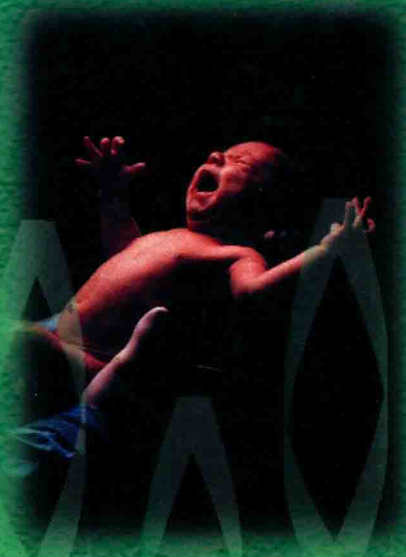
Analyzing DNA

A Laboratory Manual



Editors

Bruce Birren

Eric D. Green

Sue Klapholz

Richard M. Myers

Jane Roskams

GENOME ANALYSIS

VOLUME 1

COLD SPRING HARBOR LABORATORY PRESS

GENOME ANALYSIS

A LABORATORY MANUAL

VOLUME 1

ANALYZING DNA

Volume Editors

Bruce Birren
Whitehead Institute/MIT Center for Genome Research

Eric D. Green
National Human Genome Research Institute

Sue Klapholz
Stanford, California

Richard M. Myers
Stanford University School of Medicine

Jane Roskams
University of British Columbia

COLD SPRING HARBOR LABORATORY PRESS

GENOME ANALYSIS

A LABORATORY MANUAL
VOLUME 1 / ANALYZING DNA

Printed in the United States of America
Book design by Emily Harste
Cover design by Tony Urgo

Library of Congress Cataloging-in-Publication Data

Genome analysis: a laboratory manual / editors Eric D. Green ... [et al.]
v. <1-2 >. cm.
Includes bibliographical references and index
Contents: v. 1. Analyzing DNA — ISBN 0-87969-496-3 (comb). ISBN 0-87969-495-5 (cloth).
v. 2. Detecting genes — ISBN 0-87969-511-0 (comb). ISBN 0-87969-510-2 (cloth).
v. 3. Cloning Systems — ISBN 0-87969-513-7 (comb). ISBN 0-87969-512-9 (cloth).
v. 4. Mapping Genomes — ISBN 0-87969-515-3 (comb). ISBN 0-87969-514-5 (cloth).
1. Gene mapping -- Laboratory manuals. 2. Nucleotide sequence -- Laboratory manuals. I. Green, Eric D.
QH445.2.G4645 1997
576.5'3—dc21

97-17117
CIP

10 9 8 7 6 5 4 3 2

Cover caption for paperback
Top: Photographs of human infant and mouse were reproduced with permission from Comstock, Inc.
Middle: (*Left*) *S. cerevisiae*; (*center*) *C. elegans*; (*right*) *Drosophila*
Bottom: E. coli

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York 11803-2500. Phone: 1-800-843-4388 in Continental U.S. and Canada. All other locations: (516) 349-1946. E-mail: cshpress@cshl.org. For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our World Wide Web site: http://www.cshl.org.

GENOME ANALYSIS

A LABORATORY MANUAL

VOLUME 1

ANALYZING DNA

GENOME ANALYSIS
A LABORATORY MANUAL

VOLUME 1
ANALYZING DNA

VOLUME 2
DETECTING GENES

VOLUME 3
CLONING SYSTEMS

VOLUME 4
MAPPING GENOMES

Series Editors

Eric D. Green
National Human Genome Research Institute

Bruce Birren
Whitehead Institute/MIT Center for Genome Research

Sue Klapholz
Stanford, California

Richard M. Myers
Stanford University School of Medicine

Philip Hieter
University of British Columbia

Preface

Although biologists have studied the genomes of numerous organisms for decades, the last ten years have brought an enormous increase in the pace of genome research. Problems of a previously unthinkable scope are now routinely solved, and the results of these studies are having a profound impact on biomedical research. Many of these advances are the direct result of the coordinated effort of the Human Genome Project, which was established to produce genomic maps and sequences for a set of well-studied organisms. Recent strategic and technological advances have made it feasible to clone, genetically manipulate, and analyze very large segments of DNA; to identify expressed sequences within large genomic regions; to identify DNA sequence variation associated with phenotypic variation; and to determine the nucleotide sequences of DNA cheaper and more efficiently than before. Indeed, to date, several bacterial genomes and that of baker's yeast have been sequenced in their entirety, and the sequencing of other genomes is well under way.

The purpose of this four-volume manual is to provide newcomers and experienced practioners alike with theoretical background, laboratory protocols, and resource materials for applying these powerful new techniques of genome analysis to the study of the very large number of genes and genomes yet to be characterized. We feel that there are several compelling reasons for producing such a manual. First, there are few sources that provide detailed information on the application of methods for genome analysis, due to the recent development of these techniques and the rapid rate at which they have evolved. Second, genome research has led to a new way of thinking that allows a completely different scale of question to be addressed. Many of these techniques, while originally developed for the systematic analysis of mammalian genomes, are now applied to specific biological questions in a variety of organisms. We hope to hasten further the increased understanding of basic biological phenomena by making this technology more widely applied. We have therefore provided step-by-step protocols with detailed explanations regarding why key manipulations are performed as described and which of these steps are most critical to success or prone to failure. We also describe how to recognize and avoid common problems and provide guidelines for troubleshooting various aspects of the protocols. Finally, we believe that a manual that covers the vast array of approaches needed at the different stages of a typical genome analysis project will find application in many laboratories.

This manual consists of chapters authored and edited by genome scientists who are ex-

perts in, and in many cases, the developers of, the described experimental techniques. The editors wrote some of the chapters themselves and worked closely with the other authors and the staff at Cold Spring Harbor Laboratory Press to develop cohesiveness, a consistent style, and substantial cross-referencing. The methods are presented with a high level of detail and completeness to allow readers without experience with these methods to evaluate the strategies and successfully implement the protocols. All of the editors have led laboratory courses at Cold Spring Harbor Laboratory, and thus have an appreciation for the importance of technology transfer and the amount of background explanation and detail that must accompany a protocol to make its use a success, especially in the hands of an inexperienced user. While we have not assumed that the user is experienced in the techniques of genome research, a basic knowledge of molecular biology techniques (i.e., recombinant DNA cloning) is a prerequisite for the successful use of most of the described methods.

Volume 1 of this manual contains seven chapters describing basic techniques in genome analysis that are applicable to most of the experimental methods appearing in subsequent volumes. Chapter 1 provides basic protocols for isolating genomic DNA and performing standard manipulations, such as gel transfer for hybridization analysis. Chapter 2 describes protocols for isolating, manipulating, and analyzing high-molecular-weight DNA, many of which are based on technologies developed specifically in response to the need to clone, map, and sequence large genomes. The polymerase chain reaction is now a standard tool in almost all biomedical research laboratories, and the method has many key roles in genome research. Chapter 3 provides an overview of these PCR-based applications in genome analysis. Chapters 4, 5, and 6 describe large-scale DNA sequencing and include general protocols for dideoxy-mediated sequencing, as well as more specific methods for shotgun and directed sequencing strategies. Chapter 7 reviews perhaps the most critical set of all genome methods—those concerned with analyzing and accessing genomic information, particularly sequence data. This chapter is associated with an electronic version available at http:/www.cshl.org/books/g_a that contains supplemental information that should be useful to all researchers as well as to those interested in using the vast quantities of mapping and sequencing data generated by the genome project. Appendices containing instructions for preparing reagents, protocols for basic methods used throughout the manual, safety information, and useful reference information are included at the end of the volume.

Subsequent volumes of this manual detail an array of other methods for performing genome analysis. In Volume 2, approaches are presented for identifying, isolating, and analyzing genes, including methods for cDNA library construction, gene isolation, and mutation detection. Volume 3 contains methods for using a variety of genomic cloning systems, including cosmids, bacteriophage P1, bacterial artificial chromosomes, and yeast artificial chromosomes. Finally, Volume 4 includes methods central to the generation and use of genomic maps, including the analysis of DNA polymorphisms, meiotic mapping in humans, meiotic and comparative mapping in the mouse, mapping by fluorescence in situ hybridization, generation of PCR-based markers for genome mapping, and radiation hybrid mapping.

We strongly encourage users of this manual to heed all safety cautions noted in the protocols, in the Appendices, and in the instructions provided by manufacturers. We urge all investigators to be familiar with the safe use of reagents and laboratory equipment, as well as with national, state, local, and institutional regulations regarding the use and disposal of materials described in this manual.

The 1970s and 1980s brought the revolution of molecular biology—initially a field of research but more recently a fundamental set of techniques that have come to have key roles in virtually all studies of biological systems. The "genome revolution" of the 1990s will likely evolve in a similar fashion. Our hope is that the experimental techniques described in this four-volume manual will ultimately be useful for investigators focusing on the study of genomes, as well as for those wishing to manipulate and analyze genomes as a means of gaining insight into basic biological processes.

Eric D. Green
Bruce Birren
Philip Hieter
Sue Klapholz
Richard M. Myers

Acknowledgments

The editors and authors would like to thank the following:

Mohammad Ali Ansari-Lari
Simon Ball
Douglas Bassett
Patrick Brown
Ken Dewar
Gaby Dveksler
Joe Ecker
Fulbright Commission
Clare Huxley
Shawn Iadonato
Paul Nisson
Linda Rodgers
Stephanie Smith
Stuart Tugendreich
Bob Waterston

We would also like to acknowledge the staff at Cold Spring Harbor Laboratory Press:

Executive Director: John Inglis

Developmental Editor: Michele Ferguson
Production Editor: Dorothy Brown
Project Coordinator: Mary Cozza

Assistant to the Director: Elizabeth Powers

Book design and formatting: Nancy Ford, Emily Harste, Susan Schaefer

Contributors

Andreas D. Baxevanis *National Human Genome Research Institute, Bethesda*
Bruce Birren *Whitehead Institute/MIT Center for Genome Research, Cambridge*
Jef Boeke *Johns Hopkins University School of Medicine, Baltimore*
Mark S. Boguski *National Center for Biotechnology Information, Bethesda*
Scott E. Devine *Johns Hopkins University School of Medicine, Baltimore*
Séamus Fanning *Regional Technical College, Cork, Ireland and Baylor College of Medicine, Houston*
Robert Gemmill *University of Colorado, Denver*
Richard A. Gibbs *Baylor College of Medicine, Houston*
Andreas Gnirke *Mercator Genetics, Menlo Park, California*
Bruce E. Kimmel *Lawrence Berkeley National Laboratory, Berkeley, California*
Joan Kobori *Pasadena, California*
Elaine R. Mardis *Washington University School of Medicine, St. Louis*
Chris H. Martin *Lawrence Berkeley National Laboratory, Berkeley, California*
B.F. Francis Ouelette *National Center for Biotechnology Information, Bethesda*
Michael J. Palazzolo *Lawrence Berkeley National Laboratory, Berkeley, California*
Harold Riethman *Wistar Institute, Philadelphia*
Richard K. Wilson *Washington University School of Medicine, St. Louis*
Roger Wolff *Mercator Genetics, Menlo Park, California*

Abbreviations and Acronyms

In addition to standard abbreviations for metric measurements (e.g., ml) and chemical symbols (e.g., HCl), the abbreviations and acronyms below are used throughout this manual.

A	adenosine (RNA) or deoxyadenosine (DNA) residue
AHC medium	acid-hydrolyzed casein medium
AMCA	7-amino-4-methylcoumarin-3-acetic acid
amp^r	β-lactamase gene conferring resistance to ampicillin
Amp^r	ampicillin-resistance phenotype
AMV	avian myeloblastosis virus
AP PCR	arbitrarily primed PCR
AT-2	Artificial Transposon-2
ATCC	American Type Culture Collection
ATP	adenosine triphosphate
ATPγS	adenosine-5′-*O*-(3-thiotriphosphate)
AV-FIGE	asymmetric voltage field-inversion gel electrophoresis
BAC	bacterial artificial chromosome
BAP	bacterial alkaline phosphatase
BLAST	basic local alignment search tool
bp	base pair
Bq	Becquerel
BrdU	bromodeoxyuridine
BSA	bovine serum albumin
C	cytidine (RNA) or deoxycytidine (DNA) residue
C. albicans	*Candida albicans*
cam^r	gene conferring resistance to chloramphenicol
C-banding	centromere banding
cDNA	complementary DNA
C. elegans	*Caenorhabditis elegans*

CERN	European Nuclear Research Council
cfu	colony-forming units
CGH	comparative genome hybridization
CHAPS	3-([3-cholamidopropyl]-dimethylammonio)-1-propanesulfonate
CHEF	contour-clamped homogeneous electric field
Ci	Curie
CIP (also known as CIAP)	calf intestinal alkaline phosphatase
cM	centiMorgans
cpm	counts per minute
CTAB	cetyltrimethylammonium bromide
CTP	cytosine triphosphate
DAPI	4′,6-diamidino-2-phenylindole
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
DDBJ	DNA Database of Japan
ddNTP	dideoxynucleoside triphosphate
DD PCR	differential display PCR
DEPC	diethyl pyrocarbonate
dGTP	deoxyguanosine triphosphate
dITP	deoxyinosine triphosphate
DMD	Duchenne muscular dystrophy
D. melanogaster	*Drosophila melanogaster*
DMEM	Dulbecco's modified Eagle's medium
DMF	*N,N*-dimethylformamide
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNMP	deoxynucleoside monophosphate
dNTP	deoxynucleoside triphosphate
DOP PCR	degenerate-oligomer-primed PCR
D-PBS	Dulbecco's phosphate-buffered saline
dpm	disintegrations per minute
DTT	dithiothreitol
dTTP	deoxythymidine triphosphate
dUTP	deoxyuridine triphosphate
EBI	European Bioinformatics Institute
EBV	Epstein-Barr virus
EC number	Enzyme Commission number
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
EEO	electroendosmosis
EGTA	ethylene glycol-bis(β-amino-ethyl ether) *N,N,N′,N′*-tetraacetic acid
ELISA	enzyme-linked immunosorbent assay
E-mail	electronic mail
EMBL	European Molecular Biology Laboratory

EPPS	*N*-(2-hydroxyethyl)piperazine-*N'*-(3-propanesulfonic acid)
EST	expressed sequence tag
EUCIB	European Backcross Collaborative Group
FBS	fetal bovine serum
FCS	fetal calf serum
FIGE	field-inversion gel electrophoresis
FISH	fluorescence in situ hybridization
FITC	fluorescein isothiocyanate
ftp	file transfer protocol
F.W.	formula weight
G	guanosine (RNA) or deoxyguanosine (DNA) residue
G-banding	Giemsa banding
GSS division	GenBank division for genome survey sequences
GTP	guanosine triphosphate
HAP	hydroxyapatite
HBSS	Hanks' balanced salt solution
HEPES	*N*-(2-hydroxyethyl)piperazine-*N'*-(2-ethanesulfonic acid)
HIV	human immunodeficiency virus
HMW DNA	high-molecular-weight DNA
HPLC	high-performance liquid chromatography
HTG division	GenBank division for data from high-throughput genome sequencing centers
HTML	hypertext markup language
http	hypertext transfer protocol
H. wingei	*Hansenula wingei*
IgG	immunoglobulin G
IPTG	isopropylthio-β-D-galactoside
IRS PCR	interspersed-repetitive-sequence-based polymerase chain reaction
kan^r	gene conferring resistance to kanamycin
Kan^r	kanamycin-resistance phenotype
kb	kilobase pair
kD	kilodalton
KGB	potassium glutamate buffer
lb	pound(s)
LB medium/plate	Luria-Bertani medium/plate
LIDS	lithium dodecyl sulfate
LTR	long terminal repeat
M	molar
M13 RF	M13 replicative form
MACAW	multiple alignment construct and analysis workbench
Mb	megabase pair
Mbytes	megabytes

α-MEM	α-minimum essential medium
MES	2-(*N*-morpholino)ethanesulfonic acid
MMLV	Moloney murine leukemia virus
m.o.i.	multiplicity of infection
MOPAC	mixed oligonucleotide-primed amplification of cDNA
MOPS	3-(*N*-morpholino)propanesulfonic acid
mRNA	messenger RNA
m.w.	molecular weight
N	normal
β-NAD	β-nicotinamide adenine dinucleotide
NCBI	National Center for Biotechnology Information
NGM	nematode growth medium
NIGMS	National Institute for General Medical Sciences
NMR	nuclear magnetic resonance
NOR staining	nuclear organizing region staining
NP-40	Nonidet P-40
OFAGE	orthogonal field alternation gel electrophoresis
ORF	open reading frame
PAC	P1-derived artificial chromosome
PACE	programmable autonomously controlled electrodes
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
PFG	pulsed-field gel
PFGE	pulsed-field gel electrophoresis
pfu	plaque-forming unit
PIPES	piperazine-*N*,*N*′-bis(2-ethanesulfonic acid)
PMSF	phenylmethylsulfonyl fluoride
poly(A)$^+$	polyadenosine residues
PRINS labeling	primed in situ labeling
Q-banding	quinacrine banding
QFD-banding	Q-banding by fluorescence using DAPI
r_{avg}	average radius
RACE	rapid amplification of cDNA ends
RAPD	random amplified DNA polymorphism
RARE cleavage	RecA-assisted restriction enzyme cleavage
R-banding	replication or reverse banding
RDA	representational difference analysis
RF	replicative form
RFLP (also known as RFLV)	restriction fragment length polymorphism (variant)
RGE	rotating gel electrophoresis
RI strain	recombinant inbred strain
RNA	ribonucleic acid
RNase	ribonuclease

rpm	revolutions per minute
rRNA	ribosomal RNA
RT-PCR	reverse transcription followed by PCR
S. cerevisiae	*Saccharomyces cerevisiae*
SC medium	synthetic complete medium
SDS	sodium dodecyl sulfate
SINEs	short interspersed repeated DNA sequences
S. pombe	*Schizosaccharomyces pombe*
sq. in.	square inch(es)
SSCP	single-strand conformational polymorphism
SSLP	simple-sequence length polymorphism
ssp.	subspecies
STS	sequence-tagged site
S. typhimurium	*Salmonella typhimurium*
SV40	simian virus 40
T	thymidine (DNA) residue
TAE	Tris-acetate/EDTA
TAFE	transverse alternating field electrophoresis
TAK buffer	Tris-acetate/potassium acetate buffer
Taq DNA polymerase	DNA polymerase from *Thermus aquaticus*
T. aquaticus	*Thermus aquaticus*
TBE	Tris-borate/EDTA
TB medium	Terrific Broth medium
TCA	trichloroacetic acid
TE	Tris/EDTA
TEMED	*N,N,N′,N′*-tetramethylethylenediamine
tet^r	gene conferring resistance to tetracycline
Tet^r	tetracycline-resistance phenotype
TFA	trifluoroacetic acid
T_m	melting temperature
tRNA	transfer RNA
TTP	deoxythymidine triphosphate
U	uridine (RNA) residue
UDG	uracil DNA glycosylase
UNG	uracil-*N*-glycosylase
URL	uniform resource locator
UTP	uridine triphosphate
UV	ultraviolet
VLP	virus-like particle
VNTR	variable number of tandem repeats
V	volt
v/v	volume/volume
w/v	weight/volume
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
YAC	yeast artificial chromosome
YNB	yeast nitrogen base
ZIFE	zero integrated field electrophoresis

Contents

VOLUME 1