–Danlos syndrome, 326t, 367
 Ellis–van Creveld syndrome, 238
 Epileptic encephalopathy, 137t
- Fabry disease, 386t
 Familial hematuria, 137t
 Familial hypercholesterolemia, *see* Familial hypercholesterolemia in *main index*
 Familial thoracic aortic aneurysms and dissections, 326t
 Fanconi syndrome, 375
 Fetal alcohol syndrome, 370
 Floating Harbor syndrome, 302t
 Fragile X syndrome (FRAXA), *see* Fragile X syndrome in *main index*
 Fragile X syndrome (FRAXE), 112t
 Fragile X tremor/ataxia syndrome (FXTAS), 106
 Friedrich ataxia, 112t
 Frontotemporal dementia, 112t
- Galactosemia, 322t
 Gaucher disease, 375, 386t
 Genitopatellar syndrome, 301t
 Glutaric acidemia, 322t
 Glycogen storage disease Type II (Pompe), 322t
 Graft-versus-host disease, 263
 Guion–Almeda mandibulofacial dysostosis, 274
- Hearing loss, *see* Deafness in *main index*
 Helsmoortel–van der Aa syndrome, 302t
 Hemochromatosis, 320, 386t
 Hemophilia, 281, 386t
 Hermansky–Pudlak syndrome, 237t
 Holocarboxylase synthase deficiency, 322t
 Homocystinuria, pyridoxine-responsive, 386t
 Huntington disease, *see* Huntington disease in *main index*
 Hurler syndrome, 253
 Hydatidiform mole, 283, 283f, 291
 Hydrops fetalis, 33
 Hypercholesterolemia, 386t
 Hypertension, 334
 Hypochondrogenesis, 158
 Hypochondroplasia, 166, 166f, 166t

- Hypophosphatemia, X-linked, 18, 19f
- ICF syndrome, 299, 299f, 301t
- Immunodeficiency, 137t, 389
- Infertility, 19, 27, 35, 368b
- Intellectual disability, *see* Intellectual disability in *main index*
- Isovaleric acidemia, 322t
- Jervell and Lange–Nielsen syndrome, 139b, 163
- Kabuki syndrome, 137t, 300, 300f, 301t
- Kagami–Ogata syndrome, 291t
- Kallmann syndrome, 368b
- Kleefstra syndrome, 301t
- Klinefelter syndrome, 32b, 368b
- Kniest dysplasia, 158
- Lactose intolerance, 268b
- Larsen syndrome, 166
- Leber hereditary optic neuropathy (LHON), *see* Leber hereditary optic neuropathy in *main index*
- Legius syndrome, 78b
- Liposarcoma, 184t
- Loeys–Dietz syndrome, 326t
- Long QT syndrome, 138b, 139f
- Long-chain L-3 hydroxyacyl-CoA dehydrogenase deficiency, 322t
- Luscan–Lumish syndrome, 301t
- Lysosomal storage diseases, 253
- Malignant hyperthermia susceptibility, 326t
- Maple syrup urine disease, 322t, 386t
- Marfan syndrome, 326t, 367, 373, 388
- Marie Unna hypotrichosis, 224
- McCune–Albright syndrome, 368b
- Medium-chain acyl-CoA dehydrogenase deficiency, 322t
- Melanocytic skin nevi, congenital, 170
- Methylmalonic acidemia, 322t
- Miller syndrome, 223
- Miller–Dieker syndrome, 31, 33b
- Miscarriage, 27, 367
- MODY (maturity onset diabetes of youth), 349
- Mucopolipidosis II, 267
- Mucopolysaccharidoses, 253, 322t
- Muenke craniosynostosis, 166f, 166t
- Multilocus imprinting disturbance, 291
- Multiple epiphyseal dysplasia, 266
- Multiple malformations, 366
- Multiple sclerosis, 335t
- Multiple sulfatase deficiency, 267, 268f
- Muscular dystrophy, Becker, 156, 156t, 167t
- Muscular dystrophy, Duchenne, *see* Muscular dystrophy, Duchenne in *main index*
- Muscular dystrophy, limb-girdle, 157
- Myelodysplastic syndrome, 203
- Myotonic dystrophy 1, *see* Myotonic dystrophy 1 in *main index*
- Myotonic dystrophy 2, 112t
- Nager syndrome, 274
- Nephronophthisis, 385
- Neural tube defects, 307, 307f, 308, 374f
- Neurofibromatosis 1 (NF1), *see* Neurofibromatosis 1 in *main index*
- Neurofibromatosis 2 (NF2), 186t, 326t, 367
- Neuronal ceroid lipofuscinosis, 237t
- Nicolaides–Baraitser syndrome, 302t
- Noonan syndrome, 78b
- Organic aciduria, cobalamin-responsive, 386t
- Osteogenesis imperfecta, 374f, 386t
- Oto-palatal-digital syndrome, 166
- Ovarian teratoma, 283, 283f
- Overgrowth, 366
- Pallister–Killian syndrome, 170, 171f
- Patau syndrome (trisomy 13), 32b, 374f
- Peutz–Jeghers syndrome, 326t
- Phenylketonuria (PKU), *see* Phenylketonuria in *main index*
- Polycystic kidney disease, 367, 385
- Pompe disease, 386t
- Potocki–Lupski syndrome, 33b
- Prader–Willi syndrome, *see* Prader–Willi syndrome in *main index*
- Progressive myoclonic epilepsy, 112t
- Propionic acidemia, 322t
- Proteus syndrome, 170
- Pseudo-Friedreich ataxia, 386t
- Pseudohypoparathyroidism 1A, 291t
- RASopathies, 78b
- Refsum disease, 386t
- Retinitis pigmentosa, 273, 367
- Rett syndrome, 19, 299, 300f, 359, 372
- Rhabdomyosarcoma, 184t
- Romano–Ward long QT syndrome, 139, 140, 163, 326t
- Rubella, 370
- Rubinstein–Taybi syndrome, 300, 300f, 301t
- SBMA (spinobulbar muscular atrophy), 112t
- Schimke immuno-osseous dysplasia, 302t
- Schimmelpenning syndrome, 170
- Schinz–Giedion syndrome, 215, 301t
- Schizophrenia, 101
- Severe combined immunodeficiency, 322t, 397, 398f
- Short stature, 366
- Sickle cell disease, *see* Sickle cell disease in *main index*
- Siderius syndrome, 301t
- Silver–Russell syndrome, 291t
- Skeletal dysplasias, 166
- Smith–Lemli–Opitz syndrome, 386t
- Smith–Magenis syndrome, 33b, 372
- Sotos syndrome, 210, 301t
- Spinal muscular atrophy, *see* Spinal muscular atrophy in *main index*
- Spinobulbar muscular atrophy, 112t
- Spinocerebellar atrophy (SCA1–17), 112t
- Spondyloepiphyseal dysplasia (SED), 158
- Stickler syndrome, *see* Stickler syndrome in *main index*
- Sudden arrhythmic death syndrome (SADS), 138b
- Supravalvular aortic stenosis (SVAS), 53
- Syndrome families, 166
- Talipes, 374f
- Tatton–Brown–Rahman syndrome, 301t
- Taybi–Linder syndrome, 224, 274
- Tay–Sachs disease, *see* Tay–Sachs disease in *main index*
- Temple syndrome, 291t
- Thalassemia, *see* Alpha-thalassemia, Beta-thalassemia in *main index*
- Thanatophoric dysplasia, 166, 166f, 166t
- Toxoplasmosis, 370
- Trifunctional protein deficiency, 322t
- Triploidy, 31b, 32b, 48, 110t
- Trisomy 8, 52
- Trisomy 13 (Patau syndrome), 32b, 374f
- Trisomy 18 (Edwards syndrome), 32b, 372
- Trisomy 21, *see* Down syndrome in *main index*
- Tuberous sclerosis, 326t, 388
- Turner syndrome (45,X), *see* Turner syndrome in *main index*
- Type 2 diabetes, *see* Diabetes, Type 2 in *main index*
- Tyrosinemia Type I, 254f, 255, 386t, 387
- Usher syndrome type 1C, 237t
- VCFS (Di George) syndrome, *see* Di George–VCFS syndrome in *main index*
- Very long-chain acyl-CoA dehydrogenase deficiency, 322t
- Von Recklinghausen disease, *see* Neurofibromatosis 1 in *main index*
- Waardenburg syndrome, 210, 321
- Weaver syndrome, 301t
- Werdnig–Hoffmann disease (SMA), 149
- Williams–Beuren syndrome, 33b, 53ff, 372
- Wolf–Hirschhorn syndrome, 31, 33b
- Xanthomata, 305, 306f
- X-linked adrenoleukodystrophy, 322t

NEW CLINICAL GENETICS

FOURTH EDITION A GUIDE TO GENOMIC MEDICINE

New Clinical Genetics is used worldwide as a textbook for medical students, but also as an essential guide to the field for genetic counselors, physician assistants, clinical and nurse geneticists, and students studying healthcare courses allied to medicine. Readers love the integrated case-based approach which ties the science to real-life clinical scenarios to really aid understanding.

Clinical genetics is a fast-moving field and there have been many advances in the few years since the previous edition was published. This 4th edition has been completely updated and revised to reflect new science, new techniques and new ways of thinking.

Nowhere is this more clear than in the chapter discussing genetics services which is now significantly expanded to reflect the increasing role of genomic medicine and the use of multidisciplinary teams in the management of patients with genetic disorders.

The unique case-based structure and format remains the same, but substantial new material has been added to cover:

- polygenic risk scores – now starting to become useful clinical service tools
- preimplantation diagnosis
- noninvasive prenatal diagnosis
- companion diagnostics for prescribed drugs
- liquid biopsies in cancer
- epigenetics and gene regulation
- the widespread use of next-generation sequencing as a routine diagnostic tool
- the checking of a patient's whole exome for the cause of their problem

New Clinical Genetics continues to offer the most innovative case-based approach to investigation, diagnosis, and management in clinical genetics.

REVIEWS OF PREVIOUS EDITIONS

"this book provides a wonderful case-based learning environment. Excellent!"

Human Genetics

"this book is a very valuable tool that will be used by future geneticists all over Europe and beyond, both as a teaching material and as a source of excellent knowledge."

European Journal of Human Genetics

INSTRUCTOR COMMENTS

"I LOVED the book. I've never seen anything like it, and I've reviewed a lot of genetics texts. The way that cases are presented throughout is extremely novel."

"I am greatly pleased with the revisions. In my opinion, there is an increased clarity in the text (which will serve students well), and many welcomed updates based on current literature. Good job!"

"This is a fantastic book that I enjoy so much teaching from."

"The book looks good and we will certainly be recommending it for our medical genetics course this autumn."

"I have used this book every year since the first edition was published and it is a perfect fit for my human genetics course. I will definitely continue to use it."

"It's great. I will recommend the book as a main text for the medical student class."

