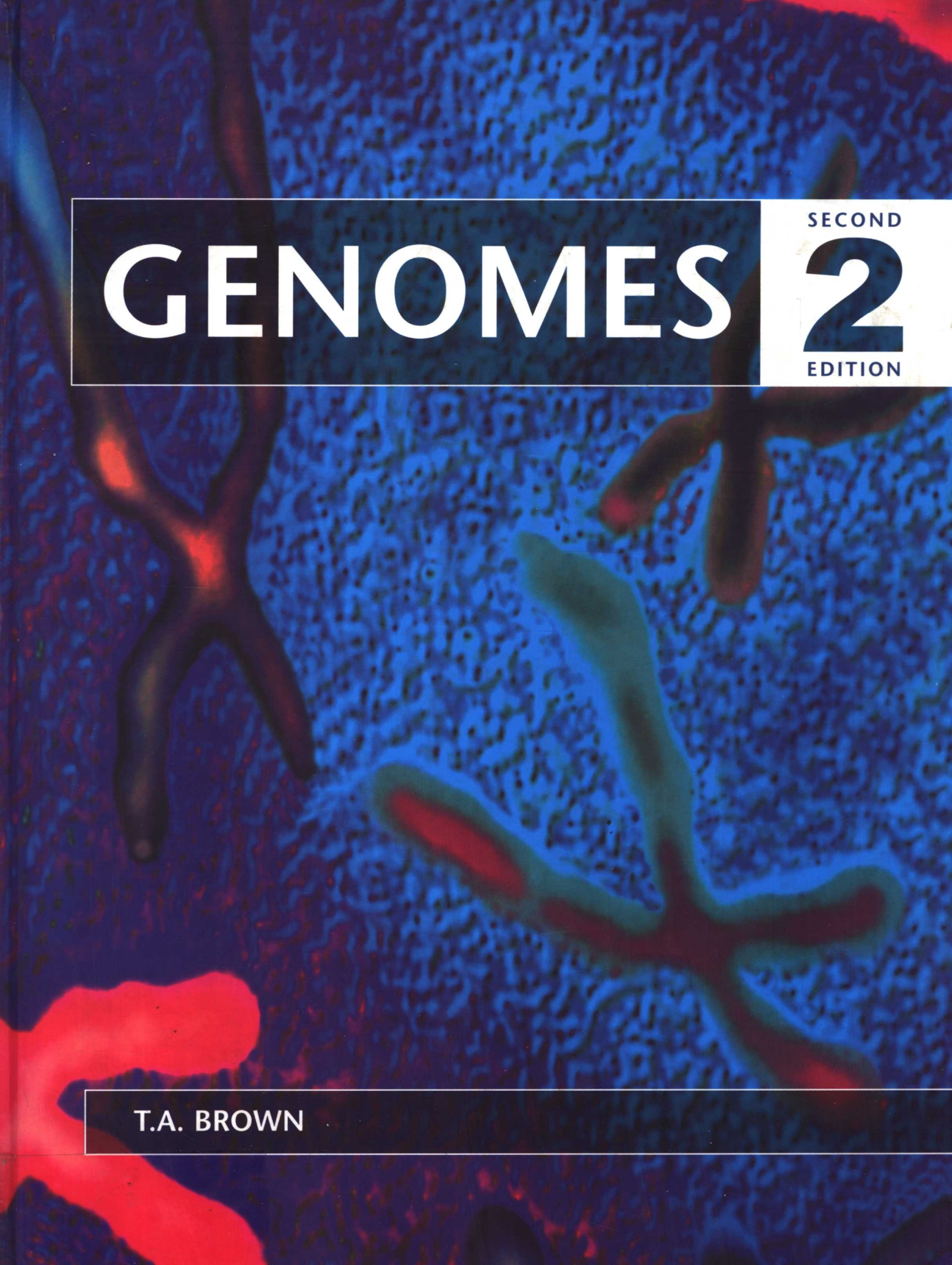
GENOMES
SECOND
2
EDITION
T.A. BROWN

GENOMES 2

SECOND EDITION

T.A. BROWN

Department of Biomolecular Sciences, UMIST, Manchester, M60 1QD, UK

WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., by arrangement with BIOS Scientific Publishers Limited, 9 Newtec Place, Magdalen Road, Oxford OX4 1RE, UK

This edition published in the United States of America, its dependent territories, Central and South America, Canada, Australia, Brunei, Cambodia, Hong Kong, India, Indonesia, Laos, Macau, Malaysia, New Zealand, People's Republic of China, Philippines, Singapore, South Korea, Taiwan, Thailand, the South Pacific and Vietnam only and not for export therefrom.

First published 1999
Second Edition 2002

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

ISBN 0-471-25046-5

Library of congress Cataloging-in-Publication Data

Genomes / edited by Terence A. Brown. -- 2nd ed.
p. ; cm.
Rev. ed. of: Genomes / T.A. Brown. 1999.
Includes bibliographical references and index
ISBN 0-471-25046-5 (alk. paper)
1. Genomes.
[DNLM: 1. Genome. QH 477 G33605 2002] I. Brown., T.A. (Terence A.)
QH447 .B76 2002
572.8'6--dc21

2002003471

USA
John Wiley & Sons Inc.,
605 Third Avenue, New York,
NY 10158–0012, USA

Canada
John Wiley & Sons (Canada) Ltd,
22 Worcester Road, Rexdale,
Ontario M9W 1L1, Canada

Project Manager: Helen Barham PhD
Production Editor: Sarah Carlson
Designed, typeset and illustrated by J&L Composition Ltd, Filey, North Yorkshire, UK
Printed by Ajanta Offset, New Delhi, India

Abbreviations

5-bU	5-bromouracil
A	adenine; alanine
ABF	ARS binding factor
Ac/Ds	activator/dissociation
ADAR	adenosine deaminase acting on RNA
ADP	adenosine 5′-diphosphate
AIDS	acquired immunodeficiency syndrome
ala	alanine
AMP	adenosine 5′-monophosphate
ANT-C	Antennapedia complex
AP	apurinic/apyrimidinic
arg	arginine
ARS	autonomously replicating sequence
asn	asparagine
ASO	allele-specific oligonucleotide
asp	aspartic acid
ATP	adenosine 5′-triphosphate
ATPase	adenosine 5′-triphosphatase
BAC	bacterial artificial chromosome
bis	*N, N*′-methylenebisacrylamide
bp	base pair
BSE	bovine spongiform encephalopathy
BX-C	Bithorax complex
C	cysteine; cytosine
cAMP	cyclic AMP
CAP	catabolite activator protein
CASP	CTD-associated SR-like protein
cDNA	complementary DNA
CEPH	Centre d'Études du Polymorphisme Humaine
cGMP	cyclic GMP
CHEF	contour-clamped homogeneous electric fields
Col	colicin
CPSF	cleavage and polyadenylation specificity factor
CRM	chromatin remodeling machine
CstF	cleavage stimulation factor
CTAB	cetyltrimethylammonium bromide
CTD	C-terminal domain
CTP	cytidine 5′-triphosphate
cys	cysteine
D	aspartic acid
DAG	diacylglycerol
Dam	DNA adenine methylase
DAPI	4, 6-diamino-2-phenylindole dihydrochloride
DASH	dynamic allele-specific hybridization
dATP	2′-deoxyadenosine 5′-triphosphate
DBS	double-stranded DNA binding site
Dcm	DNA cytosine methylase
dCTP	2′-deoxycytidine 5′-triphosphate
ddATP	2′, 3′-dideoxyadenosine 5′-triphosphate
ddCTP	2′, 3′-dideoxycytidine 5′-triphosphate
ddGTP	2′, 3′-dideoxyguanosine 5′-triphosphate
ddNTP	2′, 3′-dideoxynucleoside 5′-triphosphate
ddTTP	2′, 3′-dideoxythymidine 5′-triphosphate
dGTP	2′-deoxyguanosine 5′-triphosphate
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
Dnmt	DNA methyltransferase
dNTP	2′-deoxynucleoside 5′-triphosphate
dsRAD	double-stranded RNA adenosine deaminase
dsRBD	double-stranded RNA binding domain
dTTP	2′-deoxythymidine 5′-triphosphate
E	glutamic acid
EDTA	ethylenediamine tetraacetate
eEF	eukaryotic elongation factor
EEO	electroendosmosis value
EF	elongation factor
eIF	eukaryotic initiation factor
EMS	ethylmethane sulfonate
eRF	eukaryotic release factor
ERV	endogenous retrovirus
ES	embryonic stem
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
F	fertility; phenylalanine
FEN	flap endonuclease
FIGE	field inversion gel electrophoresis
FISH	fluorescent *in situ* hybridization
G	glycine; guanine
G1	gap phase 1
G2	gap phase 2
GABA	γ-aminobutyric acid
GAP	GTPase activating protein
Gb	gigabase pair
GDP	guanosine 5′-diphosphate
GFP	green fluorescent protein
gln	glutamine
glu	glutamic acid
gly	glycine
GMP	guanosine 5′-monophosphate
GNRP	guanine nucleotide releasing protein
GTF	general transcription factor
GTP	guanosine 5′-triphosphate

H	histidine
HAT	hypoxanthine + aminopterin + thymidine
HBS	heteroduplex binding site
HDAC	histone deacetylase
his	histidine
HIV	human immunodeficiency virus
HLA	human leukocyte antigen
HMG	high mobility group
HNPCC	hereditary non-polyposis colorectal cancer
hnRNA	heterogenous nuclear RNA
HOM-C	homeotic gene complex
HPLC	high-performance liquid chromatography
HPRT	hypoxanthine phosphoribosyl transferase
HTH	helix–turn–helix
I	isoleucine
ICF	immunodeficiency, centromere instability and facial anomalies
IF	initiation factor
IHF	integration host factor
ile	isoleucine
Inr	initiator
Ins(1,4,5)P_3	inositol-1,4,5-trisphosphate
IRE-PCR	interspersed repeat element PCR
IRES	internal ribosome entry site
IS	insertion sequence
ITR	inverted terminal repeat
JAK	Janus kinase
K	lysine
kb	kilobase pair
kDa	kilodalton
L	leucine
LCR	locus control region
leu	leucine
LINE	long interspersed nuclear element
LTR	long terminal repeat
lys	lysine
M	methionine; mitosis phase
MALDI-TOF	matrix-assisted laser desorption ionization time-of-flight
MAP	mitogen activated protein
MAR	matrix-associated region
Mb	megabase pair
MeCP	methyl-CpG-binding protein
met	methionine
MGMT	O^6-methylguanine-DNA methyltransferase
mRNA	messenger RNA
Myr	million years
N	2′-deoxynucleoside 5′-triphosphate; asparagine
NAD	nicotinamide adenine dinucleotide
NADH	reduced nicotinamide adenine dinucleotide
NHEJ	non-homologous end joining
NJ	neighbor-joining
NMD	nonsense-mediated RNA decay
NMR	nuclear magnetic resonance
NTP	nucleoside 5′-triphosphate
OFAGE	orthogonal field alternation gel electrophoresis
ORC	origin recognition complex
ORF	open reading frame
OTU	operational taxonomic unit
P	proline
PAC	P1-derived artificial chromosome
PADP	polyadenylate binding protein
PAUP	Phylogenetic Analysis Using Parsimony
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
phe	phenylalanine
PHYLIP	Phylogeny Inference Package
PIC	pre-initiation complex
PNA	peptide nucleic acid
PNPase	polynucleotide phosphorylase
pro	proline
PtdIns(4,5)P_2	phosphatidylinositol-4,5-bisphosphate
PTRF	polymerase I and transcript release factor
Pu	purine
Py	pyrimidine
Q	glutamine
R	arginine; purine
RACE	rapid amplification of cDNA ends
RAM	random access memory
RBS	RNA binding site
RC	replication complex
RF	release factor
RFC	replication factor C
RFLP	restriction fragment length polymorphism
RHB	Rel homology domain
RLF	replication licensing factor
RMP	replication mediator protein
RNA	ribonucleic acid
RNase	ribonuclease
RNP	ribonucleoprotein
RPA	replication protein A
RRF	ribosome recycling factor
rRNA	ribosomal RNA
RT-PCR	reverse transcriptase-PCR
RTVL	retroviral-like element
S	serine; synthesis phase
SAGE	serial analysis of gene expression
SAP	stress activated protein
SAR	scaffold attachment region
SCAF	SR-like CTD-associated factor
scRNA	small cytoplasmic RNA
SCS	specialized chromatin structure
SDS	sodium dodecyl sulfate
SeCys	selenocysteine
ser	serine
SINE	short interspersed nuclear element
SIV	simian immunodeficiency virus
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SRF	serum response factor

SSB	single-strand binding protein
SSLP	simple sequence length polymorphism
STAT	signal transducer and activator of transcription
STR	simple tandem repeat
STS	sequence tagged site
T	threonine; thymine
TAF	TBP-associated factor
TBP	TATA-binding protein
TEMED	*N, N, N′, N′*-tetramethylethylenediamine
TF	transcription factor
thr	threonine
Ti	tumor inducing
TIC	TAF and initiator-dependent cofactor
TK	thymidine kinase
T_m	melting temperature
tmRNA	transfer-messenger RNA
Tn	transposon
TOL	toluene
TPA	tissue plasminogen activator
TRAP	*trp* RNA-binding attenuation protein
tRNA	transfer RNA
trp	tryptophan
tyr	tyrosine
U	uracil
UCE	upstream control element
UTP	uridine 5′-triphosphate
UTR	untranslated region
UV	ultraviolet
val	valine
VNTR	variable number of tandem repeats
W	adenine or thymine; tryptophan
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Y	pyrimidine; tyrosine
YAC	yeast artificial chromosome
YIp	yeast integrative plasmid

Preface to the Second Edition

Three exciting years have elapsed since publication of the first edition of *Genomes*. Draft sequences have appeared for the fruit fly, *Arabidopsis* and human genomes, and prokaryotic genome sequences are now published at the rate of two or three per month. Experimental techniques for studying the transcriptome and proteome have begun to mature and are providing novel insights into genome expression. And as well as these new directions, the genome expression and replication processes continue to be described in ever-increasing detail. All of these advances have been incorporated into this second edition of *Genomes*. The human genome is now the central feature of Chapter 1, followed immediately by a survey of the physical and genetic organizations of genomes in general, with Part 1 completed by an overview of the transcriptome and proteome. Part 2, on the methods used to study the genome, has been supplemented by the addition of an entirely new chapter on cloning techniques and PCR, which were interspersed in a rather unsatisfactory manner throughout the first edition. The chapters on sequencing and functional analysis have been updated and extended to reflect changes in technology since 1999. Part 3, describing genome expression, has been given a thorough update, as has Part 4 on genome replication and evolution. A number of readers commented on how up-to-date the first edition of *Genomes* was, and I hope that I have been able to retain this quality in the new edition.

Other changes have been designed to make the book more user friendly. The reorganization of material in Part 1 gives a more gentle introduction for students who are encountering molecular biology for the first time, and each chapter now ends with a series of study aids that I hope will be useful both as a guide to revision and in directing supplementary tutorial work. I have also prefaced each chapter with a set of learning outcomes, these being perhaps the most useful of the teaching innovations forced on UK universities by the quality-assessment initiatives of recent years.

I would like to say a general thank you to the many people who have been kind enough to send me comments and suggestions for the second edition of *Genomes*. I hope that you will recognize the changes, large and small, that I have made in response to your feedback. Also I thank Jonathan Ray and Simon Watkins of BIOS for the tremendous support that they provided when I was writing *Genomes*, and Sarah Carlson and Helen Barham for ensuring that the production phase was not a stressful experience. Finally, this second edition of *Genomes* would not have appeared without the support of my wife, Keri. In the Acknowledgements to the First Edition I wrote, 'if you find this book useful then you should thank Keri, not me, because she is the one who ensured that it was written', and I am pleased that one or two people actually took me up on this.

T.A. Brown
Manchester

Preface to the First Edition

Genomes attempts to bring a fresh approach to the teaching of undergraduate molecular biology. It starts with the premise that the syllabus for a university course in molecular biology should reflect the major research issues of the new millennium rather than those topics that were in vogue during the 1970s and 1980s. The book is therefore centered on genomes, not genes, in recognition of the fact that today's molecular biology is driven less by research into the activities of individual genes and more by genome sequencing and functional analysis. Many of today's molecular biology undergraduates will be involved in genome research when they begin their graduate careers and all of them will find their work influenced in one way or another by genome projects. If the objective of undergraduate teaching is to prepare students for their future careers then they must be taught about genomes!

It would of course be foolish to suggest that genes are no longer important. The major challenge that I faced when writing *Genomes* was to combine the essential elements of the traditional molecular biology syllabus with the new material relating to genomes. It is not yet possible to describe adequately the events leading from DNA to protein entirely in terms of 'genome to proteome', hence a substantial part of *Genomes* is devoted to the expression pathways of individual genes. This book differs from many others in that it attempts to describe these expression pathways in the context of the activity and function of the genome as a whole. Similarly, DNA replication, mutation and recombination are dealt with largely in terms of their effects on the genome, and not simply as processes responsible for the replication and alteration of genes.

My belief that molecular biology teaching should be centered on genomes grew as I wrote this book and discovered how much more satisfying and informative the approach is compared with the traditional syllabus. A number of topics that in the past have seemed to me to be of peripheral interest have fallen into place and taken on new relevance. I hope that at least some of the excitement that I felt while writing *Genomes* is conveyed to the reader.

T.A. Brown
Manchester

An Introduction to *Genomes*

I have tried to make the second edition of *Genomes* as user friendly as possible. The book therefore includes a number of devices intended to help the reader and to make the book an effective teaching aid.

Organization of the Book

Genomes is divided into four parts:

- **Part 1 – Genomes, Transcriptomes and Proteomes** introduces the central concepts of modern molecular biology. Chapter 1 begins with DNA and then summarizes the key features of the human genome, with Chapter 2 extending the survey to the genomes of eukaryotes and prokaryotes in general. Chapter 3 then uses the new concepts of the transcriptome and the proteome to introduce the basic steps in genome expression. By the end of Part 1 the reader will have acquired a good working knowledge of the structures and organizations of genomes and will understand, in outline, how the information contained in the genome is released and made available to the cell.
- **Part 2 – Studying Genomes** begins with an orientation chapter that introduces the reader to the methods, centered on cloning and PCR, that were used in the pre-genome era to examine individual genes. The techniques that are more specifically used for studying genomes are then described in the order in which they would be used in a genome project: methods for constructing genetic and physical maps (Chapter 5); DNA sequencing methodology and the strategies used to assemble a contiguous genome sequence (Chapter 6); and methods for identifying genes in a genome sequence and determining the functions of those genes in the cell (Chapter 7). The Human Genome Project forms a continuous thread throughout Part 2, but this is not to the exclusion of all else and I have tried to give adequate coverage to the strategies that have been used, and are being used, to understand the genomes of other organisms.
- **Part 3 – How Genomes Function** covers the material that in the past has been described (inadequately in my opinion) as 'DNA goes to RNA goes to protein'. Chapter 8 addresses the increasingly important issue of how chromatin structure influences genome expression. Chapter 9 then describes the assembly of the transcription initiation complexes of prokaryotes and eukaryotes, and includes a fairly detailed discussion of DNA-binding proteins, these playing the central roles in the initial stages of genome expression. Chapters 10 and 11 give details of the synthesis of RNA and protein, and Chapter 12 surveys the regulation of genome activity. Keeping Chapter 12 to a manageable length was difficult, as many different topics are relevant to genome regulation, but I hope that by using specific examples to illustrate general themes I have managed to achieve a satisfactory balance between conciseness and breadth of coverage.
- **Part 4 – How Genomes Replicate and Evolve** links DNA replication, mutation and recombination with the gradual evolution of genomes over time. In Chapters 13 and 14 the molecular processes responsible for replication, mutation, repair and recombination are described, and in Chapter 15 the ways in which these processes are thought to have shaped the structures and genetic contents of genomes over evolutionary time are considered. Finally, Chapter 16 is devoted to the increasingly informative use of molecular phylogenetics to infer the evolutionary relationships between DNA sequences.

Organization of Chapters

Learning outcomes

Each chapter starts with a set of learning outcomes. These have been phrased very carefully. They are not merely a series of synopses of the factual content of each chapter, but instead indicate the level and type of knowledge that the student should gain from reading the chapter. Therefore, the learning outcomes state what the student should be able to describe, draw, discuss, explain, evaluate, etc., each verb having been selected to convey precisely what it is that the student is expected to be able to do. The intention is that the student is left in no doubt about what they should get out of each chapter, and hence is in no doubt about whether they have dealt satisfactorily with the material.

Figures

A good diagram is certainly worth a thousand words but a bad one can confuse the reader and a superfluous one is merely distracting. I have therefore tried to ensure that every figure is necessary and fulfils a purpose beyond simply breaking up the text and making the book look pretty. I have also tried to make figures reproducible because in my opinion this makes them much more useful as a learning aid for the student. I have never understood the penchant for making textbook diagrams into works of art because if the student cannot redraw a diagram then it is merely an illustration and does not help the student learn the information that it is designed to convey. The figures in *Genomes* are as clear, simple and uncluttered as possible.

Boxes, Technical Notes and Research Briefings

The main text in each chapter is supported and extended by additional information, separated into three distinct categories:

- **Boxes** contain discrete packages of information that I have taken out of the main text, either for emphasis or to avoid disrupting the flow of the text. Some boxes summarize the key points regarding a topic that is described at length in the text, or provide a pointer towards a later topic that has a bearing on the issues being discussed. Other boxes are used to give a more extended coverage of interesting topics, and some describe current speculation regarding areas that have not yet been resolved.
- Each **Technical Note** is a self-contained description of a technique or a group of techniques important in the study of genomes. The Technical Notes are designed to be read in conjunction with the main text, each one being located at the place in the book where an application of that technique is described for the first time.
- **Research Briefings** are designed to illustrate some of the strategies that are used to study genomes. Each Briefing is based on one or a few research papers and explains the background and rationale of a research project, describes how the resulting data were analyzed, and summarizes the conclusions that were drawn. The objective is to illustrate the way in which real research is conducted and to show how research into molecular biology has established the 'facts' about genomes.

Reading lists

The reading lists at the end of each chapter are divided into two sections:

- **References** are lists of articles that are cited in the text. *Genomes* is not itself a research publication and the text is not referenced in the way that would be appropriate for a review or scientific paper. Many points and facts are not referenced at all, and those citations that are given are often review articles rather than the relevant primary research papers. In several cases, for example, I have referred to a *Science* Perspective or *Nature* News and Views article, rather than a research paper, because these general articles are usually more helpful in explaining the context and relevance of a piece of work. My intention throughout *Genomes* has been that the reference lists should be as valuable as possible to students writing extended essays or dissertations on particular topics.
- **Further Reading** contains books and review articles that are not referred to directly in the main text but which are useful sources of additional material. In most cases I have appended a short summary stating the particular value of each item to help the reader decide which ones he or she wishes to seek out. The lists are not all-inclusive and I encourage readers to spend some time searching the shelves of their own libraries for other books and articles. Browsing is an excellent way to discover interests that you never realized you had!

Study aids

Each set of study aids is divided into three sections: key terms, self study questions and problem-based learning.

Key terms

This is a list of the important words and short phrases that the student will have encountered for the first time when reading the chapter. A short definition is required for each one. All of the terms are highlighted in the text and defined in the Glossary, so the student can check the accuracy of their answers after they have completed the exercise. Short definitions of this kind are a useful type of revision aid: if a student can accurately define every key term then they almost certainly have an excellent knowledge of the factual content of the chapter.

Self study questions

These require 100–500 word answers, or occasionally ask for an annotated diagram or a table. The questions cover the entire content of each chapter in a straightforward manner, and they can be marked simply by checking each answer against the relevant part of the text. A student can use the self study questions to work systematically through a chapter, or can select individual ones in order to evaluate their ability to answer questions on specific topics. The self study questions could also be used in closed-book examinations.

Problem-based learning

This is a student-centered activity in which a group of students research a problem and, through their studies,

obtain the information more normally delivered by a teacher-centered activity such as a lecture. Most students and teachers who have adopted this educational tool believe that it is a more effective means of learning than the traditional approaches, and is also more fun. The questions vary in nature and in difficulty. Some are reasonably straightforward and merely require a literature survey, the intention with these problems being that the students take their learning a few stages on from where *Genomes* leaves off. Some problems require that the students evaluate a statement or a hypothesis, which could be done by reading around the subject but which, hopefully, will engender a certain amount of thought and critical awareness. A few problems are very difficult, to the extent that there is no solid answer to the question posed. These are designed to stimulate debate and speculation, which stretches the knowledge of each student and forces them to think carefully about their statements. Ideally, problem-based learning is conducted as a group exercise, each group comprising 5–10 students, with an exercise lasting 1–2 weeks and being carried out through a series of meetings between the group and a facilitator, interspersed with meetings that the students conduct on their own. The facilitator helps the students to organize their thoughts, steers them away from unproductive lines of research, and points out any serious omissions in their approach. The output from the exercise is a written report, a poster, an oral presentation, or a combination of these things. Most of the problems given in *Genomes* are suitable for any type of output, and many can also be adapted for use as discussion topics in tutorials. There are no answers at the back of book! To provide answers would defeat the purpose – the intention is that the students discover a solution for themselves.

Appendix – Keeping up to Date

The Appendix gives the reader advice regarding the best way to keep up to date with the latest research discoveries. It is divided into two sections. The first section covers the various journals and other publications that include reviews and news articles on genome research, and the second section contains a list of some of the many Internet sites that contain relevant information.

Glossary

I am very much in favor of glossaries as learning aids and I have provided an extensive one for this second edition of *Genomes*. Every term that is highlighted in bold in the text is defined in the Glossary, along with a number of additional terms that the reader might come across when referring to books or articles in the reading lists. Each term in the Glossary also appears in the index, so the reader can quickly gain access to the relevant pages where the Glossary term is covered in more detail.

Contents in Brief

PART 1	**Genomes, Transcriptomes and Proteomes**	1
Chapter 1	**The Human Genome**	3
Chapter 2	**Genome Anatomies**	29
Chapter 3	**Transcriptomes and Proteomes**	69
PART 2	**Studying Genomes**	93
Chapter 4	**Studying DNA**	95
Chapter 5	**Mapping Genomes**	125
Chapter 6	**Sequencing Genomes**	163
Chapter 7	**Understanding a Genome Sequence**	187
PART 3	**How Genomes Function**	219
Chapter 8	**Accessing the Genome**	221
Chapter 9	**Assembly of the Transcription Initiation Complex**	239
Chapter 10	**Synthesis and Processing of RNA**	273
Chapter 11	**Synthesis and Processing of the Proteome**	313
Chapter 12	**Regulation of Genome Activity**	347
PART 4	**How Genomes Replicate and Evolve**	381
Chapter 13	**Genome Replication**	383
Chapter 14	**Mutation, Repair and Recombination**	417
Chapter 15	**How Genomes Evolve**	459
Chapter 16	**Molecular Phylogenetics**	483

Contents

Abbreviations xvii

Preface to the Second Edition xxi

Preface to the First Edition xxiii

An Introduction to *Genomes* xxv

PART 1 Genomes, Transcriptomes and Proteomes 1

Chapter 1 The Human Genome 3

1.1 DNA 5
1.1.1 Genes are made of DNA 6
Bacterial genes are made of DNA 6
Virus genes are made of DNA 8
1.1.2 The structure of DNA 8
Nucleotides and polynucleotides 9
RNA 10
1.1.3 The double helix 11
The evidence that led to the double helix 13
The key features of the double helix 14
Box 1.1: Base-pairing in RNA 14
The double helix has structural flexibility 14
Box 1.2: Units of length for DNA molecules 16

1.2 The Human Genome 16
1.2.1 The content of the human nuclear genome 18
Genes and related sequences 19
The functions of human genes 21
Box 1.3: How many genes are there in the human genome? 22
Pseudogenes and other evolutionary relics 22
Genome-wide repeats and microsatellites 22
Box 1.4: The organization of the human genome 23
1.2.2 The human mitochondrial genome 24

1.3 Why is the Human Genome Project Important? 24

Study Aids 26

Chapter 2 Genome Anatomies 29

2.1 An Overview of Genome Anatomies 30
2.1.1 Genomes of eukaryotes 31
2.1.2 Genomes of prokaryotes 33

2.2 The Anatomy of the Eukaryotic Genome 35
2.2.1 Eukaryotic nuclear genomes 35
Packaging of DNA into chromosomes 35
Technical Note 2.1: Agarose gel electrophoresis 37

The special features of metaphase chromosomes 37
Box 2.1: Unusual chromosome types 39
Where are the genes in a eukaryotic genome? 41
What genes are present in a eukaryotic genome? 41
Technical Note 2.2: Ultracentrifugation techniques 42
Families of genes 43
Box 2.2: Two examples of unusual gene organization 44
2.2.2 **Eukaryotic organelle genomes** 46
Physical features of organelle genomes 46
The genetic content of organelle genomes 46
The origins of organelle genomes 46

2.3 The Anatomy of the Prokaryotic Genome 49
2.3.1 **The physical structure of the prokaryotic genome** 50
The traditional view of the bacterial 'chromosome' 50
Complications on the *E. coli* theme 50
Research Briefing 2.1: Supercoiled domains in the *Escherichia coli* nucleoid 52
2.3.2 **The genetic organization of the prokaryotic genome** 54
Operons are characteristic features of prokaryotic genomes 55
Prokaryotic genomes and the species concept 56
Box 2.3: Mechanisms for gene flow between prokaryotes 57
Speculation on the minimal genome content and the identity of distinctiveness genes 58

2.4 The Repetitive DNA Content of Genomes 59
2.4.1 **Tandemly repeated DNA** 59
Satellite DNA is found at centromeres and elsewhere in eukaryotic chromosomes 59
Minisatellites and microsatellites 60
2.4.2 **Interspersed genome-wide repeats** 60
Transposition via an RNA intermediate 61
DNA transposons 63

Study Aids 66

Chapter 3 Transcriptomes and Proteomes 69

3.1 Genome Expression in Outline 70
Box 3.1: Cross-references to Part 3 of *Genomes* 71

3.2 The RNA Content of the Cell 72
3.2.1 **Coding and non-coding RNA** 72
Box 3.2: Non-coding RNA specified by the human genome 74
3.2.2 **Synthesis of RNA** 74
Processing of precursor RNA 74
3.2.3 **The transcriptome** 78
Studies of the yeast transcriptome 78
The human transcriptome 80

3.3 The Protein Content of the Cell 80
3.3.1 **Protein structure** 80
The four levels of protein structure 80
Amino acid diversity underlies protein diversity 82
Box 3.3: Non-covalent bonds in proteins 82
3.3.2 **The link between the transcriptome and the proteome** 83
The genetic code specifies how an mRNA sequence is translated into a polypeptide 84
Research Briefing 3.1: Elucidation of the genetic code 85
The genetic code is not universal 86
Box 3.4: The origin and evolution of the genetic code 87
3.3.3 **The link between the proteome and the biochemistry of the cell** 88
The amino acid sequence of a protein determines its function 88
The multiplicity of protein function 89

Study Aids 90

PART 2 Studying Genomes 93

Chapter 4 Studying DNA 95

4.1 Enzymes for DNA Manipulation 97
98
4.1.1 DNA polymerases 98
The mode of action of a template-dependent DNA polymerase 98
The types of DNA polymerases used in research 100
4.1.2 Nucleases 102
Restriction endonucleases enable DNA molecules to be cut at defined positions 102
Examining the results of a restriction digest 103
4.1.3 DNA ligases 105
4.1.4 End-modification enzymes 107

4.2 DNA Cloning 108
4.2.1 Cloning vectors and the way they are used 109
Vectors based on *E. coli* plasmids 109
110
Cloning vectors based on *E. coli* bacteriophage genomes 112
Vectors for longer pieces of DNA 113
117
Cloning in organisms other than *E. coli* 117

4.3 The Polymerase Chain Reaction (PCR) 119
120
4.3.1 Carrying out a PCR 120
4.3.2 The applications of PCR 121

Study Aids 123

Chapter 5 Mapping Genomes 125

5.1 Genetic and Physical Maps 128

5.2 Genetic Mapping 128
5.2.1 Genes were the first markers to be used 129
5.2.2 DNA markers for genetic mapping 129
Restriction fragment length polymorphisms (RFLPs) 130
Simple sequence length polymorphisms (SSLPs) 130
Single nucleotide polymorphisms (SNPs) 130
131
133
5.2.3 Linkage analysis is the basis of genetic mapping 134
The principles of inheritance and the discovery of linkage 134
Partial linkage is explained by the behavior of chromosomes during meiosis 137
From partial linkage to genetic mapping 139
5.2.4 Linkage analysis with different types of organism 140
Linkage analysis when planned breeding experiments are possible 140
141
Gene mapping by human pedigree analysis 142
Genetic mapping in bacteria 144

5.3 Physical Mapping 145
5.3.1 Restriction mapping 145
The basic methodology for restriction mapping 147
The scale of restriction mapping is limited by the sizes of the restriction fragments 147
Direct examination of DNA molecules for restriction sites 150
5.3.2 Fluorescent *in situ* hybridization (FISH) 152
In situ hybridization with radioactive or fluorescent probes 152
FISH in action 153
5.3.3 Sequence tagged site (STS) mapping 153
Any unique DNA sequence can be used as an STS 154

Fragments of DNA for STS mapping 155
156
A clone library can also be used as the mapping reagent for STS analysis 159

Study Aids 161

Chapter 6 Sequencing Genomes 163

6.1 The Methodology for DNA Sequencing 164
6.1.1 Chain termination DNA sequencing 165
Chain termination sequencing in outline 165
165
166
Chain termination sequencing requires a single-stranded DNA template 168
The primer determines the region of the template DNA that will be sequenced 169
Thermal cycle sequencing offers an alternative to the traditional methodology 170
Fluorescent primers are the basis of automated sequence reading 170
170
6.1.2 Departures from conventional DNA sequencing 171

6.2 Assembly of a Contiguous DNA Sequence 172
6.2.1 Sequence assembly by the shotgun approach 173
The potential of the shotgun approach was proven by the *Haemophilus influenzae* sequence 173
6.2.2 Sequence assembly by the clone contig approach 176
Clone contigs can be built up by chromosome walking, but the method is laborious 176
Newer more rapid methods for clone contig assembly 178
6.2.3 Whole-genome shotgun sequencing 179
Key features of whole-genome shotgun sequencing 180

6.3 The Human Genome Projects 181
6.3.1 The mapping phase of the Human Genome Project 181
6.3.2 Sequencing the human genome 182
6.3.3 The future of the human genome projects 182

Study Aids 184

Chapter 7 Understanding a Genome Sequence 187

7.1 Locating the Genes in a Genome Sequence 188
7.1.1 Gene location by sequence inspection 188
The coding regions of genes are open reading frames 189
Simple ORF scans are less effective with higher eukaryotic DNA 189
Homology searches give an extra dimension to sequence inspection 191
7.1.2 Experimental techniques for gene location 191
Hybridization tests can determine if a fragment contains transcribed sequences 192
cDNA sequencing enables genes to be mapped within DNA fragments 192
Methods are available for precise mapping of the ends of transcripts 193
Exon–intron boundaries can also be located with precision 194

7.2 Determining the Functions of Individual Genes 195
7.2.1 Computer analysis of gene function 196
Homology reflects evolutionary relationships 196
Homology analysis can provide information on the function of an entire gene or of segments within it 196
Homology analysis in the yeast genome project 197
7.2.2 Assigning gene function by experimental analysis 198
Functional analysis by gene inactivation 198
Individual genes can be inactivated by homologous recombination 198
Gene inactivation without homologous recombination 199
200
Gene overexpression can also be used to assess function 200
202
7.2.3 More detailed studies of the activity of a protein coded by an unknown gene 204
Directed mutagenesis can be used to probe gene function in detail 204

205
Reporter genes and immunocytochemistry can be used to locate where and when genes are expressed 206
7.3 Global Studies of Genome Activity 207
7.3.1 Studying the transcriptome 207
The composition of a transcriptome can be assayed by SAGE 207
Using chip and microarray technology to study a transcriptome 207
7.3.2 Studying the proteome 208
Proteomics – methodology for characterizing the protein content of a cell 208
Identifying proteins that interact with one another 211
Protein interaction maps 211
7.4 Comparative Genomics 213
7.4.1 Comparative genomics as an aid to gene mapping 214
7.4.2 Comparative genomics in the study of human disease genes 215
Study Aids 217

PART 3 How Genomes Function 219

Chapter 8 Accessing the Genome 221

8.1 Inside the Nucleus 222
8.1.1 The internal architecture of the eukaryotic nucleus 222
Box 8.1: Analysing the prokaryotic genome 223
Technique Box 8.1: Fluorescence recovery after photobleaching (FRAP) 224
8.1.2 Chromatin domains 224
Functional domains are defined by insulators 226
Some functional domains contain locus control regions 227

8.2 Chromatin Modifications and Genome Expression 228
8.2.1 Activating the genome 228
Histone modifications determine chromatin structure 228
Nucleosome remodeling influences the expression of individual genes 229
Box 8.2: Chromatin remodeling by the SWI/SNF proteins 231
8.2.2 Silencing the genome 231
Histone deacetylation is one way of repressing gene expression 231
Research Briefing 8.1: Discovery of the mammalian DNA methyltransferases 232
Genome silencing by DNA methylation 234
Methylation is involved in imprinting and X inactivation 235
Study Aids 237

Chapter 9 Assembly of the Transcription Initiation Complex 239

9.1 The Importance of DNA-binding Proteins 241
9.1.1 Locating the positions of DNA-binding sites in a genome 242
Gel retardation identifies DNA fragments that bind to proteins 242
Protection assays pinpoint binding sites with greater accuracy 242
Modification interference identifies nucleotides central to protein binding 244
9.1.2 Purifying a DNA-binding protein 245
9.1.3 Studying the structures of proteins and DNA–protein complexes 246
X-ray crystallography has broad applications in structure determination 246
NMR gives detailed structural information for small proteins 248
9.1.4 The special features of DNA-binding proteins 248
The helix–turn–helix motif is present in prokaryotic and eukaryotic proteins 249
Box 9.1: bZIP binding motifs 250
Zinc fingers are common in eukaryotic proteins 250
Other DNA-binding motifs 251
Box 9.2: How sequences specific to bases are read from the major groove of B-conformation helix 252
9.1.5 The interaction between DNA and its binding proteins 252
Direct readout of the nucleotide sequence 252