

GENOMES **2**
SECOND
EDITION

T.A. BROWN

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



WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

NanoString EX2015.001

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.

© Bios Scientific Publishers Limited 2002

First published 1999
Second Edition 2002

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 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)

I. Genomes.

[DNLM: I. 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

NanoString EX2015.002

Contents

Abbreviations	xvii
Preface to the Second Edition	xxi
Preface to the First Edition	xxiii
An Introduction to Genomes	xxv
 PART I Genomes, Transcriptomes and Proteomes	 I
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	---

	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 <i>E. coli</i> theme	50
	Research Briefing 2.1: Supercoiled domains in the <i>Escherichia coli</i> 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 <i>Genomes</i>	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	

PART 2	Studying Genomes	93
Chapter 4	Studying DNA	95
4.1	Enzymes for DNA Manipulation	97
	Technical Note 4.1: DNA labeling	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 <i>E. coli</i> plasmids	109
	Technical Note 4.2: DNA purification	110
	Cloning vectors based on <i>E. coli</i> bacteriophage genomes	112
	Vectors for longer pieces of DNA	113
	Technical Note 4.3: Working with a clone library	117
	Cloning in organisms other than <i>E. coli</i>	117
4.3	The Polymerase Chain Reaction (PCR)	119
	Technical Note 4.4: Techniques for studying RNA	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
	Box 5.1: Why do SNPs have only two alleles?	131
	Technical Note 5.1: DNA microarrays and chips	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
	Box 5.2: Multipoint crosses	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 <i>in situ</i> hybridization (FISH)	152
	<i>In situ</i> 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
Research Briefing 5.1: The radiation hybrid map of the mouse genome	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
Technical Note 6.1: Polyacrylamide gel electrophoresis	165
Box 6.1: DNA polymerases for chain termination sequencing	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
Box 6.2: The chemical degradation sequencing method	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 <i>Haemophilus influenzae</i> 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
Box 7.1: The phenotypic effect of gene inactivation is sometimes difficult to discern	200
Gene overexpression can also be used to assess function	200
Research Briefing 7.1: Analysis of chromosome I of <i>Caenorhabditis elegans</i> by RNA interference	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

Technical Note 7.1: Site-directed mutagenesis	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: Accessing the prokaryotic genome	223
Technical Note 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 modification by the HMGN 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: RNA-binding motifs	250
Zinc fingers are common in eukaryotic proteins	250
Other DNA-binding motifs	251
Box 9.2: Can sequence specificity be predicted from the structure of a recognition helix?	252
9.1.5 The interaction between DNA and its binding proteins	252
Direct readout of the nucleotide sequence	252

The nucleotide sequence has a number of indirect effects on helix structure	253
Contacts between DNA and proteins	253
9.2 DNA-Protein Interactions During Transcription Initiation	254
9.2.1 RNA polymerases	254
Box 9.3: Mitochondrial and chloroplast RNA polymerases	255
9.2.2 Recognition sequences for transcription initiation	255
Bacterial RNA polymerases bind to promoter sequences	255
Eukaryotic promoters are more complex	256
9.2.3 Assembly of the transcription initiation complex	257
Transcription initiation in <i>E. coli</i>	257
Transcription initiation with RNA polymerase II	258
Box 9.4: Initiation of transcription in the archaea	258
Transcription initiation with RNA polymerases I and III	259
Research Briefing 9.1: Similarities between TFIID and the histone core octamer	260
9.3 Regulation of Transcription Initiation	261
9.3.1 Strategies for controlling transcription initiation in bacteria	261
Promoter structure determines the basal level of transcription initiation	261
Regulatory control over bacterial transcription initiation	262
Box 9.5: <i>Cis</i> and <i>trans</i>	263
9.3.2 Control of transcription initiation in eukaryotes	265
Activators of eukaryotic transcription initiation	265
Contacts between activators and the pre-initiation complex	266
Repressors of eukaryotic transcription initiation	266
Box 9.6: The modular structures of RNA polymerase II promoters	267
Controlling the activities of activators and repressors	268
Study Aids	271
Chapter 10 Synthesis and Processing of RNA	273
10.1 Synthesis and Processing of mRNA	274
10.1.1 Synthesis of bacterial mRNAs	274
The elongation phase of bacterial transcription	274
Termination of bacterial transcription	275
Control over the choice between elongation and termination	277
Research Briefing 10.1: The structure of the bacterial RNA polymerase	278
Box 10.1: Antitermination during the infection cycle of bacteriophage λ	280
10.1.2 Synthesis of eukaryotic mRNAs by RNA polymerase II	283
Capping of RNA polymerase II transcripts occurs immediately after initiation	283
Elongation of eukaryotic mRNAs	284
Termination of mRNA synthesis is combined with polyadenylation	286
10.1.3 Intron splicing	287
Conserved sequence motifs indicate the key sites in GU-AG introns	288
Outline of the splicing pathway for GU-AG introns	288
snRNAs and their associated proteins are the central components of the splicing apparatus	289
Alternative splicing is common in many eukaryotes	291
AU-AC introns are similar to GU-AG introns but require a different splicing apparatus	294
10.2 Synthesis and Processing of Non-coding RNAs	295
10.2.1 Transcript elongation and termination by RNA polymerases I and III	295
10.2.2 Cutting events involved in processing of bacterial and eukaryotic pre-rRNA and pre-tRNA	295
10.2.3 Introns in eukaryotic pre-rRNA and pre-tRNA	296
Eukaryotic pre-rRNA introns are self-splicing	296
Eukaryotic tRNA introns are variable but all are spliced by the same mechanism	298
10.3 Processing of Pre-RNA by Chemical Modification	298
10.3.1 Chemical modification of pre-rRNAs	298
Box 10.2: Other types of intron	298

10.3.2 RNA editing	303
10.4 Degradation of mRNAs	304
Box 10.3: More complex forms of RNA editing	305
10.4.1 Bacterial mRNAs are degraded in the 3'→5' direction	305
10.4.2 Eukaryotes have more diverse mechanisms for RNA degradation	306
10.5 Transport of RNA Within the Eukaryotic Cell	308
Study Aids	310
Chapter 11 Synthesis and Processing of the Proteome	313
11.1 The Role of tRNA in Protein Synthesis	314
11.1.1 Aminoacylation: the attachment of amino acids to tRNAs	315
All tRNAs have a similar structure	315
Aminoacyl-tRNA synthetases attach amino acids to tRNAs	316
11.1.2 Codon-anticodon interactions: the attachment of tRNAs to mRNA	317
11.2 The Role of the Ribosome in Protein Synthesis	319
11.2.1 Ribosome structure	321
Ultracentrifugation was used to measure the sizes of ribosomes and their components	321
Probing the fine structure of the ribosome	322
11.2.2 Initiation of translation	323
Initiation in bacteria requires an internal ribosome binding site	324
Initiation in eukaryotes is mediated by the cap structure and poly(A) tail	325
Initiation of eukaryotic translation without scanning	328
Regulation of translation initiation	328
11.2.3 Elongation of translation	328
Elongation in bacteria and eukaryotes	328
Frameshifting and other unusual events during elongation	330
Research Briefing 11.1: Peptidyl transferase is a ribozyme	332
11.2.4 Termination of translation	333
Box 11.1: Translation in the archaea	334
11.3 Post-translational Processing of Proteins	335
11.3.1 Protein folding	335
Not all proteins fold spontaneously in the test tube	335
In cells, folding is aided by molecular chaperones	337
11.3.2 Processing by proteolytic cleavage	339
Cleavage of the ends of polypeptides	339
Proteolytic processing of polypeptides	339
11.3.3 Processing by chemical modification	340
11.3.4 Inteins	342
11.4 Protein Degradation	343
Study Aids	346
Chapter 12 Regulation of Genome Activity	347
12.1 Transient Changes in Genome Activity	350
12.1.1 Signal transmission by import of the extracellular signaling compound	351
Lactoferrin is an extracellular signaling protein which acts as a transcription activator	352
Some imported signaling compounds directly influence the activity of pre-existing protein factors	352
Some imported signaling compounds influence genome activity indirectly	354
12.1.2 Signal transmission mediated by cell surface receptors	354
Signal transduction with one step between receptor and genome	356
Signal transduction with many steps between receptor and genome	358
Signal transduction via second messengers	359
12.2 Permanent and Semipermanent Changes in Genome Activity	
12.2.1 Genome rearrangements	---

Yeast mating types are determined by gene conversion events	360
Genome rearrangements are responsible for immunoglobulin and T-cell receptor diversities	361
12.2.2 Changes in chromatin structure	362
12.2.3 Genome regulation by feedback loops	363
Research Briefing 12.1: Unraveling a signal transduction pathway	364
12.3 Regulation of Genome Activity During Development	365
12.3.1 Sporulation in <i>Bacillus</i>	366
Sporulation involves coordinated activities in two distinct cell types	366
Special σ subunits control genome activity during sporulation	366
12.3.2 Vulva development in <i>Caenorhabditis elegans</i>	369
<i>C. elegans</i> is a model for multicellular eukaryotic development	369
Determination of cell fate during development of the <i>C. elegans</i> vulva	369
Research Briefing 12.2: The link between genome replication and sporulation in <i>Bacillus</i>	370
12.3.3 Development in <i>Drosophila melanogaster</i>	372
Maternal genes establish protein gradients in the <i>Drosophila</i> embryo	372
A cascade of gene expression converts positional information into a segmentation pattern	374
Box 12.1: The genetic basis of flower development	375
Segment identity is determined by the homeotic selector genes	375
Homeotic selector genes are universal features of higher eukaryotic development	376
Study Aids	378
PART 4 How Genomes Replicate and Evolve	381
Chapter 13 Genome Replication	383
13.1 The Topological Problem	384
13.1.1 Experimental proof for the Watson–Crick scheme for DNA replication	385
The Meselson–Stahl experiment	386
13.1.2 DNA topoisomerases provide a solution to the topological problem	388
13.1.3 Variations on the semiconservative theme	389
13.2 The Replication Process	391
13.2.1 Initiation of genome replication	391
Initiation at the <i>E. coli</i> origin of replication	391
Origins of replication in yeast have been clearly defined	392
Replication origins in higher eukaryotes have been less easy to identify	393
13.2.2 The elongation phase of replication	393
The DNA polymerases of bacteria and eukaryotes	395
Discontinuous strand synthesis and the priming problem	396
Events at the bacterial replication fork	397
The eukaryotic replication fork: variations on the bacterial theme	400
13.2.3 Termination of replication	401
Replication of the <i>E. coli</i> genome terminates within a defined region	402
Little is known about termination of replication in eukaryotes	403
Box 13.1: Genome replication in the archaea	403
13.2.4 Maintaining the ends of a linear DNA molecule	404
Telomeric DNA is synthesized by the telomerase enzyme	405
Telomere length is implicated in senescence and cancer	405
Box 13.2: Telomeres in <i>Drosophila</i>	408
13.3 Regulation of Eukaryotic Genome Replication	409
13.3.1 Coordination of genome replication and cell division	409
Establishment of the pre-replication complex enables genome replication to commence	409
Research Briefing 13.1: Replication of the yeast genome	410
Regulation of pre-RC assembly	412
13.3.2 Control within S phase	412
Early and late replication origins	412
Checkpoints within S phase	413
Study Aids	

Chapter 14 Mutation, Repair and Recombination	417
Box 14.1: Terminology for describing point mutations	419
14.1 Mutations	420
14.1.1 The causes of mutations	420
Errors in replication are a source of point mutations	420
Replication errors can also lead to insertion and deletion mutations	423
Mutations are also caused by chemical and physical mutagens	424
14.1.2 The effects of mutations	428
The effects of mutations on genomes	428
Technical Note 14.1: Mutation detection	429
The effects of mutations on multicellular organisms	431
The effects of mutations on microorganisms	432
14.1.3 Hypermutation and the possibility of programmed mutations	433
14.2 DNA Repair	434
14.2.1 Direct repair systems fill in nicks and correct some types of nucleotide modification	434
Research Briefing 14.1: Programmed mutations?	435
14.2.2 Excision repair	437
Base excision repairs many types of damaged nucleotide	437
Nucleotide excision repair is used to correct more extensive types of damage	437
14.2.3 Mismatch repair: correcting errors of replication	440
14.2.4 Repair of double-stranded DNA breaks	441
14.2.5 Bypassing DNA damage during genome replication	442
14.2.6 Defects in DNA repair underlie human diseases, including cancers	444
14.3 Recombination	444
14.3.1 Homologous recombination	444
The Holliday model for homologous recombination	445
Proteins involved in homologous recombination in <i>E. coli</i>	446
The double-strand break model for recombination in yeast	447
Box 14.2: The RecE and RecF recombination pathways of <i>Escherichia coli</i>	447
14.3.2 Site-specific recombination	448
Integration of λ DNA into the <i>E. coli</i> genome involves site-specific recombination	448
14.3.3 Transposition	450
Box 14.3: DNA methylation and transposition	450
Replicative and conservative transposition of DNA transposons	451
Transposition of retroelements	454
Study Aids	456
Chapter 15 How Genomes Evolve	459
15.1 Genomes: the First 10 Billion Years	460
15.1.1 The origins of genomes	460
The first biochemical systems were centered on RNA	461
The first DNA genomes	462
How unique is life?	463
15.2 Acquisition of New Genes	465
15.2.1 Acquisition of new genes by gene duplication	465
Whole-genome duplications can result in sudden expansions in gene number	465
Research Briefing 15.1: Segmental duplications in the yeast and human genomes	468
Duplications of individual genes and groups of genes have occurred frequently in the past	470
Box 15.1: Gene duplication and genetic redundancy	471
Genome evolution also involves rearrangement of existing genes	472
15.2.2 Acquisition of new genes from other species	473
15.3 Non-coding DNA and Genome Evolution	476
15.3.1 Transposable elements and genome evolution	476
Box 15.2: The origin of a microsatellite	476
15.3.2 The origins of introns	477

'Introns early' and 'introns late': two competing hypotheses	477
The current evidence disproves neither hypothesis	478
Box 15.3 The role of non-coding DNA	479
15.4 The Human Genome: the Last 5 Million Years	479
Study Aids	482
Chapter 16 Molecular Phylogenetics	483
16.1 The Origins of Molecular Phylogenetics	484
Box 16.1: Phenetics and cladistics	486
16.2 The Reconstruction of DNA-based Phylogenetic Trees	487
16.2.1 The key features of DNA-based phylogenetic trees	487
Gene trees are not the same as species trees	488
Box 16.2: Terminology for molecular phylogenetics	488
16.2.2 Tree reconstruction	489
Sequence alignment is the essential preliminary to tree reconstruction	490
Converting alignment data into a phylogenetic tree	491
Technical Note 16.1: Phylogenetic analysis	491
Assessing the accuracy of a reconstructed tree	493
Molecular clocks enable the time of divergence of ancestral sequences to be estimated	493
16.3 The Applications of Molecular Phylogenetics	494
16.3.1 Examples of the use of phylogenetic trees	494
DNA phylogenetics has clarified the evolutionary relationships between humans and other primates	494
The origins of AIDS	495
16.3.2 Molecular phylogenetics as a tool in the study of human prehistory	496
Intraspecific studies require highly variable genetic loci	496
The origins of modern humans – out of Africa or not?	496
Box 16.3: Genes in populations	497
The patterns of more recent migrations into Europe are also controversial	498
Research Briefing 16.1: Neandertal DNA	499
Prehistoric human migrations into the New World	502
Study Aids	504
Appendix	507
Keeping up to Date	507
Keeping up to Date by Reading the Literature	507
Keeping up to Date using the Internet	507
Glossary	511
Index	551

4

Studying DNA

Chapter Contents

4.1	<i>Enzymes for DNA Manipulation</i>	97
4.1.1	DNA polymerases	98
4.1.2	Nucleases	102
4.1.3	DNA ligases	105
4.1.4	End-modification enzymes	107
4.2	<i>DNA Cloning</i>	108
4.2.1	Cloning vectors and the way they are used	109
4.3	<i>The Polymerase Chain Reaction (PCR)</i>	119
4.3.1	Carrying out a PCR	120
4.3.2	The applications of PCR	121

Learning outcomes

When you have read Chapter 4, you should be able to:

- Give outline descriptions of the events involved in DNA cloning and the polymerase chain reaction (PCR), and state the applications and limitations of these techniques
- Describe the activities and main applications of the different types of enzyme used in recombinant DNA research
- Identify the important features of DNA polymerases and distinguish between the various DNA polymerases used in genomics research
- Describe, with examples, the way that restriction endonucleases cut DNA and explain how the results of a restriction digest are examined
- Distinguish between blunt- and sticky-end ligation and explain how the efficiency of blunt-end ligation can be increased
- Give details of the key features of plasmid cloning vectors and describe how these vectors are used in cloning experiments, using pBR322 and pUC8 as examples
- Describe how bacteriophage λ vectors are used to clone DNA
- Give examples of vectors used to clone long pieces of DNA, and evaluate the strengths and weaknesses of each type
- Summarize how DNA is cloned in yeast, animals and plants
- Describe how a PCR is performed, paying particular attention to the importance of the primers and the temperatures used during the thermal cycling

THE TOOLKIT OF TECHNIQUES used by molecular biologists to study DNA molecules was assembled during the 1970s and 1980s. Before then, the only way in which individual genes could be studied was by classical genetics, using the procedures that we will examine in Chapter 5. Classical genetics is a powerful approach to gene analysis and many of the fundamental discoveries in molecular biology were made in this way. The **operon theory** proposed by Jacob and Monod in 1961 (Section 9.3.1), which describes how the expression of some bacterial genes is regulated, was perhaps the most heroic achievement of this era of genetics. But the classical approach is limited because it does not involve the direct examination of genes, information on gene structure and activity being inferred from the biological characteristics of the organism being studied. By the late 1960s these indirect methods had become insufficient for answering the more detailed questions that molecular biologists had begun to ask about the expression pathways of individual genes. These questions could only be addressed by examining directly the segments of DNA containing the genes of interest.

This was not possible using the current technology, so a new set of techniques had to be invented.

The development of these new techniques was stimulated by breakthroughs in biochemical research which, in the early 1970s, provided molecular biologists with enzymes that could be used to manipulate DNA molecules in the test tube. These enzymes occur naturally in living cells and are involved in processes such as DNA replication, repair and recombination (see Chapters 13 and 14). In order to determine the functions of these enzymes, many of them were purified and the reactions that they catalyze studied in the test tube. Molecular biologists then adopted the pure enzymes as tools for manipulating DNA molecules in pre-determined ways, using them to make copies of DNA molecules, to cut DNA molecules into shorter fragments, and to join them together again in combinations that do not exist in nature (Figure 4.1). These manipulations, which are described in Section 4.1, form the basis of **recombinant DNA technology**, in which new or 'recombinant' DNA molecules are constructed in the test tube from pieces of naturally occurring chromosomes and plasmids. Recomb

DNA methodology led to the development of **DNA or gene cloning**, in which short DNA fragments, possibly containing a single gene, are inserted into a plasmid or virus chromosome and then replicated in a bacterial or eukaryotic host (Figure 4.2). We will examine exactly how gene cloning is performed, and the reasons why this technique resulted in a revolution in molecular biology, in Section 4.2.

Gene cloning was well established by the end of the 1970s. The next major technical breakthrough came some 5 years later when the **polymerase chain reaction (PCR)** was invented (Mullis, 1990). PCR is not a complicated technique – all that it achieves is the repeated copying of a short segment of a DNA molecule (Figure 4.3) – but it has become immensely important in many areas of bio-

logical research, not least the study of genomes. PCR is covered in detail in Section 4.3.

4.1 Enzymes for DNA Manipulation

Recombinant DNA technology was one of the main factors that contributed to the rapid advance in knowledge concerning gene expression that occurred during the 1970s and 1980s. The basis of recombinant DNA technology is the ability to manipulate DNA molecules in the test tube. This, in turn, depends on the availability of purified enzymes whose activities are known and can be controlled, and which can therefore be used to make specified changes to the DNA molecules that are being manipulated. The enzymes available to the molecular biologist fall into four broad categories:

- **DNA polymerases** (Section 4.1.1), which are enzymes that synthesize new polynucleotides complementary to an existing DNA or RNA template (Figure 4.4A);
- **Nucleases** (Section 4.1.2), which degrade DNA molecules by breaking the phosphodiester bonds that link one nucleotide to the next (Figure 4.4B);
- **Ligases** (Section 4.1.3), which join DNA molecules together by synthesizing phosphodiester bonds between nucleotides at the ends of two different molecules, or at the two ends of a single molecule (Figure 4.4C);
- **End-modification enzymes** (Section 4.1.4), which make changes to the ends of DNA molecules, adding an important dimension to the design of ligation experiments, and providing one means of labeling DNA molecules with radioactive and other markers (Technical Note 4.1).

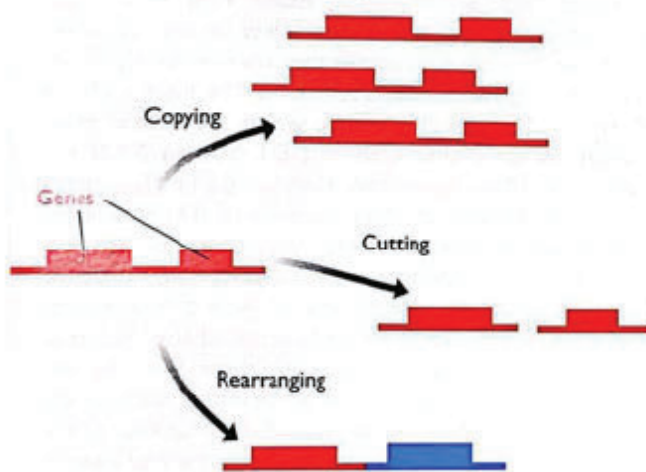


Figure 4.1 Examples of the manipulations that can be carried out with DNA molecules.

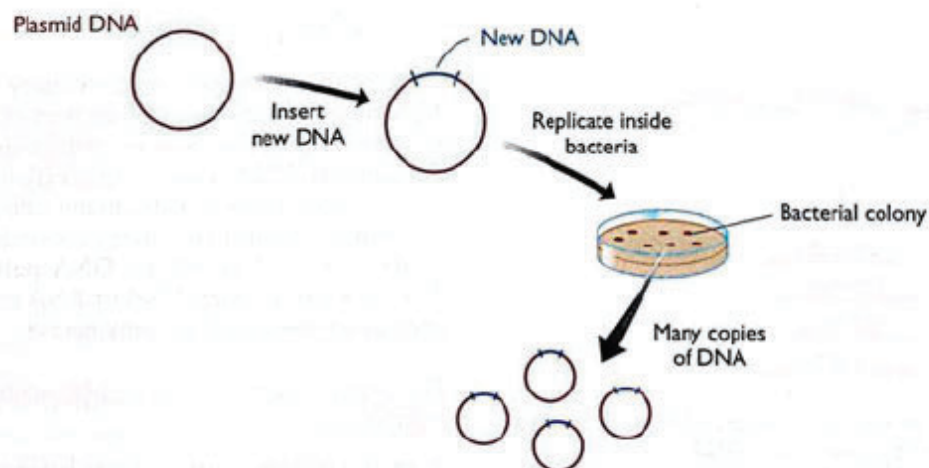
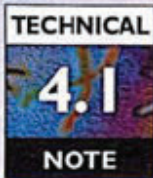


Figure 4.2 DNA cloning.

In this example, the fragment of DNA to be cloned is inserted into a plasmid vector which is subsequently replicated inside a bacterial host.



DNA labeling

Attachment of radioactive, fluorescent or other types of marker to DNA molecules.

DNA labeling is a central part of many molecular biology procedures, including Southern hybridization (Section 4.1.2), fluorescent *in situ* hybridization (FISH; Section 5.3.2) and DNA sequencing (Section 6.1). It enables the location of a particular DNA molecule – on a nitrocellulose or nylon membrane, in a chromosome or in a gel – to be determined by detecting the signal emitted by the marker. Labeled RNA molecules are also used in some applications (Technical Note 4.4, page 120).

Radioactive markers are frequently used for labeling DNA molecules. Nucleotides can be synthesized in which one of the phosphorus atoms is replaced with ^{32}P or ^{33}P , one of the oxygen atoms in the phosphate group is replaced with ^{35}S , or one or more of the hydrogen atoms is replaced with ^3H (see Figure 1.6, page 11). Radioactive nucleotides still act as substrates for DNA polymerases and so are incorporated into a DNA molecule by any strand-synthesis reaction catalyzed by a DNA polymerase. Labeled nucleotides or individual phosphate groups can also be attached to one or both ends of a DNA molecule by the reactions catalyzed by T4 polynucleotide kinase or terminal deoxynucleotidyl transferase (Section 4.1.4). The radioactive signal can be detected by scintillation counting, but for most molecular biology applications positional information is needed, so detection is by exposure of an X-ray-sensitive film (**autoradiography**) or a radiation-

sensitive phosphorescent screen (**phosphorimaging**). The choice between the various radioactive labels depends on the requirements of the procedure. High sensitivity is possible with ^{32}P because this isotope has a high emission energy, but sensitivity is accompanied by low resolution because of signal scattering. Low-emission isotopes such as ^{33}S or ^3H give less sensitivity but greater resolution.

Health and environmental issues have meant that radioactive markers have become less popular in recent years and for many procedures they are now largely superseded by non-radioactive alternatives. The most useful of these are fluorescent markers, which are central components of techniques such as FISH (Section 5.3.2) and automated DNA sequencing (Section 6.1.1). Fluorescent labels with various emission wavelengths (i.e. of different colors) are incorporated into nucleotides or attached directly to DNA molecules, and are detected with a suitable film, by fluorescence microscopy, or with a fluorescence detector. Other types of non-radioactive labeling make use of chemiluminescent emissions, but these have the disadvantage that the signal is not generated directly by the label, but instead must be 'developed' by treatment of the labeled molecule with chemicals. A popular method involves labeling the DNA with the enzyme alkaline phosphatase, which is detected by applying diioxetane, which the enzyme dephosphorylates to produce the chemiluminescence.

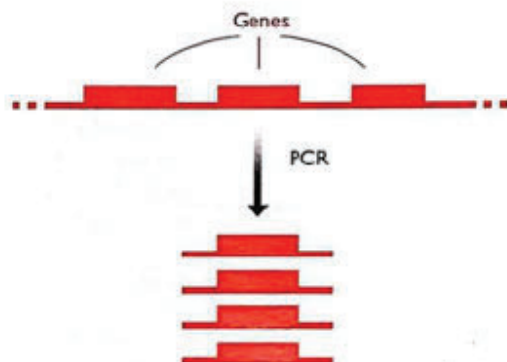


Figure 4.3 The polymerase chain reaction (PCR) is used to make copies of a selected segment of a DNA molecule.

In this example, a single gene is copied.

4.1.1 DNA polymerases

Many of the techniques used to study DNA depend on the synthesis of copies of all or part of existing DNA or RNA molecules. This is an essential requirement for PCR (Section 4.3), DNA sequencing (Section 6.1), DNA labeling (Technical Note 4.1) and many other procedures that are central to molecular biology research. An enzyme that synthesizes DNA is called a **DNA polymerase** and one that copies an existing DNA or RNA molecule is called a **template-dependent DNA polymerase**.

The mode of action of a template-dependent DNA polymerase

A template-dependent DNA polymerase makes a new DNA polynucleotide whose sequence is dictated, via the base-pairing rules, by the sequence of nucleotides in the DNA molecule that is being copied (Figure 4.5). The mode of action is very similar to that of a template-dependent RNA polymerase (Section 3.2.2), the new polynucleotide...

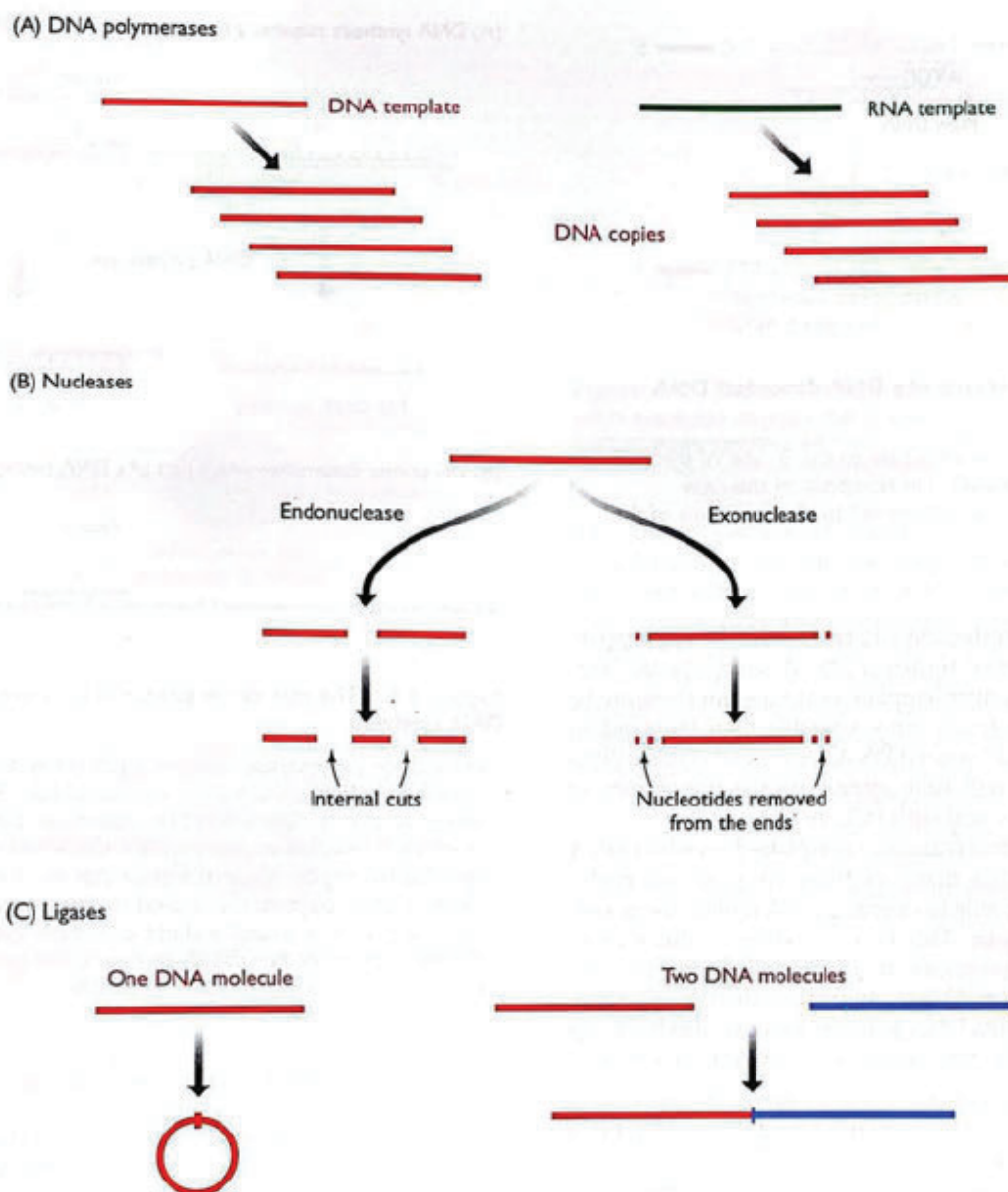


Figure 4.4 The activities of (A) DNA polymerases, (B) nucleases, and (C) ligases.

In (A), the activity of a DNA-dependent DNA polymerase is shown on the left and that of an RNA-dependent DNA polymerase on the right. In (B), the activities of endonucleases and exonucleases are shown. In (C) the red DNA molecule is ligated to itself on the left, and to a second molecule on the right.

being synthesized in the 5'→3' direction: DNA polymerases that make DNA in the other direction are unknown in nature.

One important difference between template-dependent DNA synthesis and the equivalent process for synthesis of RNA is that a DNA polymerase is unable to use an entirely single-stranded molecule as the template. In order to initiate DNA synthesis there must be a short double-stranded region to provide a 3' end onto which the enzyme will add new nucleotides (Figure 4.6A). The way in which this requirement is met in living cells when

the genome is replicated is described in Chapter 13. In the test tube, a DNA copying reaction is initiated by attaching to the template a short synthetic **oligonucleotide**, usually about 20 nucleotides in length, which acts as a **primer** for DNA synthesis. At first glance, the need for a primer might appear to be an undesired complication in the use of DNA polymerases in recombinant DNA technology, but nothing could be further from the truth. Because annealing of the primer to the template depends on complementary base-pairing, the position within the template molecule at which DNA copying is initiated

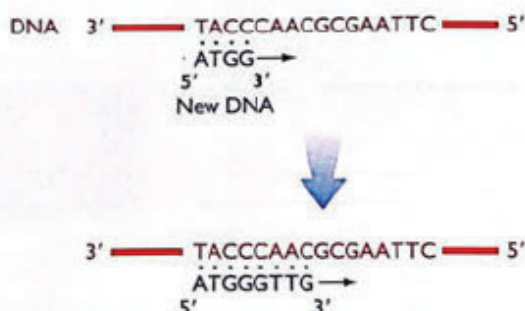


Figure 4.5 The activity of a DNA-dependent DNA polymerase.

New nucleotides are added on to the 3' end of the growing polynucleotide, the sequence of this new polynucleotide being determined by the sequence of the template DNA.

be specified by synthesizing a primer with the appropriate nucleotide sequence (Figure 4.6B). A short specific segment of a much longer template molecule can therefore be copied, which is much more valuable than the random copying that would occur if DNA synthesis did not need to be primed. You will fully appreciate the importance of priming when we deal with PCR in Section 4.3.

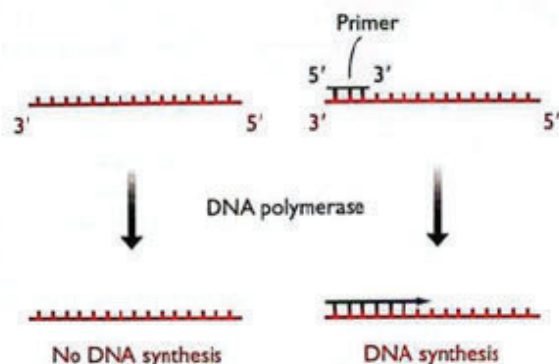
A second general feature of template-dependent DNA polymerases is that many of these enzymes are multifunctional, being able to degrade DNA molecules as well as synthesize them. This is a reflection of the way in which DNA polymerases act in the cell during genome replication (Section 13.2.2). As well as their 5'→3' DNA synthesis capability, DNA polymerases can also have one or both of the following exonuclease activities (Figure 4.7):

- A 3'→5' exonuclease activity enables the enzyme to remove nucleotides from the 3' end of the strand that it has just synthesized. This is called the **proof-reading** activity because it allows the polymerase to correct errors by removing a nucleotide that has been inserted incorrectly.
- A 5'→3' exonuclease activity is less common, but is possessed by some DNA polymerases whose natural function in genome replication requires that they must be able to remove at least part of a polynucleotide that is already attached to the template strand that the polymerase is copying.

The types of DNA polymerases used in research

Several of the template-dependent DNA polymerases that are used in molecular biology research (Table 4.1) are versions of the *Escherichia coli* DNA polymerase I enzyme, which plays a central role in replication of this bacterium's genome (Section 13.2.2). This enzyme, sometimes called the **Kornberg polymerase**, after its discoverer Arthur Kornberg (Kornberg, 1960), has both the 3'→5' and 5'→3' exonuclease activities, which limits

(A) DNA synthesis requires a primer



(B) The primer determines which part of a DNA molecule is copied



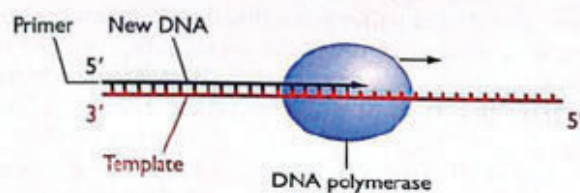
Figure 4.6 The role of the primer in template-dependent DNA synthesis.

(A) A DNA polymerase requires a primer in order to initiate the synthesis of a new polynucleotide. (B) The sequence of this oligonucleotide determines the position at which it attaches to the template DNA and hence specifies the region of the template that will be copied. When a DNA polymerase is used to make new DNA *in vitro*, the primer is usually a short oligonucleotide made by chemical synthesis. For details of how DNA synthesis is primed *in vivo*, see Figure 13.12, page 396.

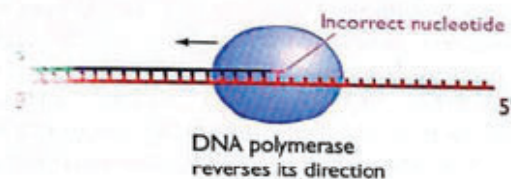
its usefulness in DNA manipulation. Its main application is in DNA labeling, as described in Technical Note 4.1.

Of the two exonuclease activities, it is the 5'→3' version that causes most problems when a DNA polymerase is used to manipulate molecules in the test tube. This is because an enzyme that possesses this activity is able to remove nucleotides from the 5' ends of polynucleotides that have just been synthesized (Figure 4.8). It is unlikely that the polynucleotides will be completely degraded, because the polymerase function is usually much more active than the exonuclease, but some techniques will not work if the 5' ends of the new polynucleotides are shortened in any way. In particular, DNA sequencing is based on synthesis of new polynucleotides, all of which share exactly the same 5' end, marked by the primer used to initiate the sequencing reactions. If any nibbling of the 5' ends occurs, then it is impossible to determine the correct DNA sequence. When DNA sequencing was first introduced in the late 1970s, it made use of a modified version of the Kornberg enzyme called the **Klenow polymerase**. The Klenow polymerase was initially prepared by cutting the natural *E. coli* DNA polymerase I enzyme into two segments with a protease. One of these segments retained the polymerase and 3'→5' exonuclease activities,

(A) 5'→3' DNA synthesis



(B) 3'→5' exonuclease activity



(C) 5'→3' exonuclease activity

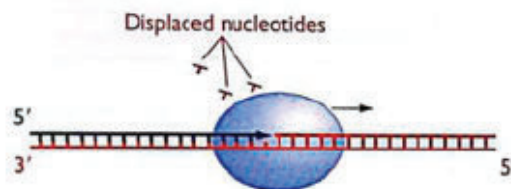


Figure 4.7 The DNA synthesis and exonuclease activities of DNA polymerases.

All DNA polymerases can make DNA and many also have one or both of the exonuclease activities.

lacked the 5'→3' exonuclease of the untreated enzyme. The enzyme is still often called the Klenow *fragment* in memory of this old method of preparation, but nowadays it is almost always prepared from *E. coli* cells whose polymerase gene has been engineered so that the resulting enzyme has the desired properties. But in fact the Klenow polymerase is now rarely used in sequencing and has its major application in DNA labeling (see Technical Note

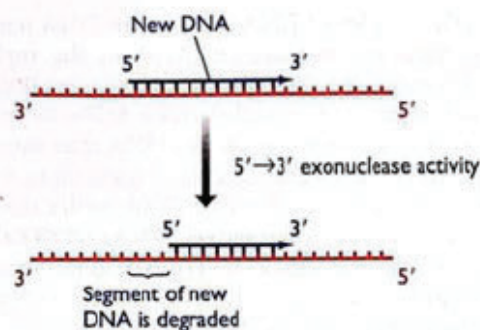


Figure 4.8 The 5'→3' exonuclease activity of a DNA polymerase can degrade the 5' end of a polynucleotide that has just been synthesized.

4.1). This is because an enzyme called **Sequenase** (see Table 4.1), which has superior properties as far as sequencing is concerned, was developed in the 1980s. We will return to the features of Sequenase, and why they make the enzyme ideal for sequencing, in Box 6.1 (page 166).

The *E. coli* DNA polymerase I enzyme has an optimum reaction temperature of 37 °C, this being the usual temperature of the natural environment of the bacterium, inside the intestines of mammals such as humans. Test-tube reactions with either the Kornberg or Klenow polymerases, and with Sequenase, are therefore incubated at 37 °C, and terminated by raising the temperature to 75 °C or above, which causes the protein to unfold or **denature**, destroying its activity. This regimen is perfectly adequate for most molecular biology techniques but, for reasons that will become clear in Section 4.3, PCR requires a **thermostable** DNA polymerase – one that is able to function at temperatures much higher than 37 °C. Suitable enzymes can be obtained from bacteria such as *Thermus aquaticus*, which live in hot springs at temperatures up to 95 °C, and whose DNA polymerase I enzyme has an optimum working temperature of 72 °C. The biochemical basis of protein thermostability is not fully understood, but probably centers on structural features that reduce the amount of protein unfolding that occurs at elevated temperatures.

One additional type of DNA polymerase is important in molecular biology research. This is **reverse transcriptase**, which is an **RNA-dependent DNA polymerase** and so

Table 4.1 Features of template-dependent DNA polymerases used in molecular biology research

Polymerase	Description	Main use	Cross reference
DNA polymerase I	Unmodified <i>E. coli</i> enzyme	DNA labeling	Technical Note 4.1 (page 98)
Klenow polymerase	Modified version of <i>E. coli</i> DNA polymerase I	DNA labeling	Technical Note 4.1 (page 98)
Sequenase	Modified version of phage T7 DNA polymerase I	DNA sequencing	Box 6.1 (page 166)
Taq polymerase	<i>Thermus aquaticus</i> DNA polymerase I	PCR	Section 4.3
Reverse transcriptase	RNA-dependent DNA polymerase, obtained from various retroviruses	cDNA synthesis	Section 5.3.3

makes DNA copies of RNA rather than DNA templates. Reverse transcriptases are involved in the replication cycles of retroviruses (Section 2.4.2), including the human immunodeficiency viruses that cause AIDS, these having RNA genomes that are copied into DNA after infection of the host. In the test tube, a reverse transcriptase can be used to make DNA copies of mRNA molecules. These copies are called **complementary DNAs (cDNAs)**. Their synthesis is important in some types of gene cloning and in techniques used to map the regions of a genome that specify particular mRNAs (Section 7.1.2).

4.1.2 Nucleases

A range of nucleases have found applications in recombinant DNA technology (Table 4.2). Some nucleases have a broad range of activities but most are either **exonucleases**, removing nucleotides from the ends of DNA and/or RNA molecules, or **endonucleases**, making cuts at internal phosphodiester bonds. Some nucleases are specific for DNA and some for RNA; some work only on double-stranded DNA and others only on single-stranded DNA, and some are not fussy what they work on. We will encounter various examples of nucleases in later chapters when we deal with the techniques in which they used. Only one type of nuclease will be considered in detail here: the **restriction endonucleases**, which play a central role in all aspects of recombinant DNA technology.

Restriction endonucleases enable DNA molecules to be cut at defined positions

A restriction endonuclease is an enzyme that binds to a DNA molecule at a specific sequence and makes a double-stranded cut at or near that sequence. Because of the sequence specificity, the positions of cuts within a DNA molecule can be predicted, assuming that the DNA sequence is known, enabling defined segments to be excised from a larger molecule. This ability underlies gene cloning and all other aspects of recombinant DNA technology in which DNA fragments of known sequence are required.

There are three types of restriction endonuclease. With Types I and III there is no strict control over the position of the cut relative to the specific sequence in the DNA molecule that is recognized by the enzyme. These

enzymes are therefore less useful because the sequences of the resulting fragments are not precisely known. Type II enzymes do not suffer from this disadvantage because the cut is always at the same place, either within the recognition sequence or very close to it (Figure 4.9). For example, the Type II enzyme called *EcoRI* (isolated from *E. coli*) cuts DNA only at the hexanucleotide 5'-GAATTC-3'. Digestion of DNA with a Type II enzyme therefore gives a reproducible set of fragments whose sequences are predictable if the sequence of the target DNA molecule is known. Over 2500 Type II enzymes have been isolated and more than 300 are available for use in the laboratory (Brown, 1998). Many enzymes have hexanucleotide target sites, but others recognize shorter or longer sequences (Table 4.3). There are also examples of enzymes with degenerate recognition sequences, meaning that they cut DNA at any of a family of related sites. *HinfI* (from *Haemophilus influenzae*), for example, recognizes 5'-GANTC-3', where 'N' is any nucleotide, and so

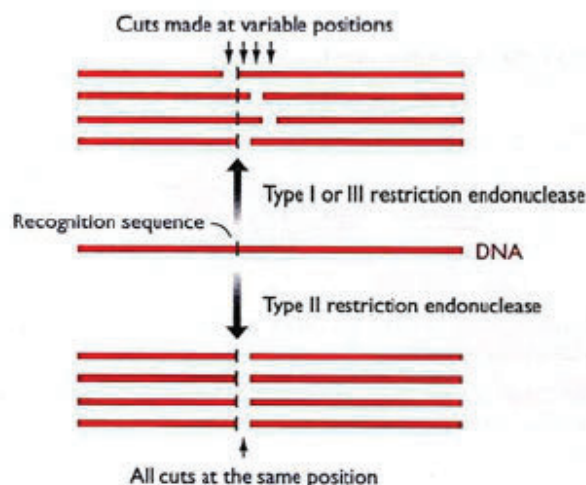


Figure 4.9 Cuts produced by restriction endonucleases.

In the top part of the diagram, the DNA is cut by a Type I or Type III restriction endonuclease. The cuts are made in slightly different positions relative to the recognition sequence, so the resulting fragments have different lengths. In the lower part, a Type II enzyme is used. Each molecule is cut at exactly the same position to give exactly the same pair of fragments.

Table 4.2 Features of important nucleases used in molecular biology research

Nuclease	Description	Main use	Cross reference
Restriction endonucleases	Sequence-specific DNA endonucleases, from many sources	Many applications	Section 4.1.2
SI nuclease	Endonuclease specific for single-stranded DNA and RNA, from the fungus <i>Aspergillus oryzae</i>	Transcript mapping	Section 7.1.2
Deoxyribonuclease I	Endonuclease specific for double-stranded DNA and RNA, from <i>Escherichia coli</i>	Nuclease footprinting	Section 9.1.1

Table 4.3 Some examples of restriction endonucleases

Enzyme	Recognition sequence	Type of ends	End sequences
<i>AluI</i>	5'-AGCT-3' 3'-TCGA-5'	Blunt	5'-AG CT-3' 3'-TC GA-5'
<i>Sau3AI</i>	5'-GATC-3' 3'-CTAG-5'	Sticky, 5' overhang	5'- GATC-3' 3'-CTAG -5'
<i>Hinfi</i>	5'-GANTC-3' 3'-CTNAG-5'	Sticky, 5' overhang	5'-G ANT C-3' 3'-CTNA G-5'
<i>BamHI</i>	5'-GGATCC-3' 3'-CCTAGG-5'	Sticky, 5' overhang	5'-G GATCC-3' 3'-CCTAG G-5'
<i>BsrBI</i>	5'-CCGCTC-3' 3'-GGCGAG-5'	Blunt	5'- NNNCCGCTC-3' 3'- NNNGGCGAG-5'
<i>EcoRI</i>	5'-GAATTC-3' 3'-CTTAAG-5'	Sticky, 5' overhang	5'-G AATTC-3' 3'-CTTAA G-5'
<i>PstI</i>	5'-CTGCAG-3' 3'-GACGTC-5'	Sticky, 3' overhang	5'-CTGCA G-3' 3'-G ACGTC-5'
<i>NotI</i>	5'-GCGGCCGC-3' 3'-CGCCGGCG-5'	Sticky, 5' overhang	5'-GC GGCCGC-3' 3'-CGCCGG CG-5'
<i>BglII</i>	5'-GCCNNNNNGGC-3' 3'-CGGNNNNNCCG-5'	Sticky, 3' overhang	5'-GCCNNNN NGGC-3' 3'-CGGN NNNNCCG-5'

N = any nucleotide.

Note that most, but not all, recognition sequences have inverted symmetry; when read in the 5'→3' direction, the sequence is the same in both strands.

cuts at 5'-GAATC-3', 5'-GATTC-3', 5'-GAGTC-3' and 5'-GACTC-3'. Most enzymes cut within the recognition sequence, but a few, such as a *BsrBI*, cut at a specified position outside of this sequence.

Restriction enzymes cut DNA in two different ways. Many make a simple double-stranded cut, giving a **blunt** or **flush end**; others cut the two DNA strands at different positions, usually two or four nucleotides apart, so that the resulting DNA fragments have short single-stranded overhangs at each end. These are called **sticky** or **cohesive ends** because base-pairing between them can stick the DNA molecule back together again (Figure 4.10A). Some sticky-end cutters give 5' overhangs (e.g. *Sau3AI*, *Hinfi*) whereas others leave 3' overhangs (e.g. *PstI*) (Figure 4.10B). One feature that is particularly important in recombinant DNA technology is that some pairs of restriction enzymes have different recognition sequences but give the same sticky ends, examples being *Sau3AI* and *BamHI*, which both give a 5'-GATC-3' sticky end even though *Sau3AI* has a 4-bp recognition sequence and *BamHI* recognizes a 6-bp sequence (Figure 4.10C).

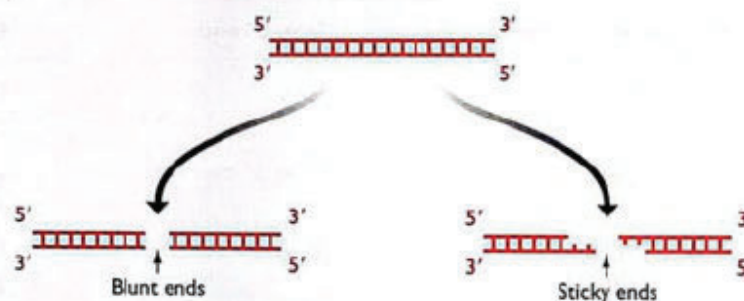
Examining the results of a restriction digest

After treatment with a restriction endonuclease, the resulting DNA fragments can be examined by agarose gel electrophoresis (see Technical Note 2.1, page 37) to

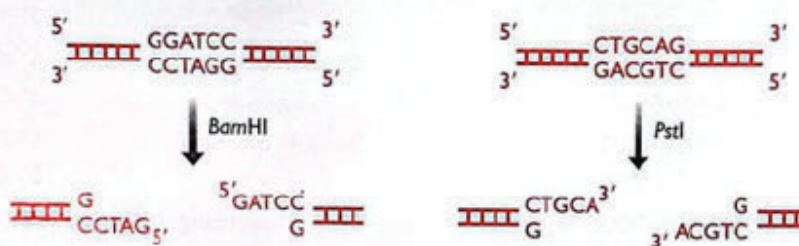
determine their sizes. Depending on the concentration of agarose in the gel, fragments between 100 bp and 50 kb can be separated into sharp bands after electrophoresis (Figure 4.11). Fragments less than 150 bp can be separated in a 4% or 5% agarose gel, making it possible to distinguish bands representing molecules that differ in size by just a single nucleotide. With larger fragments, however, it is not always possible to separate molecules of similar size, even in gels of lower agarose concentration. If the starting DNA is long, and so gives rise to many fragments after digestion with a restriction enzyme, then the gel may simply show a smear of DNA because there are fragments of every possible length that all merge together. This is the usual result when genomic DNA is restricted.

If the sequence of the starting DNA is known then the sequences, and hence the sizes, of the fragments resulting from treatment with a particular restriction enzyme can be predicted. The band for a desired fragment (for example, one containing a gene) can then be identified, cut out of the gel, and the DNA purified. Even if its size is unknown, a fragment containing a gene or another segment of DNA of interest can be identified by the technique called **Southern hybridization**, providing that some of the sequence within the fragment is known. The first step is to transfer the restriction fragments from the agarose gel to a nitrocellulose or

(A) Blunt and sticky ends



(B) 5' and 3' overhangs



(C) The same sticky end produced by different enzymes

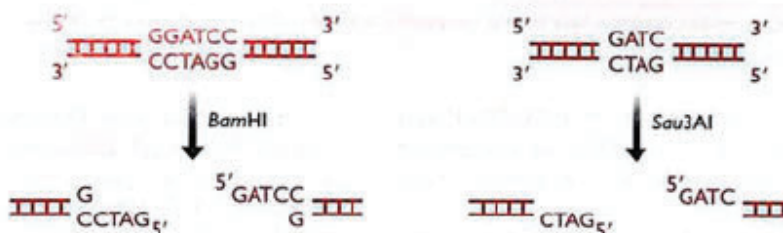


Figure 4.10 The results of digestion of DNA with different restriction endonucleases.

(A) Blunt ends and sticky ends. (B) Different types of sticky end: the 5' overhangs produced by *Bam*HI and the 3' overhangs produced by *Pst*I. (C) The same sticky end produced by two different restriction endonucleases: a 5' overhang with the sequence 5'–GATC–3' is produced by both *Bam*HI (recognizes 5'–GGATCC–3') and *Sau*3AI (recognizes 5'–GATC–3').

nylon membrane. This is done by placing the membrane on the gel and allowing buffer to soak through, taking the DNA from the gel to the membrane, where it becomes bound (Figure 4.12A). This process results in the DNA bands becoming immobilized in the same relative positions on the surface of the membrane.

The next step is to prepare a **hybridization probe**, which is a labeled DNA molecule whose sequence is complementary to the target DNA that we wish to detect. The probe could, for example, be a synthetic oligonucleotide whose sequence matches part of an interesting gene. Because the probe and target DNAs are complementary, they can base-pair or **hybridize**, the position of the hybridized probe on the membrane being identified by

detecting the signal given out by a label attached to the probe. To carry out the hybridization, the membrane is placed in a glass bottle with the labeled probe and some buffer, and the bottle gently rotated for several hours so that the probe has plenty of opportunity to hybridize to its target DNA. The membrane is then washed to remove any probe that has not become hybridized, and the signal from the label is detected (see Technical Note 4.1, page 98). In the example shown in Figure 4.12B the probe is radioactively labeled and the signal is detected by **autoradiography**. The band that is seen on the autoradiograph is the one that corresponds to the restriction fragment that hybridizes to the probe and which therefore contains gene that we are searching for.

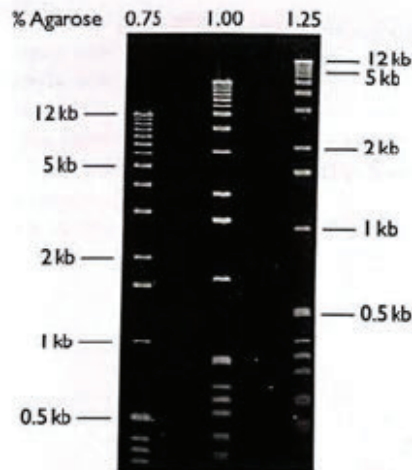


Figure 4.11 Separation of DNA molecules by agarose gel electrophoresis.

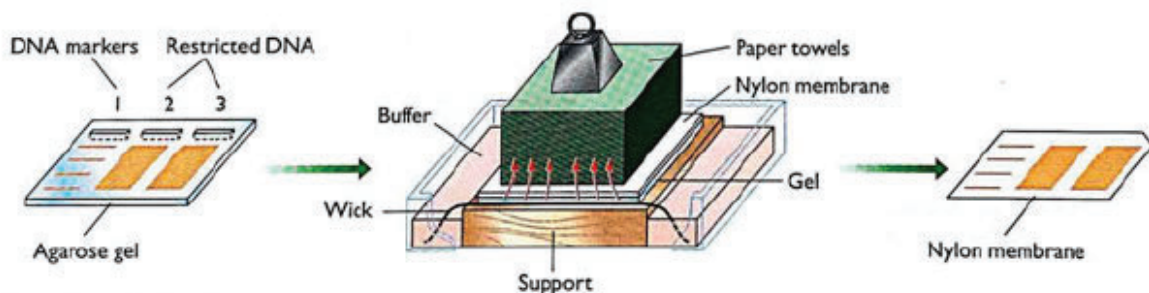
The range of fragment sizes that can be resolved depends on the concentration of agarose in the gel. Electrophoresis has been performed with three different concentrations of agarose. The labels indicate the sizes of bands in the left and right lanes. Photo courtesy of BioWhittaker Molecular Applications.

4.1.3 DNA ligases

DNA fragments that have been generated by treatment with a restriction endonuclease can be joined back together again, or attached to a new partner, by a DNA ligase. The reaction requires energy, which is provided by adding either ATP or NAD to the reaction mixture, depending on the type of ligase that is being used.

The most widely used DNA ligase is obtained from *E. coli* cells infected with T4 bacteriophage. It is involved in replication of the phage DNA and is encoded by the T4 genome. Its natural role is to synthesize phosphodiester bonds between unlinked nucleotides present in one polynucleotide of a double-stranded molecule (Figure 4.13A). In order to join together two restriction fragments, the ligase has to synthesize two phosphodiester bonds, one in each strand (Figure 4.13B). This is by no means beyond the capabilities of the enzyme, but the reaction can occur only if the ends to be joined come close enough to one another by chance – the ligase is not able to catch hold of them and bring them together. If the two molecules have complementary sticky ends, and the ends come together by random diffusion events in the ligation mixture, then transient base pairs might form between the two overhangs. These base pairs are not particularly stable but they may persist for sufficient time for a ligase enzyme to attach to the junction and synthesize phosphodiester bonds to fuse the ends together (Figure 4.13C). If the molecules are blunt ended, then they cannot base-pair to one another, not even temporarily, and ligation is a much less efficient process, even when the DNA concentration is high and pairs of ends are in relatively close proximity.

(A) Transfer of DNA from gel to membrane

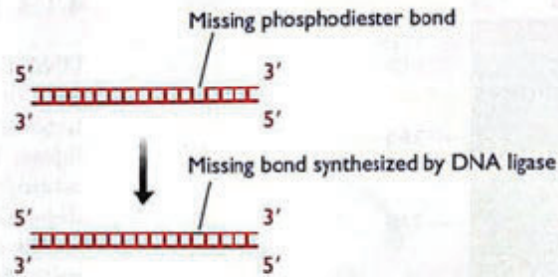
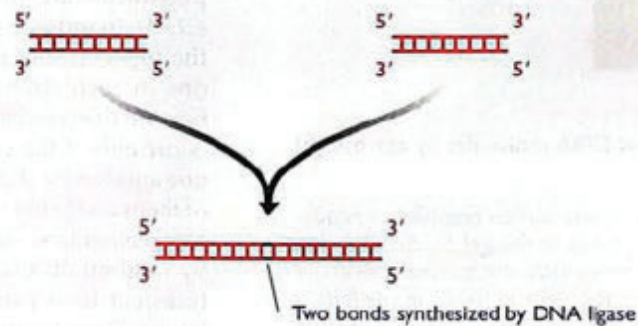


(B) Hybridization analysis

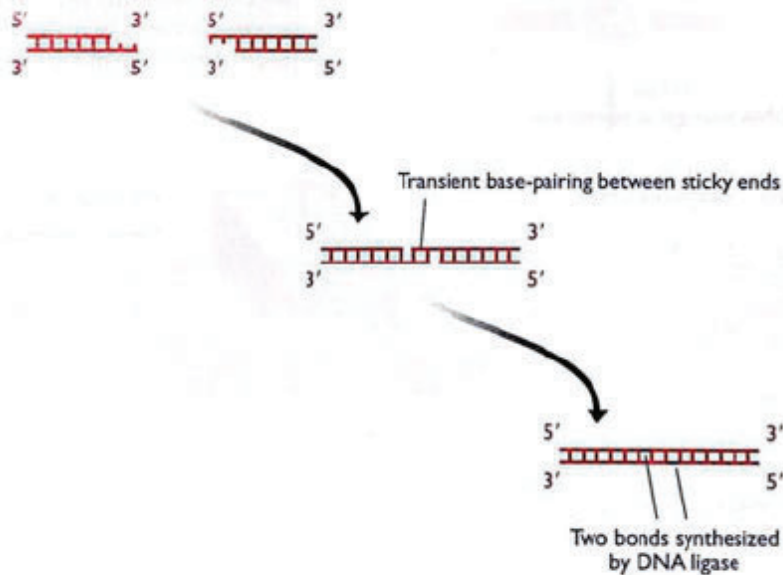


Figure 4.12 Southern hybridization.

(A) Transfer of DNA from the gel to the membrane. (B) The membrane is probed with a radioactively labeled DNA molecule. On the resulting autoradiograph, one hybridizing band is seen in lane 2, and two in lane 3.

(A) The role of DNA ligase *in vivo*(B) Ligation *in vitro*

(C) Sticky-end ligation is more efficient

**Figure 4.13** Ligation of DNA molecules with DNA ligase.

(A) In living cells, DNA ligase synthesizes a missing phosphodiester bond in one strand of a double-stranded DNA molecule. (B) To link two DNA molecules *in vitro*, DNA ligase must make two phosphodiester bonds, one in each strand. (C) Ligation *in vitro* is more efficient when the molecules have compatible sticky ends, because transient base-pairing between these ends holds the molecules together and so increases the opportunity for DNA ligase to attach and synthesize the new phosphodiester bonds. For the role of DNA ligase during DNA replication *in vivo*, see Figures 13.17 and 13.19.

The greater efficiency of sticky-end ligation has stimulated the development of methods for converting blunt ends into sticky ends. In one method, short double-stranded molecules called **linkers** or **adaptors** are attached to the blunt ends. Linkers and adaptors work in slightly different ways but both contain a recognition sequence for a restriction endonuclease and so produce a sticky end after treatment with the appropriate enzyme (Figure 4.14). Another way to create a sticky end is by

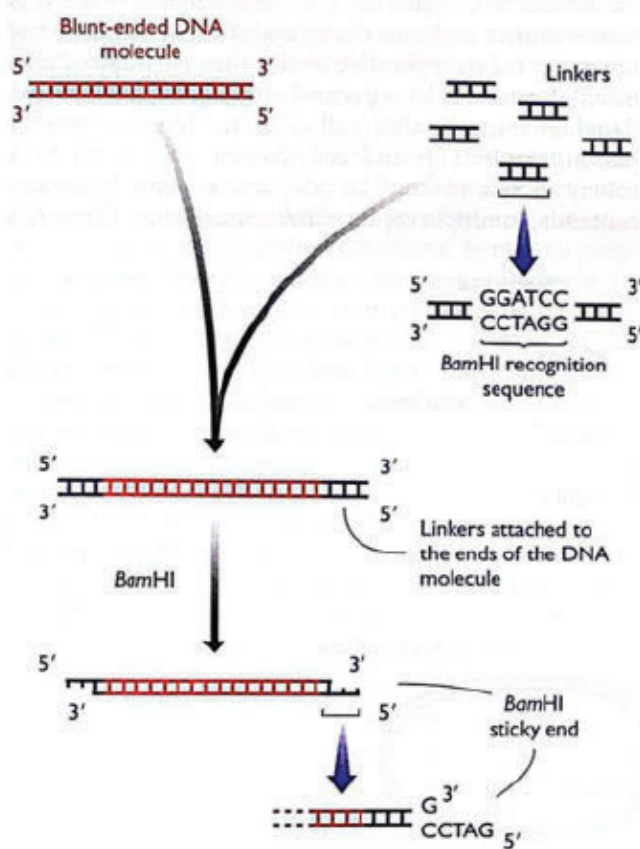


Figure 4.14 Linkers are used to place sticky ends on to a blunt-ended molecule.

In this example, each linker contains the recognition sequence for the *Bam*HI restriction endonuclease. DNA ligase attaches the linkers to the ends of the blunt-ended molecule in a reaction that is made relatively efficient because the linkers are present at a high concentration. The restriction enzyme is then added to cleave the linkers and produce the sticky ends. Note that during the ligation the linkers ligate to one another, so a series of linkers (a **concatamer**) is attached to each end of the blunt molecule. When the restriction enzyme is added, these linker concatamers are cut into segments, with half of the innermost linker left attached to the DNA molecule. Adaptors are similar to linkers but each one has one blunt end and one sticky end. The blunt-ended DNA is therefore given sticky ends simply by ligating it to the adaptors; there is no need to carry out the restriction step. For more details, see Brown (2001).

homopolymer tailing, in which nucleotides are added one after the other to the 3' terminus at a blunt end (Figure 4.15). The enzyme involved is called **terminal deoxynucleotidyl transferase**, which we will meet in the next section. If the reaction contains the DNA, enzyme, and only one of the four nucleotides, then the new stretch of single-stranded DNA that is made consists entirely of just that single nucleotide. It could, for example, be a poly(G) tail, which would enable the molecule to base-pair to other molecules that carry poly(C) tails, created in the same way but with dCTP rather than dGTP in the reaction mixture.

4.1.4 End-modification enzymes

Terminal deoxynucleotidyl transferase (see Figure 4.15), obtained from calf thymus tissue, is one example of an end-modification enzyme. It is, in fact, a **template-independent DNA polymerase**, because it is able to synthesize a new DNA polynucleotide without base-pairing of the incoming nucleotides to an existing strand of DNA or RNA. Its main role in recombinant DNA technology is in homopolymer tailing, as described above.

Two other end-modification enzymes are also frequently used. These are **alkaline phosphatase** and **T4 polynucleotide kinase**, which act in complementary ways. Alkaline phosphatase, which is obtained from various sources, including *E. coli* and calf intestinal tissue, removes phosphate groups from the 5' ends of DNA molecules, which prevents these molecules from being ligated to one another. Two ends carrying 5' phosphates can be ligated to one another, and a phosphatased end can ligate to a non-phosphatased end, but a link cannot be formed between a pair of ends if neither carries a

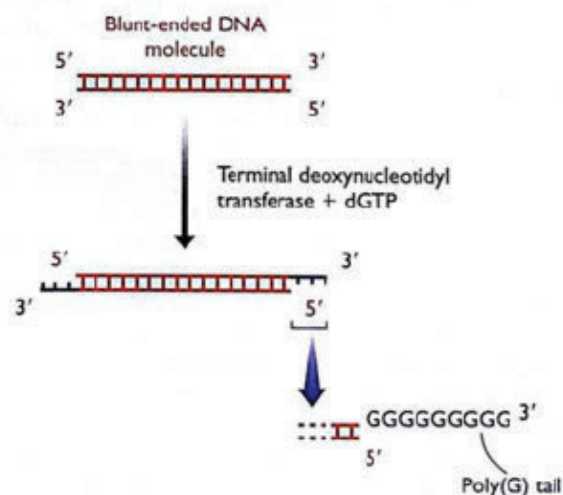


Figure 4.15 Homopolymer tailing.

In this example, a poly(G) tail is synthesized at each end of a blunt-ended DNA molecule. Tails comprising other nucleotides are synthesized by including the appropriate dNTP in the reaction mixture.

5' phosphate. Judicious use of alkaline phosphatase can therefore direct the action of a DNA ligase in a predetermined way so that only desired ligation products are obtained. T4 polynucleotide kinase, obtained from *E. coli* cells infected with T4 phage, performs the reverse reaction to alkaline phosphatase, adding phosphates to 5' ends. Like alkaline phosphatase, the enzyme is used during complicated ligation experiments, but its main application is in the end-labeling of DNA molecules (see Technical Note 4.1, page 98).

4.2 DNA Cloning

DNA cloning is a logical extension of the ability to manipulate DNA molecules with restriction endonucleases and ligases. Imagine that an animal gene has been obtained as a single restriction fragment after digestion of a larger

molecule with the restriction enzyme *Bam*HI, which leaves 5'-GATC-3' sticky ends (Figure 4.16). Imagine also that a small *E. coli* plasmid has been purified and treated with *Bam*HI, which cuts the plasmid in a single position. The circular plasmid has therefore been converted into a linear molecule, again with 5'-GATC-3' sticky ends. Mix the two DNA molecules together and add DNA ligase. Various recombinant ligation products will be obtained, one of which comprises the circularized plasmid with the animal gene inserted into the position originally taken by the *Bam*HI restriction site. If the recombinant plasmid is now re-introduced into *E. coli*, and the inserted gene has not disrupted its replicative ability, then the plasmid plus inserted gene will be replicated and copies passed to the daughter bacteria after cell division. More rounds of plasmid replication and cell division will result in a colony of recombinant *E. coli* bacteria, each bacterium containing multiple copies of the animal gene. This series

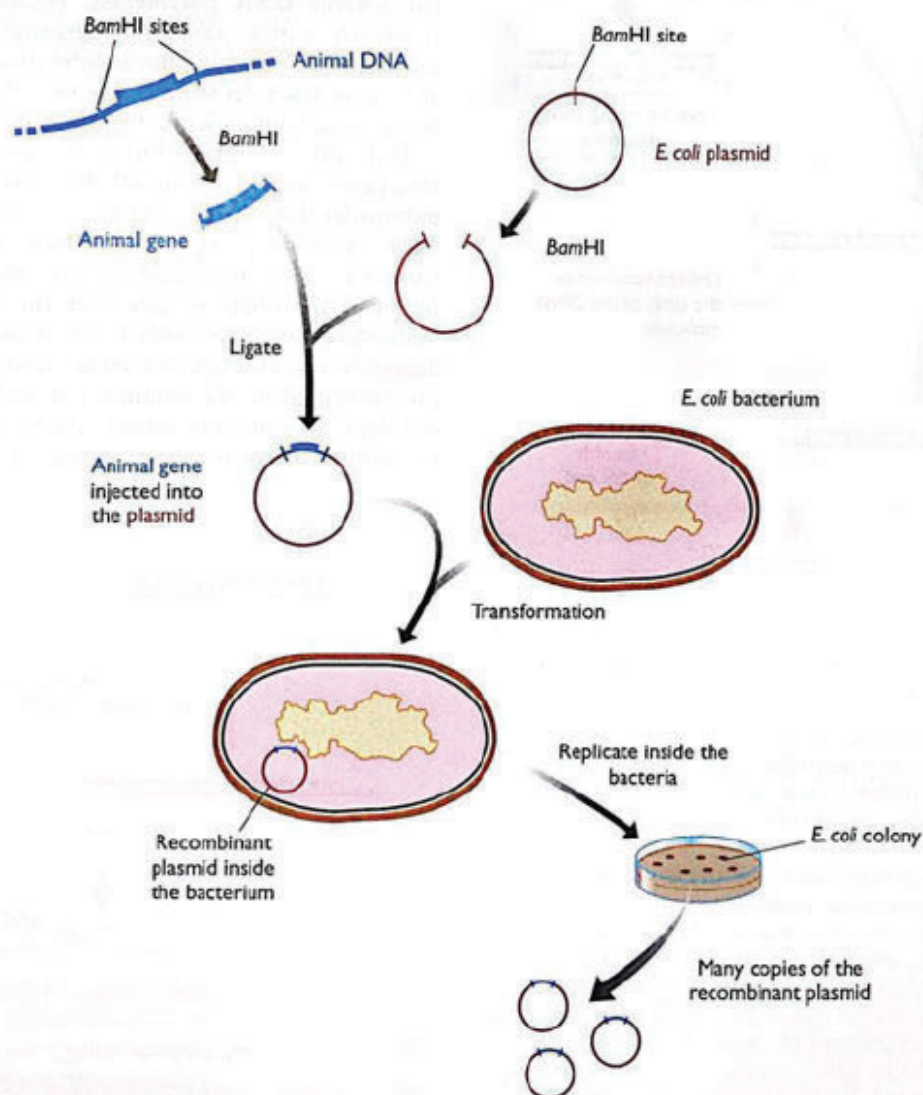


Figure 4.16 An outline of gene cloning.

See the text for details.

of events, as illustrated in Figure 4.16, constitutes the process called DNA or gene cloning.

4.2.1 Cloning vectors and the way they are used

In the experiment shown in Figure 4.16, the plasmid acts as a **cloning vector**, providing the replicative ability that enables the cloned gene to be propagated inside the host cell. Plasmids replicate efficiently in bacterial hosts because each plasmid possesses an **origin of replication** which is recognized by the DNA polymerases and other proteins that normally replicate the bacterium's chromosomes (Section 13.2.1). The host cell's replicative machinery therefore propagates the plasmid, plus any new genes that have been inserted into it. Bacteriophage genomes can also be used as cloning vectors because they too possess origins of replication that enable them to be propagated inside bacteria, either by the host enzymes or by DNA polymerases and other proteins specified by phage genes. The next two sections describe how plasmid and phage vectors are used to clone DNA in *E. coli*.

Plasmids are uncommon in eukaryotes, although *Saccharomyces cerevisiae* possesses one that is sometimes used for cloning purposes; most eukaryotic vectors are therefore based on virus genomes. Alternatively, with a eukaryotic host the replication requirement can be bypassed by performing the experiment in such a way that the DNA to be cloned becomes inserted into one of the host chromosomes. These approaches to cloning in eukaryotic cells are described later in the chapter.

Vectors based on *E. coli* plasmids

The easiest way to understand how a cloning vector is used is to start with the simplest *E. coli* plasmid vectors, which illustrate all of the basic principles of DNA cloning. We will then be able to turn our attention to the special features of phage vectors and vectors used with eukaryotes.

One of the first plasmid vectors to be developed was pBR322 (Bolivar *et al.*, 1977), which was constructed by ligating restriction fragments from three naturally occurring *E. coli* plasmids: R1, R6.5 and pMB1. The pBR322 plasmid is small (just 4363 bp) and, as well as the origin of replication, it carries genes coding for enzymes that enable the host bacterium to withstand the growth-inhibitory effects of two antibiotics: ampicillin and tetracycline (Figure 4.17). This means that cells containing a pBR322 plasmid can be distinguished from those that do not by plating the bacteria onto agar medium containing ampicillin and/or tetracycline. Normal *E. coli* cells are sensitive to these antibiotics and cannot grow when either of the two antibiotics is present. Ampicillin and tetracycline resistance are therefore **selectable markers** for pBR322.

The manipulations shown in Figure 4.16, resulting in construction of a recombinant plasmid, are carried out in the test tube with purified DNA. Pure pBR322 DNA can

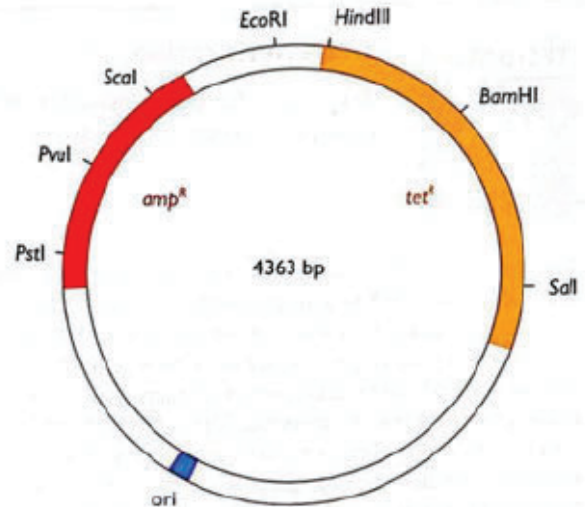


Figure 4.17 pBR322.

The map shows the positions of the ampicillin-resistance gene (amp^r), the tetracycline-resistance gene (tet^r), the origin of replication (ori) and the recognition sequences for seven restriction endonucleases.

be obtained quite easily from extracts of bacterial cells (Technical Note 4.2), but how can the manipulated plasmids be re-introduced into the bacteria? The answer is to make use of the natural processes for **transformation** of bacteria, which result in the uptake of 'naked' DNA by a bacterial cell. This is the process studied by Avery and his colleagues in the experiments which showed that bacterial genes are made of DNA (Section 1.1.1). Transformation is not a particularly efficient process in many bacteria, including *E. coli*, but the rate of uptake can be enhanced significantly by suspending the cells in calcium chloride before adding the DNA, and then briefly incubating the mixture at 42 °C. Even after this enhancement, only a very small proportion of the cells take up a plasmid. This is why the antibiotic-resistance markers are so important – they allow the small number of transformants to be selected from the large background of non-transformed cells.

The map of pBR322 shown in Figure 4.17 indicates the positions of the recognition sequences for seven restriction enzymes, each of which cuts the plasmid at just one location. Note that six of these sites lie within one or other of the genes for antibiotic resistance. This means that if a new fragment of DNA is ligated into one of these six sites, then the antibiotic-resistance properties of the plasmid become altered – the plasmid loses the ability to confer either ampicillin or tetracycline resistance on the host cells. This is called **insertional inactivation** of the selectable marker and is the key to distinguishing a **recombinant** plasmid – one that contains an inserted piece of DNA – from a non-recombinant plasmid that has no new DNA. Identifying recombinants is important because the manipulations illustrated in Figure 4.16 result in a mixture of ligation products, including plasmids that have recir-

TECHNICAL

4.2

NOTE

DNA purification

Techniques for the preparation of pure samples of DNA from living cells play a central role in molecular biology research.

The first step in DNA purification is to break open the cells from which the DNA will be obtained. With some types of material this step is easy: cultured animal cells, for example, are broken open simply by adding a detergent such as sodium dodecyl sulfate (SDS), which disrupts the cell membranes, releasing the cell contents. Other types of cell have strong walls and so demand a harsher treatment. Plant cells are usually frozen and then ground in a mortar and pestle, this being the only effective way of breaking their cellulose walls. Bacteria such as *Escherichia coli* can be lysed by a combination of enzymatic and chemical treatment. The enzyme is **lysozyme**, obtained from egg white, which breaks down the polymeric compounds in the bacterial cell wall; the chemical is ethylenediamine tetra-acetate (EDTA), which chelates magnesium ions, further reducing the integrity of the cell wall. Disruption of the cell membrane by adding a detergent then causes the cells to burst.

Once the cells have been broken, two different methods can be used to purify the DNA from the resulting extract. The first involves degrading or removing all the cellular components other than the DNA, an approach that works best if the cells do not contain large amounts of lipid or carbohydrate. The extract is first centrifuged at low speed to remove debris such as pieces of cell wall, which form a pellet at the bottom of the tube. The supernatant is transferred to another test tube and mixed with phenol, which causes the protein to precipitate at the interface between the organic and aqueous layers. The aqueous layer, which contains the dissolved nucleic acids, is collected and a ribonuclease enzyme added, which breaks the RNA into a mixture of nucleotides and short oligonucleotides. The DNA polynucleotides, which remain intact, can now be precipitated by adding ethanol, pelleted by centrifugation, and resuspended in an appropriate volume of buffer.

In the second method for DNA purification, rather than degrading everything other than DNA, the DNA itself is selectively removed from the extract. One way of doing this

is by adding the detergent cetyltrimethylammonium bromide (CTAB), which forms an insoluble complex with nucleic acids. The precipitate is collected by centrifugation and resuspended in a high-salt solution, which causes the complex to break down, releasing the nucleic acids. Ribonuclease treatment followed by ethanol precipitation yields pure DNA from this mixture. Another popular technique makes use of the tight binding between DNA and silica particles that occurs in the presence of a denaturing chemical such as guanidinium thiocyanate. The silica particles and guanidinium thiocyanate can be added directly to the extract and the DNA collected by centrifugation. Alternatively, the silica can be placed in a chromatography column and the extract, plus guanidinium thiocyanate, passed through. The DNA binds to the silica particles in the column and is subsequently recovered by washing away the guanidinium thiocyanate with water; the DNA now detaching from the silica and eluting from the column.

The two approaches described above purify all the DNA in a cell. Special methods are needed if the aim is to obtain just plasmid DNA (for example, recombinant cloning vectors) from bacterial cells. One popular method makes use of the fact that, although both plasmids and the bacterial chromosome are made up of supercoiled DNA, lysis of the bacterial cell inevitably leads to a certain amount of disruption of the nucleoid, leading to breakage of the loops of supercoiled chromosomal DNA (Section 2.3.1). A cell extract therefore contains supercoiled plasmid DNA and *non-supercoiled* chromosomal DNA, and the plasmids can be purified by a method that distinguishes between DNA molecules with these different conformations. One technique involves adding sodium hydroxide until the pH of the cell extract reaches 12.0–12.5, which causes the base pairs in non-supercoiled DNA to break. The resulting single strands tangle up into an insoluble network that can be removed by centrifugation, leaving the supercoiled plasmids in the supernatant.

cularized without insertion of new DNA. To identify recombinants, the resistance properties of colonies are tested by transferring cells from agar containing one antibiotic onto agar containing the second antibiotic. For example, if the *Bam*HI site has been used then recombinants will be ampicillin resistant but tetracycline sensitive, because the *Bam*HI site lies within the region that specifies resistance to tetracycline. After transformation, cells are plated onto ampicillin agar (Figure 4.18). All cells that contain a pBR322 plasmid, whether recombinant or

not, divide and produce a colony. The colonies are then transferred onto tetracycline agar by **replica plating**, which results in the colonies on the second plate retaining the relative positions that they had on the first plate. Some colonies do not grow on the tetracycline agar because their cells contain recombinant pBR322 molecules with a disrupted tetracycline-resistance gene. These are the colonies we are looking for because they contain the cloned gene, so we return to the ampicillin plate, from which samples of the cells can be recovered.

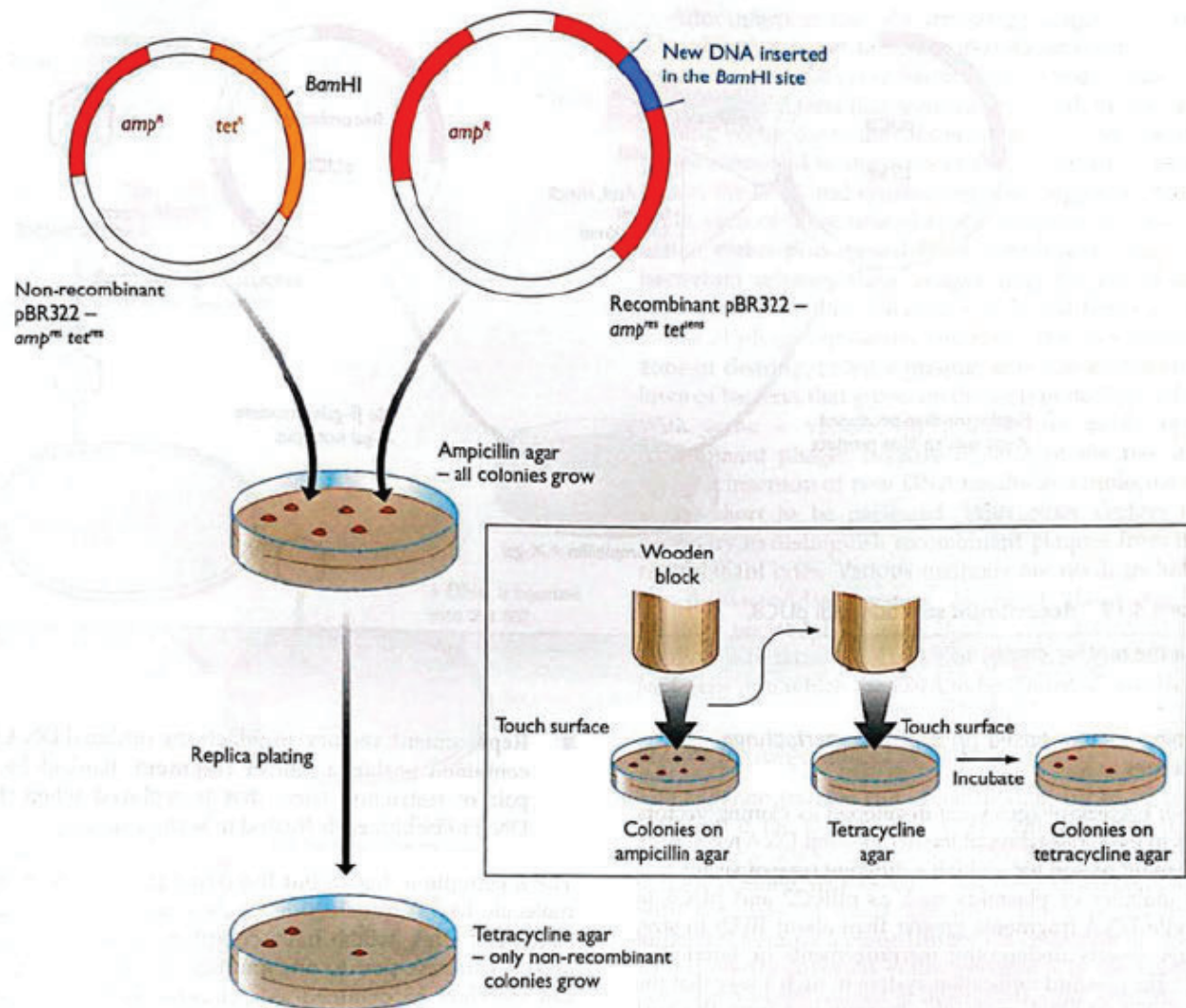


Figure 4.18 Recombinant selection with pBR322.

See text for details. The inset shows how replica plating is performed.

Replica plating is not a difficult technique but it takes time. It would be much better if recombinants could be distinguished from non-recombinants simply by plating onto a single agar medium. This is possible with most of the modern plasmid cloning vectors, including pUC8 (Figure 4.19; Vieira and Messing, 1982). This vector carries the ampicillin-resistance gene from pBR322, along with a second gene, called *lacZ'*, which is part of the *E. coli* gene for the enzyme β -galactosidase. The remainder of the *lacZ* gene is located in the chromosome of the special *E. coli* strain that is used when cloning genes with pUC8. The proteins specified by the gene segments on the plasmid and on the chromosome are able to combine to produce a functional β -galactosidase enzyme. The presence of functional β -galactosidase molecules in the cells can be checked by a histochemical test with a compound called X-gal (5-bromo-4-chloro-

3-indolyl- β -D-galactopyranoside), which the enzyme converts into a blue product. The *lacZ'* gene contains a cluster of unique restriction sites; insertion of new DNA into any one of these sites results in insertional inactivation of the gene and hence loss of β -galactosidase activity. Recombinants and non-recombinants can therefore be distinguished simply by plating the transformed cells onto agar containing ampicillin and X-gal (Figure 4.19). All colonies that grow on this medium are made up of transformed cells because only transformants are ampicillin resistant. Some colonies are blue and some are white. Those that are blue contain cells with functional β -galactosidase enzymes and hence with undisturbed *lacZ'* genes; these colonies are the non-recombinants. The white colonies comprise cells without β -galactosidase activity and hence with disrupted *lacZ'* genes; these are recombinants.

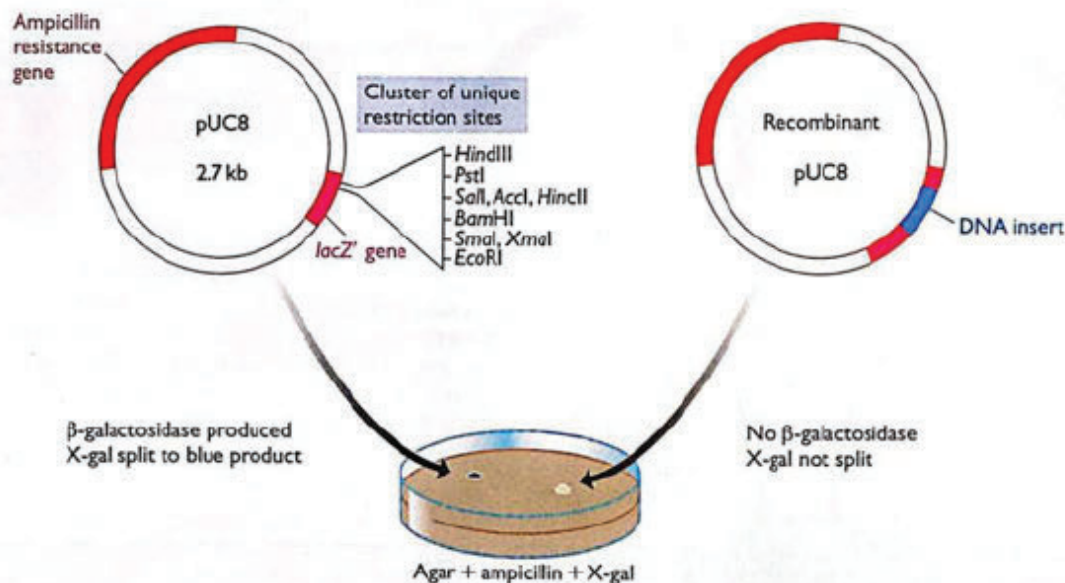


Figure 4.19 Recombinant selection with pUC8.

See the text for details.

Cloning vectors based on *E. coli* bacteriophage genomes

E. coli bacteriophages were developed as cloning vectors back in the earliest days of the recombinant DNA revolution. The main reason for seeking a different type of vector was the inability of plasmids such as pBR322 and pUC8 to handle DNA fragments greater than about 10 kb in size, larger inserts undergoing rearrangements or interfering with the plasmid replication system in such a way that the recombinant DNA molecules become lost from the host cells. The first attempts to develop vectors able to handle larger fragments of DNA centered on bacteriophage λ . The infection cycle of λ is similar to that of the T2 phages studied by Hershey and Chase in the experiments that alerted molecular biologists to the fact that genes are made of DNA (Section 1.1.1), but there is one important difference. As well as following the **lytic infection cycle** (see Figure 1.4B, page 9), the λ genome is able to integrate into the bacterial chromosome, where it can remain quiescent for many generations, being replicated along with the host chromosome every time the cell divides. This is called the **lysogenic infection cycle** (Figure 4.20).

The λ genome is 48.5 kb, of which some 15 kb or so is 'optional' in that it contains genes that are only needed for integration of the phage DNA into the *E. coli* chromosome (Figure 4.21A). These segments can therefore be deleted without impairing the ability of the phage to infect bacteria and direct synthesis of new λ particles by the lytic cycle. Two types of vector have been developed (Figure 4.21B):

- **Insertion vectors**, in which part or all of the optional DNA has been removed and a unique restriction site introduced at some position within the trimmed down genome;

- **Replacement vectors**, in which the optional DNA is contained within a **stuffer fragment**, flanked by a pair of restriction sites, that is replaced when the DNA to be cloned is ligated into the vector.

The λ genome is linear, but the two natural ends of the molecule have 12-nucleotide single-stranded overhangs, called *cos* sites, which have complementary sequences and so can base-pair to one another. A λ cloning vector can therefore be obtained as a circular molecule which can be manipulated in the test tube in the same way as a plasmid, and re-introduced into *E. coli* by **transfection**, the term used for uptake of naked phage DNA. Alternatively, a more efficient uptake system called ***in vitro* packaging** can be utilized (Hohn and Murray, 1977). This procedure starts with the linear version of the cloning vector, the initial restriction cutting the molecule into two segments, the left and right arms, each with a *cos* site at one end. The ligation is carried out with carefully measured quantities of each arm and the DNA to be cloned, the aim being to produce concatamers in which the different fragments are linked together in the order left arm–new DNA–right arm, as shown in Figure 4.22. The concatamers are then added to an *in vitro* packaging mix, which contains all the proteins needed to make a λ phage particle. These proteins form phage particles spontaneously, and will place inside the particles any DNA fragment that is between 37 and 52 kb in length and is flanked by *cos* sites. The packaging mix therefore cuts left arm–new DNA–right arm combinations of 37–52 kb out of the concatamers and constructs λ phages around them. The phages are then mixed with *E. coli* cells, and the natural infection process transports the vector plus new DNA into the bacteria.

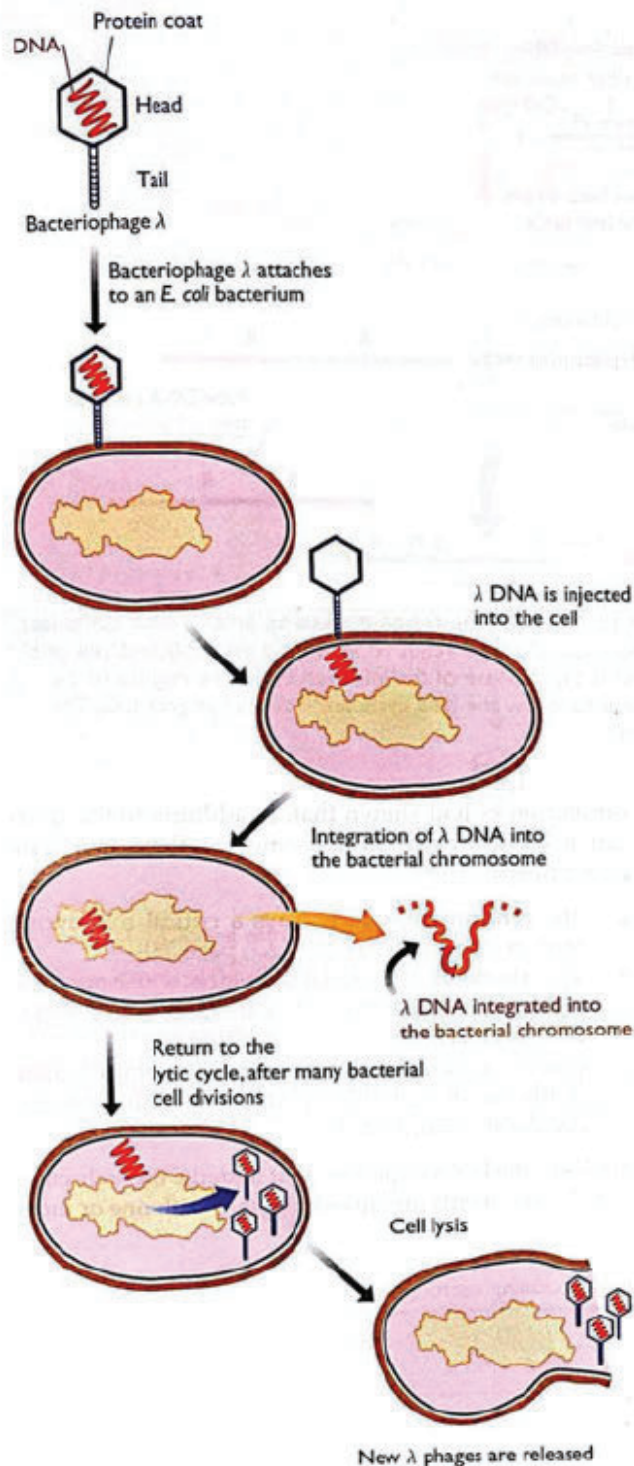


Figure 4.20 The lysogenic infection cycle of bacteriophage λ .

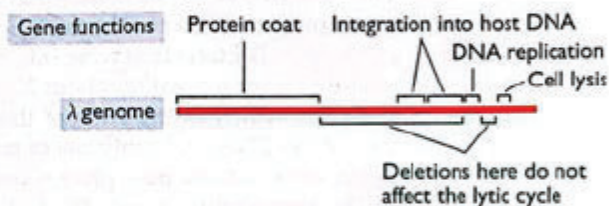
Compare with the lytic infection cycle of T2 bacteriophage, shown in Figure 1.4B (page 9). The special feature of the lysogenic cycle is the insertion of the phage genome into the bacterium's chromosomal DNA, where it can remain quiescent for many generations.

After infection, the cells are spread onto an agar plate. The objective is not to obtain individual colonies but to produce an even layer of bacteria across the entire surface of the agar. Bacteria that were infected with the packaged cloning vector die within about 20 minutes because the λ genes contained in the arms of the vector direct replication of the DNA and synthesis of new phages by the lytic cycle, each of these new phages containing its own copy of the vector plus cloned DNA. Death and lysis of the bacterium releases these phages into the surrounding medium, where they infect new cells and begin another round of phage replication and lysis. The end result is a zone of clearing, called a **plaque**, which is visible on the lawn of bacteria that grows on the agar plate (Figure 4.23). With some λ vectors, all plaques are made up of recombinant phages because ligation of the two arms without insertion of new DNA results in a molecule that is too short to be packaged. With other vectors it is necessary to distinguish recombinant plaques from non-recombinant ones. Various methods are used, including the β -galactosidase system described above for the plasmid vector pUC8 (see Figure 4.19), which is also applicable to those λ vectors that carry a fragment of the *lacZ* gene into which the DNA to be cloned is inserted.

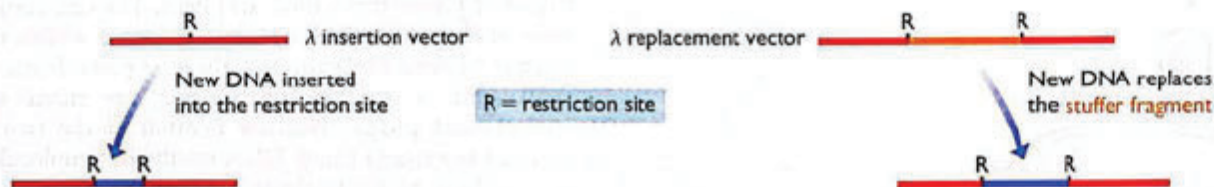
Vectors for longer pieces of DNA

The λ phage particle can accommodate up to 52 kb of DNA, so if the genome has 15 kb removed then up to 18 kb of new DNA can be cloned. This limit is higher than that for plasmid vectors, but is still very small compared with the sizes of intact genomes. The comparison is important because a **clone library** – a collection of clones whose inserts cover an entire genome – is the starting point for a project aimed at determining the sequence of that genome (Chapter 6). If a λ vector is used with human DNA, then over half a million clones are needed for there to be a 95% chance of any particular part of the genome being present in the library (Table 4.4). It is possible to prepare a library comprising half a million clones, especially if automated techniques are used, but such a large collection is far from ideal. It would be much better to reduce the number of clones by using a vector that is able to handle fragments of DNA longer than 18 kb. Many of the developments in cloning technology over the last 20 years have been aimed at finding ways of doing this.

One possibility is to use a **cosmid** – a plasmid that carries a λ *cos* site (Figure 4.24). Concatamers of cosmid molecules, linked at their *cos* sites, act as substrates for *in vitro* packaging because the *cos* site is the only sequence that a DNA molecule needs in order to be recognized as a ' λ genome' by the proteins that package DNA into λ phage particles. Particles containing cosmid DNA are as infective as real λ phages, but once inside the cell the cosmid cannot direct synthesis of new phage particles and instead replicates as a plasmid. Recombinant DNA is therefore obtained from colonies rather than plaques. As with other types of λ vector, the upper limit for the size of the cloned DNA is set by the space available

(A) The λ genome contains 'optional' DNA

(B) Insertion and replacement vectors

**Figure 4.21** Cloning vectors based on bacteriophage λ .

(A) In the λ genome, the genes are arranged into functional groups. For example, the region marked as 'protein coat' comprises 21 genes coding for proteins that are either components of the phage capsid or are required for capsid assembly, and 'cell lysis' comprises four genes involved in lysis of the bacterium at the end of the lytic phase of the infection cycle. The regions of the genome that can be deleted without impairing the ability of the phage to follow the lytic cycle are indicated in green. (B) The differences between a λ insertion vector and a λ replacement vector.

λ phage particle. A cosmid can be 8 kb or less in size, so up to 44 kb of new DNA can be inserted before the packaging limit of the λ phage particle is reached. This reduces the size of the human genomic library to about a quarter of a million clones, which is an improvement compared with a λ library, but is still a massive number of clones to have to work with.

The first major breakthrough in attempts to clone DNA fragments much longer than 50 kb came with the invention of **yeast artificial chromosomes** or **YACs** (Burke *et al.*, 1987). These vectors are propagated in *S. cerevisiae* rather than in *E. coli* and are based on chromosomes, rather than on plasmids or viruses. The first YACs were constructed after studies of natural

chromosomes had shown that, in addition to the genes that it carries, each chromosome has three important components:

- The **centromere**, which plays a critical role during nuclear division (see Figure 2.7, page 38);
- The **telomeres**, the special sequences which mark the ends of chromosomal DNA molecules (see Figure 2.10, page 41);
- One or more **origins of replication**, which initiate synthesis of new DNA when the chromosome divides (Section 13.2.1).

In a YAC, the DNA sequences that underlie these chromosomal components are linked together with one or more

Table 4.4 Sizes of human genomic libraries prepared in different types of cloning vector

Type of vector	Insert size (kb)	Number of clones*	
		P = 95%	P = 99%
λ replacement	18	532 500	820 000
Cosmid, fosmid	40	240 000	370 000
PI	125	77 000	118 000
BAC, PAC	300	32 000	50 000
YAC	600	16 000	24 500
Mega-YAC	1400	6850	10 500

*Calculated from the equation:

$$N = \frac{\ln(1-P)}{\ln(1-\frac{c}{b})}$$

where N is the number of clones required, P is the probability that any given segment of the genome is present in the library, c is the average size of the DNA fragments inserted into the vector, and b is the size of the genome.

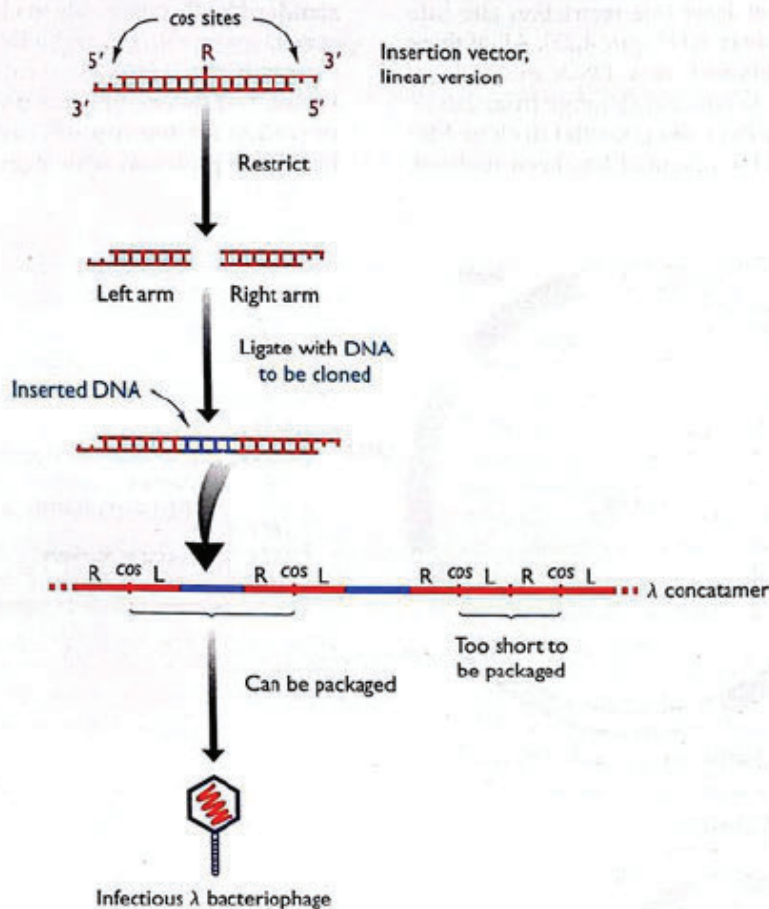
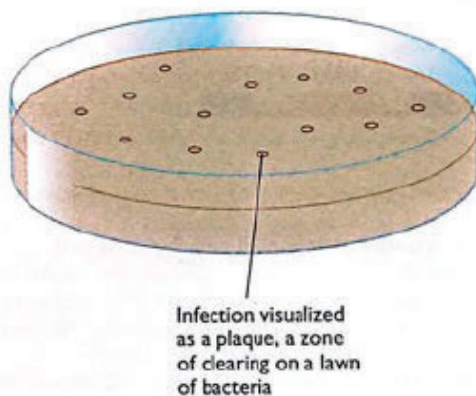


Figure 4.22 Cloning with a λ insertion vector.

The linear form of the vector is shown at the top of the diagram. Treatment with the appropriate restriction endonuclease produces the left and right arms, both of which have one blunt end and one end with the 12-nucleotide overhang of the *cos* site. The DNA to be cloned is blunt ended and so is inserted between the two arms during the ligation step. These arms also ligate to one another via their *cos* sites, forming a concatamer. Some parts of the concatamer comprise left arm–insert DNA–right arm and, assuming this combination is 37–52 kb in length, will be enclosed inside the capsid by the *in vitro* packaging mix. Parts of the concatamer made up of left arm ligated directly to right arm, without new DNA, are too short to be packaged.



Infection visualized as a plaque, a zone of clearing on a lawn of bacteria

Figure 4.23 Bacteriophage infection is visualized as a plaque on a lawn of bacteria.

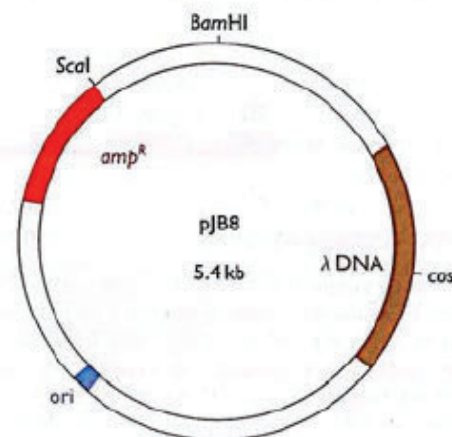


Figure 4.24 A typical cosmid.

pJB8 is 5.4 kb in size and carries the ampicillin-resistance gene (*amp^R*), a segment of λ DNA containing the *cos* site, and an *Escherichia coli* origin of replication (*ori*).

selectable markers and at least one restriction site into which new DNA can be inserted (Figure 4.25). All of these components can be contained in a DNA molecule of 10–15 kb. Natural yeast chromosomes range from 230 kb to over 1700 kb, so YACs have the potential to clone Mb-sized DNA fragments. This potential has been realized,

standard YACs being able to clone 600 kb fragments, with special types able to handle DNA up to 1400 kb in length. Currently this is the highest capacity of any type of cloning vector, and several genome projects have made extensive use of YACs. Unfortunately, with some types of YAC there have been problems with insert stability, the cloned DNA

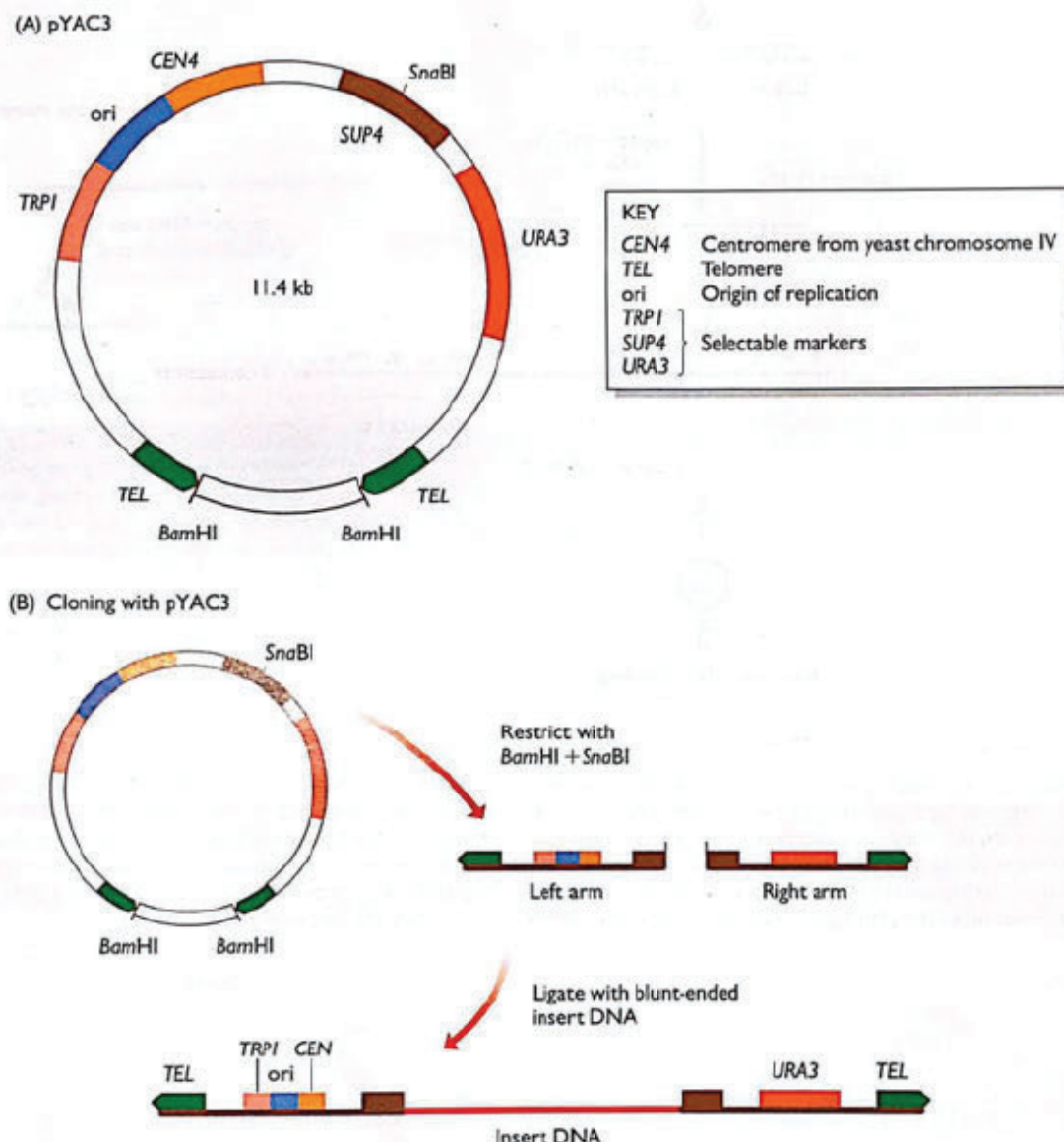


Figure 4.25 Working with a YAC.

(A) The cloning vector pYAC3. (B) To clone with pYAC3, the circular vector is digested with *Bam*HI and *Sna*BI. *Bam*HI restriction removes the stuffer fragment held between the two telomeres in the circular molecule. *Sna*BI cuts within the *SUP4* gene and provides the site into which new DNA will be inserted. Ligation of the two vector arms with new DNA produces the structure shown at the bottom. This structure carries functional copies of the *TRP1* and *URA3* selectable markers. The host strain has inactivated copies of these genes, which means that it requires tryptophan and uracil as nutrients. After transformation, cells are plated onto a minimal medium, lacking tryptophan and uracil. Only cells that contain the vector, and so can synthesize tryptophan and uracil, are able to survive on this medium and produce colonies. Note that if a vector comprises two right arms, or two left arms, then it will not give rise to colonies because the transformed cells will still require one of the nutrients. The presence of insert DNA in the cloned vector molecules is checked by testing for inactivation of *SUP4*. This is done by a color test: on the appropriate medium, colonies containing recombinant vectors (i.e. with an insert) are white; non-recombinants (vector but no insert) are red.

TECHNICAL

4.3

NOTE

Working with a clone library

Clone collections used as a source of genes and other DNA segments.

Clone libraries have been used in molecular biology research for many years and their importance extends well beyond their role as the starting point for a genome sequencing project. Since the 1970s, clone libraries have been prepared from different organisms as a means of obtaining individual genes and other DNA segments for further study by sequencing and other recombinant DNA techniques.

Clone libraries can be prepared from either genomic DNA or cDNA (Section 7.1.2), using a plasmid or bacteriophage vector. The clones are usually stored as bacterial colonies or plaques on 23×23 cm agar plates, with 100 000–150 000 clones per plate. A complete human library can therefore be contained in just 1–8 plates, depending on the type of cloning vector that has been used (see Table 4.4). Three methods can identify the clone that contains the gene or other piece of DNA that is being sought:

- **Hybridization analysis** can be performed with an oligonucleotide or other DNA molecule that is known to hybridize to the sequence of interest. To do this, a nylon or nitrocellulose membrane is placed on the surface of the agar dish and then carefully removed to 'lift off' the colonies or plaques. Treatment with alkali and protease degrades the cellular material, leaving behind the DNA from each clone, which is then bound tightly to the surface of the membrane by heating or ultraviolet irradiation. The labeled probe is now applied to the membrane, in the same way as in Southern hybridization (Figure 4.12), and the position at which the probe attaches determined by the appropriate detection method. The position of the hybridization signal on the membrane corresponds to the location of the clone of interest on the agar plate.
- **PCR** (Section 4.3) can be used to screen clones for the sequence of interest. This cannot be done *in situ*, so individual clones must be transferred to the wells of microtiter trays. The PCR approach to clone identification is therefore relatively cumbersome because only a few hundred clones can be accommodated in a single tray. PCRs using primers specific for the sequence of interest are performed with each clone in turn, possibly using a combinatorial approach, which reduces the number of PCRs that are needed in order to identify the one that gives a positive result (see Figure 6.14, page 178).
- **Immunological techniques** can be used if the sequence being sought is a gene that is expressed in the cell in which the clone library has been prepared. If gene expression is occurring then the protein product will be made, and this can be detected by screening the library with a labeled antibody that binds only to that protein. As in hybridization analysis, the clones are first transferred onto a membrane and then treated to break down the cells and bind the protein to the membrane surface. Exposure of the membrane to the labeled antibody then reveals the position of the clone containing the gene of interest.

becoming rearranged into new sequence combinations (Anderson, 1993). For this reason there is also great interest in other types of vectors, ones that cannot clone such large pieces of DNA but which suffer less from instability problems. These vectors include the following:

- **Bacteriophage P1 vectors** (Sternberg, 1990) are very similar to λ vectors, being based on a deleted version of a natural phage genome, the capacity of the cloning vector being determined by the size of the deletion and the space within the phage particle. The P1 genome is larger than the λ genome, and the phage particle is bigger, so a P1 vector can clone larger fragments of DNA than a λ vector, up to 125 kb using current technology.
- **Bacterial artificial chromosomes or BACs** (Shizuya *et al.*, 1992) are based on the naturally occurring F plasmid of *E. coli*. Unlike the plasmids used to construct the early cloning vectors, the F plasmid is relatively large and vectors based on it have a higher capacity for accepting inserted DNA. BACs can be used to clone fragments of 300 kb and longer.
- **P1-derived artificial chromosomes or PACs** (Ioannou *et al.*, 1994) combine features of P1 vectors and BACs and have a capacity of up to 300 kb.
- **Fosmids** (Kim *et al.*, 1992) contain the F plasmid origin of replication and a λ *cos* site. They are similar to cosmids but have a lower copy number in *E. coli*, which means that they are less prone to instability problems.

The sizes of human genome libraries prepared in these various types of vector are given in Table 4.4.

Cloning in organisms other than *E. coli*

Cloning is not merely an aid to DNA sequencing, it provides a means of studying the mode of expression of a gene and the way in which expression is regulated, of

carrying out genetic engineering experiments aimed at modifying the biological characteristics of the host organism, and of synthesizing important animal proteins, such as pharmaceuticals, in a new host cell from which the proteins can be obtained in larger quantities than is possible by conventional purification from

animal tissue. These multifarious applications demand that genes must frequently be cloned in organisms other than *E. coli*.

Cloning vectors based on plasmids or phages have been developed for most of the well studied species of bacteria such as *Bacillus*, *Streptomyces* and *Pseudomonas*,

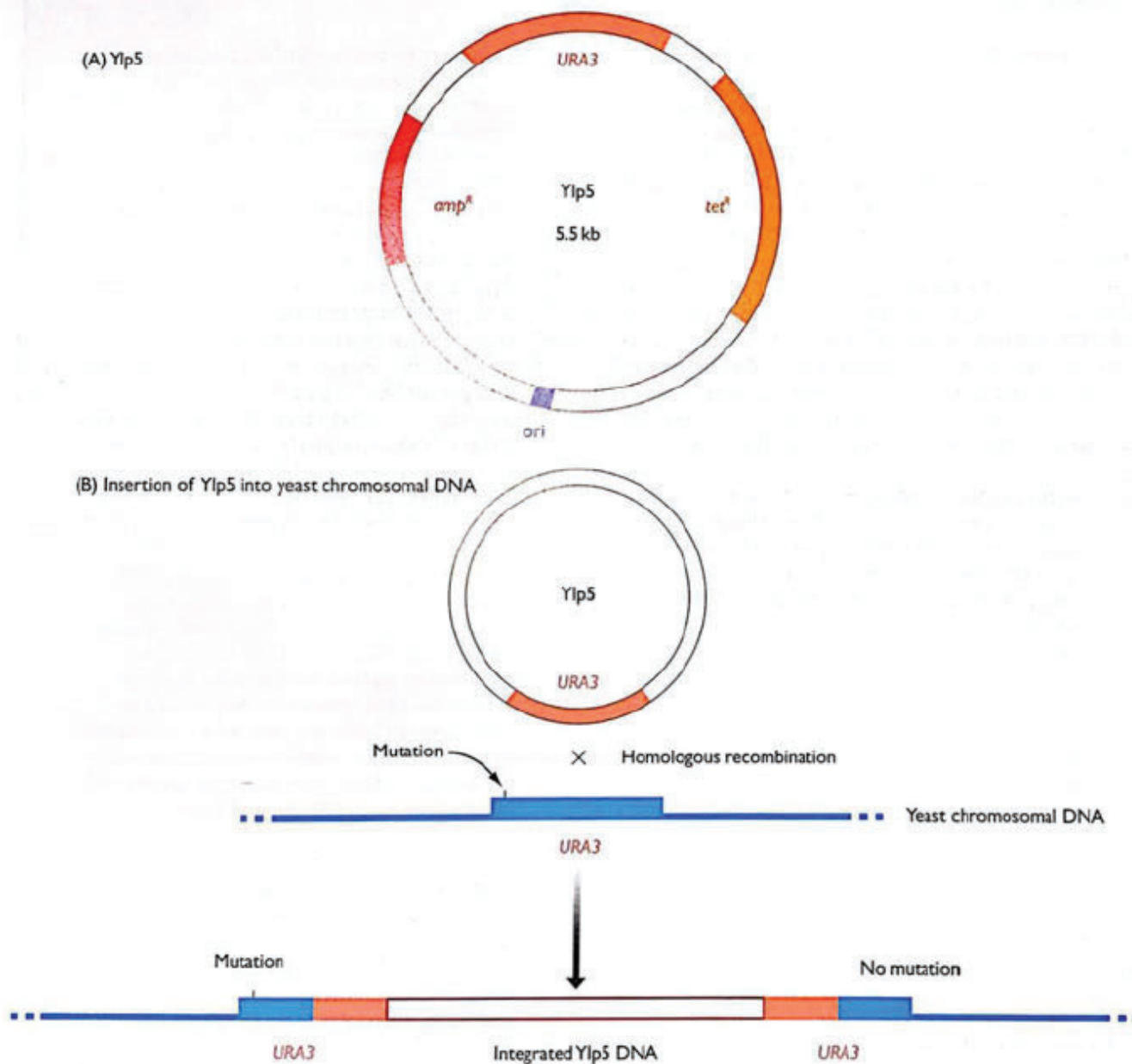


Figure 4.26 Cloning with a Ylp.

(A) Ylp5, a typical yeast integrative plasmid. The plasmid contains the ampicillin-resistance gene (*amp^r*), the tetracycline-resistance gene (*tet^r*), the yeast gene *URA3*, and an *Escherichia coli* origin of replication (*ori*). The presence of the *E. coli* *ori* means that recombinant Ylp5 molecules can be constructed in *E. coli* before their transfer into yeast cells. Ylp5 is therefore a **shuttle vector** – it can be shuttled between two species. (B) Ylp5 has no origin of replication that can function inside yeast cells, but can survive if it integrates into the yeast chromosomal DNA by homologous recombination between the plasmid and chromosomal copies of the *URA3* gene. The chromosomal gene carries a small mutation which means that it is non-functional and the host cells are *ura3⁻*. One of the pair of *URA3* genes that are formed after integration of the plasmid DNA is mutated, but the other is not. Recombinant cells are therefore *ura3⁺* and can be selected by plating on to minimal medium, which does not contain uracil.

these vectors being used in exactly the same way as the *E. coli* analogs. Plasmid vectors are also available for yeasts and fungi. Some of these carry the origin of replication from the 2 μ m circle, a plasmid present in many strains of *S. cerevisiae*, but other plasmid vectors only have an *E. coli* origin. An example is YIp5, an *S. cerevisiae* vector that is simply a pBR322 plasmid that contains a copy of the yeast gene called *URA3* (Figure 4.26A). What is the logic behind the construction of YIp5? When used in a cloning experiment, the vector is initially used with *E. coli* as the host, up to the stage where the desired recombinant molecule has been constructed by restriction and ligation. The recombinant vector is then purified from *E. coli* and transferred into *S. cerevisiae*, usually by mixing the DNA with **protoplasts** – yeast cells whose walls have been removed by enzyme treatment. Without an origin of replication the vector is unable to propagate independently inside yeast cells, but it can survive if it becomes integrated into one of the yeast chromosomes, which can occur by **homologous recombination** (Section 7.2.2) between the *URA3* gene carried by the vector and the chromosomal copy of this gene (Figure 4.26B). 'YIp' in fact stands for 'yeast integrative plasmid'. Once integrated the YIp, plus any DNA that has been inserted into it, replicates along with the host chromosomes.

Integration into chromosomal DNA is also a feature of many of the cloning systems used with animals and plants, and forms the basis of the construction of **knockout mice**, which are used to determine the functions of previously unknown genes that are discovered in the human genome (Section 7.2.2). The vectors are animal equivalents of YIps. Adenoviruses and retroviruses are used to clone genes in animals when the objective is to treat a genetic disease or a cancer by **gene therapy** (Lemoine and Cooper, 1998). A similar range of vectors has been developed for cloning genes in plants. Plasmids can be introduced into plant embryos by bombardment with DNA-coated microprojectiles, a process called **biolistics** (Klein *et al.*, 1987), integration of the plasmid DNA into the plant chromosomes, followed by growth of the embryo, resulting in a plant that contains the cloned DNA in most or all of its cells. Some success has also been achieved with plant vectors based on the genomes of caulimoviruses and geminiviruses (Timmermans *et al.*, 1994; Viaplana *et al.*, 2001), but the most interesting types of plant cloning vector are those derived from the **Ti plasmid** (Hansen and Wright, 1999). This is a large bacterial plasmid found in the soil microorganism *Agrobacterium tumefaciens*, part of which, the **T-DNA**, becomes integrated into a plant chromosome when the bacterium infects a plant stem and causes crown root disease. The T-DNA carries a number of genes that are expressed inside the plant cells and induce the various physiological changes that characterize the disease. Vectors such as pBIN19 (Figure 4.27) have been designed to make use of this natural genetic engineering system (Bevan, 1984). The recombinant vector is introduced into *A. tumefaciens* cells, which are allowed to infect a cell suspension or

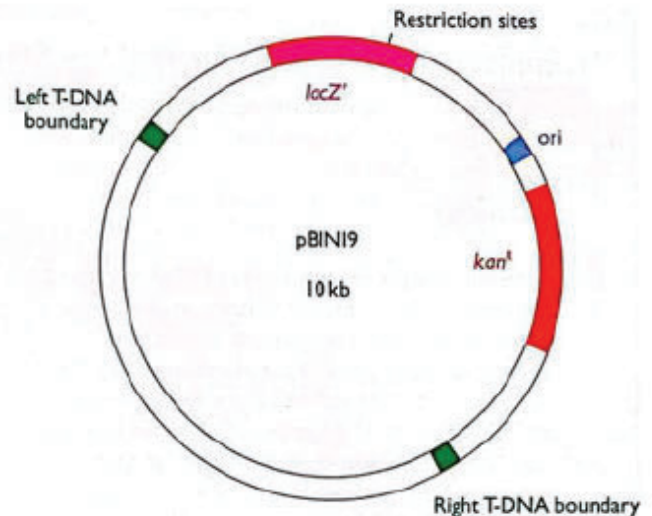


Figure 4.27 The plant cloning vector pBIN19.

pBIN19 carries the *lacZ'* gene (see Figure 4.19), the kanamycin-resistance gene (*kan^r*), an *Escherichia coli* origin of replication (*ori*), and the two boundary sequences from the T-DNA region of the Ti plasmid. These two boundary sequences recombine with plant chromosomal DNA, inserting the segment of DNA between them into the plant DNA. The orientation of the boundary sequences in pBIN19 means that the *lacZ'* and *kan^r* genes, as well as any new DNA ligated into the restriction sites within *lacZ'*, are transferred to the plant DNA. Recombinant plant cells are selected by plating onto kanamycin agar, and then regenerated into whole plants. Note that pBIN19 is another example of a shuttle vector, recombinant molecules being constructed in *E. coli*, using the *lacZ'* selection system, before transfer to *Agrobacterium tumefaciens* and thence to the plant. For more details, see Brown (2001).

plant callus culture, from which mature transformed plants can be regenerated.

4.3 The Polymerase Chain Reaction (PCR)

In essence, DNA cloning results in the purification of a single fragment of DNA from a complex mixture of DNA molecules. Cloning is a powerful technique and its impact on our understanding of genes and genomes has been immeasurable. Cloning does, however, have one major disadvantage: it is a time-consuming and, in parts, difficult procedure. It takes several days to perform the manipulations needed to insert DNA fragments into a cloning vector and then introduce the ligated molecules into the host cells and select recombinants. If the experimental strategy involves generation of a large clone library, followed by screening of the library to identify a clone that contains a gene of interest (see Technical Note 4.3), then several more weeks or even months might be needed to complete the project.

TECHNICAL

4.4

NOTE

Techniques for studying RNA

Many of the techniques devised for studying DNA molecules can be adapted for use with RNA.

- **Agarose gel electrophoresis** of RNA is carried out after denaturation of the RNA so that the migration rate of each molecule is dependent entirely on its length, and is not influenced by the intramolecular base pairs that can form in many RNAs (e.g. Figure 11.11, page 323). The denaturant, usually formaldehyde or glyoxal, is added to the sample before it is loaded onto the gel.
- **Northern hybridization** refers to the procedure whereby an RNA gel is blotted onto a nylon membrane and hybridized to a labeled probe (see Figure 7.4, page 192). This is equivalent to Southern hybridization and is done in a similar way.
- **Labeled RNA** molecules are usually prepared by copying a DNA template into RNA in the presence of a labeled ribonucleotide. The RNA polymerase enzymes of SP6, T3 or T7 bacteriophages are used because they can produce up to 30 µg of labeled RNA from 1 µg of DNA in 30 minutes. RNA can also be end-labeled by treatment with purified poly(A) polymerase (Section 10.1.2).
- **PCR** of RNA molecules requires a modification to the first step of the normal reaction. *Taq* polymerase cannot copy an RNA molecule so the first step is catalyzed by a reverse transcriptase, which makes a DNA copy of the RNA template. This DNA copy is then amplified by *Taq* polymerase. The technique is called **reverse transcriptase-PCR** or **RT-PCR**. The discovery of thermostable enzymes that make DNA copies of both RNA and DNA templates (e.g. the *Tth*

DNA polymerase from the bacterium *Thermus thermophilus*) raises the possibility of carrying out RT-PCR in a single reaction with just one enzyme.

- **RNA sequencing** methods exist but are difficult to perform and are applicable only to small molecules. The methods are similar to chemical degradation sequencing of DNA (Box 6.2, page 170) but employ sequence-specific endonucleases rather than chemicals to generate the cleaved molecules. In practice, the sequence of an RNA molecule is usually obtained by converting it into cDNA (see Figure 5.32, page 155) and sequencing by the chain termination method (Section 6.1.1).
- **Specialist methods** have been developed for mapping the positions of RNA molecules on to DNA sequences, for example to determine the start and end points of transcription and to locate the positions of introns in a DNA sequence. These methods are described in Section 7.1.2.

The only major deficiency in the RNA toolkit is the absence of enzymes with the degree of sequence specificity displayed by the restriction endonucleases that are so important in DNA manipulations. Other than this, the only drawback with RNA work is the ease with which RNAs are degraded by ribonucleases that are released when cells are disrupted (as during RNA extraction), and which are also present on the hands of laboratory workers and which tend to contaminate glassware and solutions. This means that rigorous laboratory procedures (e.g. cleaning of glassware with chemicals that destroy ribonucleases) have to be adopted in order to keep RNA molecules intact.

PCR complements DNA cloning in that it enables the same result to be achieved – purification of a specified DNA fragment – but in a much shorter time, perhaps just a few hours (Saiki *et al.*, 1988). PCR is complementary to, not a replacement for, cloning because it has its own limitations, the most important of which is the need to know the sequence of at least part of the fragment that is to be purified. Despite this constraint, PCR has acquired central importance in many areas of molecular biology research. We will examine the technique first and then survey its applications.

4.3.1 Carrying out a PCR

PCR results in the repeated copying of a selected region of a DNA molecule (see Figure 4.3, page 98). Unlike cloning,

PCR is a test-tube reaction and does not involve the use of living cells: the copying is carried out not by cellular enzymes but by the purified, thermostable DNA polymerase of *T. aquaticus* (Section 4.1.1). The reason why a thermostable enzyme is needed will become clear when we look in more detail at the events that occur during PCR.

To carry out a PCR experiment, the target DNA is mixed with *Taq* DNA polymerase, a pair of oligonucleotide primers, and a supply of nucleotides. The amount of target DNA can be very small because PCR is extremely sensitive and will work with just a single starting molecule. The primers are needed to initiate the DNA synthesis reactions that will be carried out by the *Taq* polymerase (see Figure 4.6, page 100). They must attach to the target DNA at either side of the segment that is to be copied; the sequences of these attachment sites must

therefore be known so that primers of the appropriate sequences can be synthesized.

The reaction is started by heating the mixture to 94 °C. At this temperature the hydrogen bonds that hold together the two polynucleotides of the double helix are broken, so the target DNA becomes denatured into single-stranded molecules (Figure 4.28). The temperature is then reduced to 50–60 °C, which results in some rejoining of the single strands of the target DNA, but also allows the primers to attach to their annealing positions. DNA synthesis can now begin, so the temperature is raised to 72 °C, the optimum for *Taq* polymerase. In this first stage of the PCR, a set of 'long products' is synthesized from each strand of the target DNA. These polynucleotides have identical 5' ends but random 3' ends, the latter representing positions where DNA synthesis terminates by chance. When the cycle of denaturation–annealing–synthesis is repeated, the long products act as templates for new DNA synthesis, giving rise to 'short products' whose 5' and 3' ends are both set by the primer annealing positions (Figure 4.29). In subsequent cycles, the number of short products accumulates in an exponential fashion (doubling during each cycle) until one of the components of the reaction becomes depleted. This means that after 30 cycles, there will be over 250 million short products

derived from each starting molecule. In real terms, this equates to several micrograms of PCR product from a few nanograms or less of target DNA.

The results of a PCR can be determined in various ways. Usually, the products are analyzed by agarose gel electrophoresis, which will reveal a single band if the PCR has worked as expected and has amplified a single segment of the target DNA (Figure 4.30). Alternatively, the sequence of the product can be determined, using techniques described in Section 6.1.1.

4.3.2 The applications of PCR

PCR is such a straightforward procedure that it is sometimes difficult to understand how it can have become so important in modern research. First we will deal with its limitations. In order to synthesize primers that will anneal at the correct positions, the sequences of the boundary regions of the DNA to be amplified must be known. This means that PCR cannot be used to purify fragments of genes or other parts of a genome that have never been studied before. A second constraint is the length of DNA that can be copied. Regions of up to 5 kb can be amplified without too much difficulty, and longer amplifications – up to 40 kb – are possible using modifications of the standard technique. However, the >100 kb fragments that are needed for genome sequencing projects are unattainable by PCR.

What are the strengths of PCR? Primary among these is the ease with which products representing a single segment of the genome can be obtained from a number of different DNA samples. We will encounter one important example of this in the next chapter when we look at how DNA markers are typed in genetic mapping projects (Section 5.2.2). PCR is used in a similar way to screen human DNA samples for mutations associated with genetic diseases such as thalassemia and cystic fibrosis. It also forms the basis of genetic profiling, in which variations in microsatellite length are typed (see Figure 2.25, page 61).

A second important feature of PCR is its ability to work with minuscule amounts of starting DNA. This means that PCR can be used to obtain sequences from the trace amounts of DNA that are present in hairs, bloodstains and other forensic specimens, and from bones and other remains preserved at archaeological sites. In clinical diagnosis, PCR is able to detect the presence of viral DNA well before the virus has reached the levels needed to initiate a disease response. This is particularly important in the early identification of viral-induced cancers because it means that treatment programs can be initiated before the cancer becomes established.

The above are just a few of the applications of PCR. The technique is now a major component of the molecular biologist's toolkit and we will discover many more examples of its use in the study of genomes as we progress through the remaining chapters of this book.

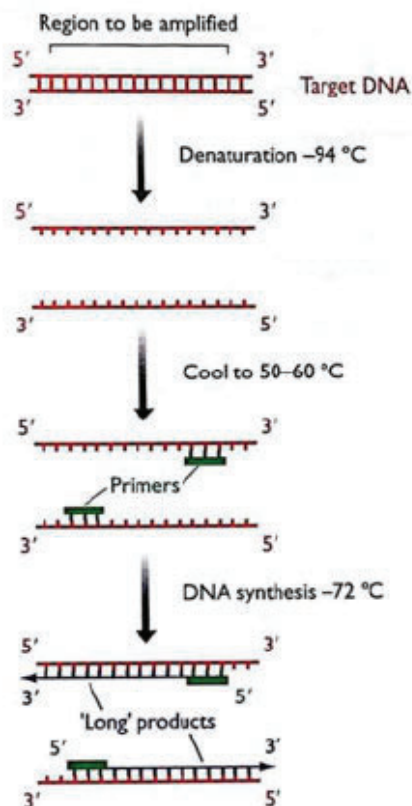


Figure 4.28 The first stage of a PCR.

See the text for details.

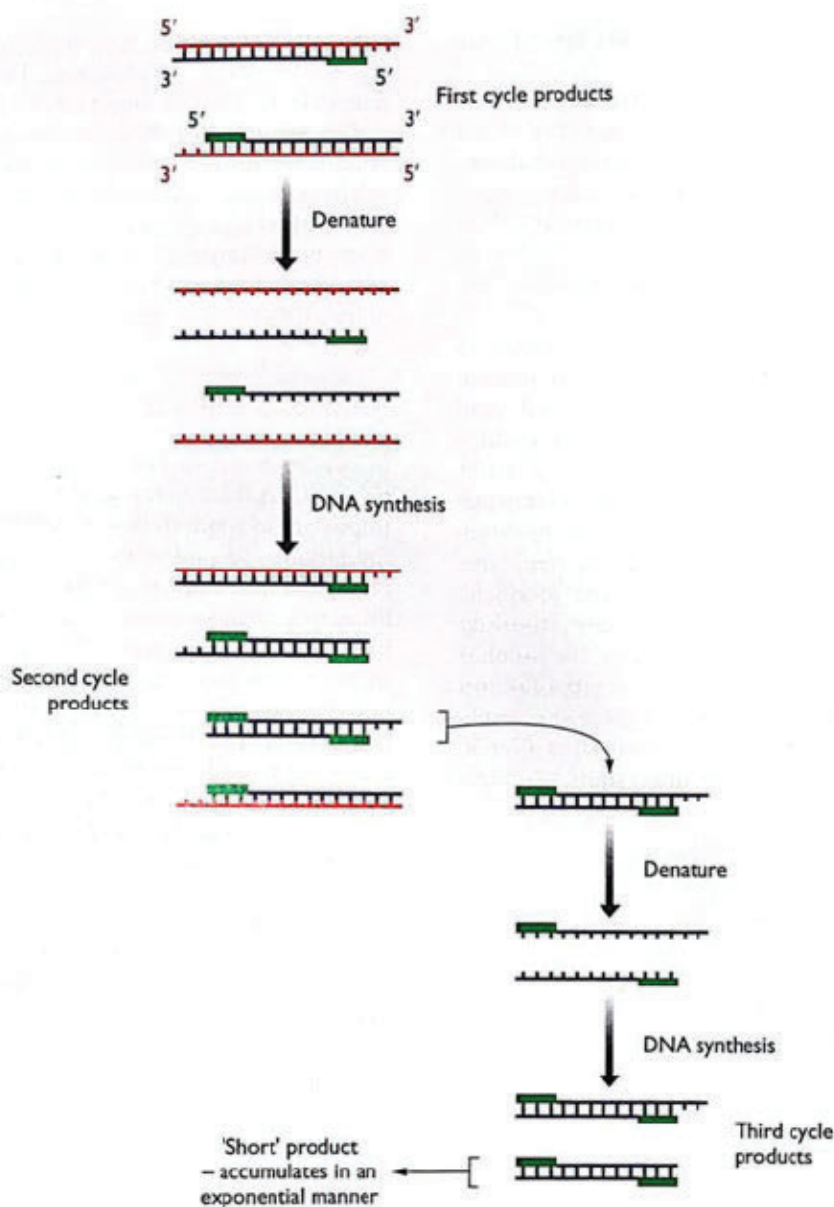


Figure 4.29 The synthesis of 'short' products in a PCR.

The first cycle products from Figure 4.28 are shown at the top. The next cycle of denaturation–annealing–synthesis leads to four products, two of which are identical to the first cycle products and two of which are made entirely of new DNA. During the third cycle, the latter give rise to 'short' products which, in subsequent cycles, accumulate in an exponential fashion.

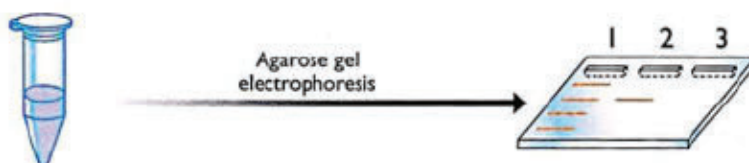


Figure 4.30 Analysing the results of a PCR by agarose gel electrophoresis.

The PCR has been carried out in a microfuge tube. A sample is loaded into lane 2 of an agarose gel. Lane 1 contains DNA size markers, and lane 3 contains a sample of a PCR done by a colleague. After electrophoresis, the gel is stained with ethidium bromide (see Technical Note 2.1, page 37). Lane 2 contains a single band of the expected size, showing that the PCR has been successful. In lane 3 there is no band – this PCR has not worked.

References

- Anderson C** (1993) Genome shortcut leads to problems. *Science*, **259**, 1684–1687.
- Bevan M** (1984) Binary *Agrobacterium* vectors for plant transformation. *Nucleic Acids Res.*, **12**, 8711–8721.
- Bolivar F, Rodriguez RL, Greene PJ, et al.** (1977) Construction and characterisation of new cloning vectors. II. A multi-purpose cloning system. *Gene*, **2**, 95–113.
- Brown TA** (1998) *Molecular Biology Labfax*, 2nd edition, Vol. I. Academic Press, London.
- Brown TA** (2001) *Gene Cloning and DNA Analysis: An Introduction*, 4th edition. Blackwell Scientific Publishers, Oxford.
- Burke DT, Carle GF and Olson MV** (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, **236**, 806–812.
- Hansen G and Wright MS** (1999) Recent advances in the transformation of plants. *Trends Plant Sci.*, **4**, 226–231.
- Hohn B and Murray K** (1977) Packaging recombinant DNA molecules into bacteriophage particles *in vitro*. *Proc. Natl Acad. Sci. USA*, **74**, 3259–3263.
- Ioannou PA, Amemiya CT, Garnes J, et al.** (1994) P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.*, **6**, 84–89.
- Kim U-J, Shizuya H, de Jong PJ, Birren B and Simon MI** (1992) Stable propagation of cosmid and human DNA inserts in an F factor based vector. *Nucleic Acids Res.*, **20**, 1083–1085.
- Klein TM, Wolf ED, Wu R and Sanford JC** (1987) High velocity microprojectiles for delivering nucleic acids into living cells. *Nature*, **327**, 70–73.
- Kornberg A** (1960) Biologic synthesis of deoxyribonucleic acid. *Science*, **131**, 1503–1508.
- Lemoine N and Cooper D** (1998) *Gene Therapy*. BIOS Scientific Publishers, Oxford.
- Mullis KB** (1990) The unusual origins of the polymerase chain reaction. *Sci. Am.*, **262**(4), 56–65.
- Saiki RK, Gelfand DH, Stoffel S, et al.** (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
- Shizuya H, Birren B, Kim UJ, et al.** (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA*, **89**, 8794–8797.
- Sternberg N** (1990) Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc. Natl Acad. Sci. USA*, **87**, 103–107.
- Timmermans MCP, Das OP and Messing J** (1994). Geminiviruses and their uses as extrachromosomal replicons. *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, **45**, 79–112.
- Viaplana R, Turner DS and Covey SN** (2001) Transient expression of a GUS reporter gene from cauliflower mosaic virus replacement vectors in the presence and absence of helper virus. *J. Gen. Virol.*, **82**, 59–65.
- Vieira J and Messing J** (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene*, **19**, 259–268.

Further Reading

- Blackman K** (2001) The advent of genetic engineering. *Trends Biochem. Sci.*, **26**, 268–270. — A personal account of the early days of gene cloning.
- Dale JW** (1998) *Molecular Genetics of Bacteria*, 3rd edition. Wiley, Chichester. — Provides a detailed description of plasmids and bacteriophages.
- Monaco AP and Larin Z** (1994) YACs, BACs, PACs and MACs — artificial chromosomes as research tools. *Trends Biotechnol.*, **12**, 280–286. — A good review of high-capacity cloning vectors.
- Southern EM** (2000) Blotting at 25. *Trends Biochem. Sci.*, **25**, 585–588. — The origins of Southern hybridization.
- Watson JD, Gilman M, Witkowski J and Zoller M** (1992) *Recombinant DNA*, 2nd edition. W.H. Freeman, New York. — Detailed descriptions of basic recombinant DNA methodology.

STUDY AIDS FOR CHAPTER 4

Key terms

Give short definitions of the following terms:

2 µm circle	Complementary DNA (cDNA)
Adaptor	Concatamer
Alkaline phosphatase	cos site
Autoradiography	Cosmid
Bacterial artificial chromosome (BAC)	Denaturation of protein
Bacteriophage P1 vector	DNA cloning
Biolistics	DNA polymerase
Blunt end	End-modification enzyme
Clone library	Endonuclease
Cloning vector	Exonuclease
Cohesive end	Flush end

Fosmid
Gene cloning
Gene therapy
Homologous recombination
Homopolymer tailing
Hybridization
Hybridization probe
In vitro packaging
Insertion vector
Insertional inactivation
Klenow polymerase
Knockout mice
Kornberg polymerase
Ligase

Linker
Lysogenic infection cycle
Lysozyme
Lytic infection cycle
Northern hybridization
Nuclease
Oligonucleotide
Origin of replication
P1-derived artificial chromosome (PAC)
Phosphorimaging
Plaque
Polymerase chain reaction (PCR)

Primer	Sticky end
Proofreading	Stuffer fragment
Protoplast	T4 polynucleotide kinase
Recombinant	T-DNA
Recombinant DNA technology	Template-dependent DNA polymerase
Replacement vector	Template-independent DNA polymerase
Replica plating	Terminal deoxynucleotidyl transferase
Restriction endonuclease	Thermostable
Reverse transcriptase	Ti plasmid
Reverse transcriptase-PCR	Transfection
RNA-dependent DNA polymerase	Transformation
Selectable marker	Yeast artificial chromosome (YAC)
Sequenase	
Shuttle vector	
Southern hybridization	

Self study questions

1. Draw diagrams that outline the events that occur during (a) DNA cloning, and (b) PCR. What are the limitations of each of these two techniques?
2. List the types of enzyme used in recombinant DNA research.
3. Distinguish between the two types of exonuclease activity that can be possessed by a DNA polymerase, and explain how these activities influence the potential applications of individual DNA polymerases in recombinant DNA research.
4. Using examples, describe the various types of end produced after digestion of DNA with a restriction endonuclease.
5. How are agarose gel electrophoresis and Southern hybridization used to examine the results of a restriction digest?
6. Explain why the efficiency of blunt-end ligation is less than that of sticky-end ligation. What steps can be taken to improve the efficiency of blunt-end ligation?

7. Draw diagrams of (a) pBR322, and (b) pUC8. Explain how the differences between these two vectors influence the ways in which they are used to clone DNA fragments.
8. Distinguish between the lytic and lysogenic infection cycles for a bacteriophage.
9. Write a short description of the way in which a bacteriophage λ vector is used to clone DNA. How does a cosmid differ from a standard λ vector?
10. Draw a diagram showing a typical YAC. Indicate the key features and explain how a YAC is used to clone DNA.
11. What problems might arise when a YAC is used to clone a large fragment of DNA? To what extent can these problems be solved by the use of other types of high-capacity cloning vector?
12. How is DNA cloned in organisms other than *Escherichia coli*?
13. Describe how a PCR is carried out, paying particular attention to the role of the primers and the temperatures used during the thermal cycling.

Problem-based learning

1. Soon after the first gene cloning experiments were carried out in the early 1970s, a number of scientists argued that there should be a temporary moratorium on this type of research. What was the basis of these scientists' fears and to what extent were these fears justified?
2. What would be the features of an ideal cloning vector? To what extent are these requirements met by any of the existing cloning vectors?
3. The specificity of the primers is a critical feature of a successful PCR. If the primers anneal at more than one position in the target DNA then products additional to the one being sought will be synthesized. Explore the factors that determine primer specificity and evaluate the influence of the annealing temperature on the outcome of a PCR.