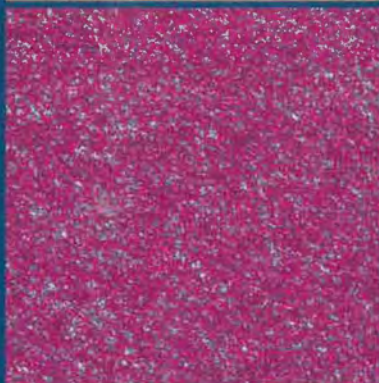
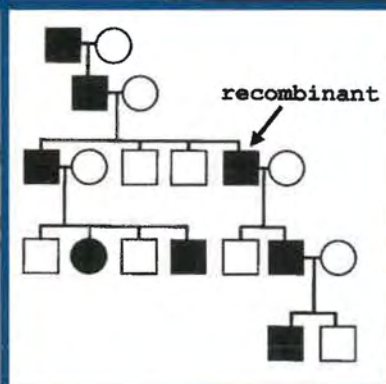
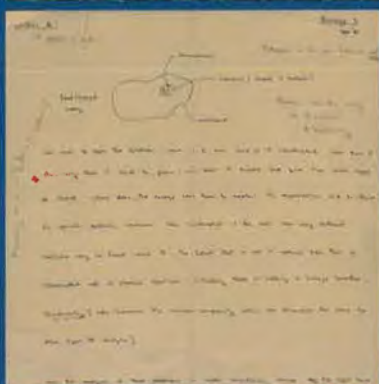
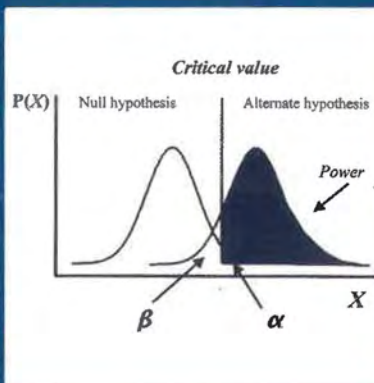


GENETIC VARIATION

A LABORATORY MANUAL



EDITED BY

M.P. Weiner • S.B. Gabriel • J.C. Stephens

Genetic Variation

A LABORATORY MANUAL

ALSO FROM COLD SPRING HARBOR LABORATORY PRESS

RELATED LABORATORY MANUALS

Basic Methods in Microscopy: Protocols and Concepts from Cells—A Laboratory Manual

Gene Transfer: Delivery and Expression of DNA and RNA—A Laboratory Manual

Molecular Cloning: A Laboratory Manual

The Condensed Protocols from Molecular Cloning: A Laboratory Manual

Protein-Protein Interactions: A Molecular Cloning Manual, Second Edition

WEB SITE

www.cshprotocols.org

OTHER RELATED TITLES

At the Bench: A Laboratory Navigator, Updated Edition

Binding and Kinetics for Molecular Biologists

Bioinformatics: Sequence and Genome Analysis, Second Edition

Epigenetics

Experimental Design for Biologists

Lab Math: A Handbook of Measurements, Calculations, and Other Quantitative Skills for Use at the Bench

Lab Ref, Volume 1: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench

Lab Ref, Volume 2: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench

The Dog and Its Genome

Genetic Variation

A LABORATORY MANUAL

EDITED BY

Michael P. Weiner

RainDance Technologies

Stacey B. Gabriel

The Broad Institute of MIT and Harvard

J. Claiborne Stephens

Motif BioSciences Inc.

<http://www.cshprotocols.org>



COLD SPRING HARBOR LABORATORY PRESS

Cold Spring Harbor, New York • <http://www.cshlpress.com>

Genetic Variation
A LABORATORY MANUAL

All rights reserved

© 2007 by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Printed in the United States of America

Publisher

Acquisition Editor and Managing Editor

Development, Marketing, & Sales Director

Developmental Editors

Project Coordinator

Production Manager

Production Editor

Desktop Editor

Sales Account Manager

Cover Designer

John Inglis

Kaaren Janssen

Jan Argentine

Irene Pech, Martin Winer, and Laurie Goodman

Inez Sialiano

Denise Weiss

Patricia Barker

Lauren Heller

Elizabeth Powers

Lewis Agrell

Front cover artwork: (*Top left*) Confocal imaging of cells in mitosis (tubulin stained green, chromosome stained orange), from P. Waadsworth, in *Live cell imaging: A laboratory manual* (ed. R.D. Goldman and D. Spector); (*top center*) diagram redrawn from Chapter 3 by D. Evans and S. Purcell; (*top right*) *Mus musculus*, Gary Henderson; (*middle left*) J.D. Watson lecture note #1 entitled "What is Life," courtesy James D. Watson Collection, Cold Spring Harbor Laboratory Archives; (*middle center*) double helix by Michael P. Weiner; (*middle right*) phylogeny, Michael P. Weiner; (*bottom left*) data from an experiment showing the expression of thousands of genes on a single GeneChip® probe array, image courtesy of Affymetrix; (*bottom center*) Bonobo chimpanzee ©2005 William H. Calvin, University of Washington; (*bottom right*) maize photographs, courtesy Cold Spring Harbor Laboratory Archives.

Library of Congress Cataloging-in-Publication Data

Genetic variation : a laboratory manual / edited by Michael P. Weiner,
Stacey Gabriel, J. Claiborne Stephens.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-87969-780-8 (pbk. : alk. paper) -- ISBN 978-0-87969-779-2

(printed hardcover : alk. paper)

1. Molecular genetics--Laboratory manuals. 2. Variation

(Biology)--Laboratory manuals. I. Weiner, Michael P., 1958- II. Gabriel,

Stacey. III. Stephens, J. Claiborne. IV. Title.

QH440.5.G455 2007

572.8078--dc22

2007023448

10 9 8 7 6 5 4 3 2 1

Students and researchers using the procedures in this manual do so at their own risk. Cold Spring Harbor Laboratory makes no representations or warranties with respect to the material set forth in this manual and has no liability in connection with the use of these materials. All registered trademarks, trade names, and brand names mentioned in this book are the property of the respective owners. Readers should please consult individual manufacturers and other resources for current and specific product information.

With the exception of those suppliers listed in the text with their addresses, all suppliers mentioned in this manual can be found on the BioSupplyNet Web site at <http://www.biosupplynet.com/>.

All World Wide Web addresses are accurate to the best of our knowledge at the time of printing.

Procedures for the humane treatment of animals must be observed at all times. Check with the local animal facility for guidelines.

Certain experimental procedures in this manual may be the subject of national or local legislation or agency restrictions. Users of this manual are responsible for obtaining the relevant permissions, certificates, or licenses in these cases. Neither the authors of this manual nor Cold Spring Harbor Laboratory assumes any responsibility for failure of a user to do so.

The polymerase chain reaction process is covered by certain patent and proprietary rights. Users of this manual are responsible for obtaining any licenses necessary to practice PCR or to commercialize the results of such use. COLD SPRING HARBOR LABORATORY MAKES NO REPRESENTATION THAT USE OF THE INFORMATION IN THIS MANUAL WILL NOT INFRINGE ANY PATENT OR OTHER PROPRIETARY RIGHT.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press, provided that the appropriate fee is paid directly to the Copyright Clearance Center (CCC). Write or call CCC at 222 Rosewood Drive, Danvers, MA 01923 (978-750-8400) for information about fees and regulations. Prior to photocopying items for educational classroom use, contact CCC at the above address. Additional information on CCC can be obtained at CCC Online at <http://www.copyright.com/>.

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 500 Sunnyside Blvd., Woodbury, New York 11797-2924. Phone: 1-800-843-4388 in the Continental U.S. and Canada. All other locations: (516) 422-4100. FAX: (516) 422-4097. E-mail: cshpress@cshl.edu. For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our World Wide Web site <http://www.cshlpress.com/>.

*Our collective gratitude goes out to the investigators who strive
to alleviate human suffering through an increased
understanding of the biological principles of the world around us*

Contents

Foreword, *xiii*

Preface, *xv*

Contributors, *xvii*

- 1 Ethical Issues in Human Genetic Research: The Global Experience, *1*
K. Arnold and J. van der Walt

SECTION 1: STUDY DESIGN

Section 1 covers issues that deal with the importance of study design in addressing the sheer volume of data to be evaluated. Chapter 2 highlights that, for human genetic analyses, certain populations may have more favorable characteristics for associating traits with genes or precise genomic regions (e.g., disease-gene mapping). The size of the sample from a given population is critical for ensuring that a study is adequately powered to find a genetic association when it exists. Chapter 3 considers sample size and all other relevant factors for determining the power of genetic studies. Chapter 4 reviews the analytical methodology available for human genetic studies. In particular, strengths and weaknesses of linkage mapping, where pedigrees play an essential role, are contrasted with genetic association analyses, in which study subjects are typically unrelated. Both types of studies have their place, and the choice of method is strongly dependent on the type of disease or trait being addressed. Chapter 5 gives detailed insights into one of the most important resources available: the dbSNP database, a resource that maintains the current inventory of known variable sites in the human genome. Chapter 5 is an introduction to accessing the multitude of data cataloged in dbSNP. Section 1 closes with Chapter 6, which provides a user's guide to the suite of tools for manipulating HapMap data. The HapMap data set is a unique resource in that it provides an enormous, consistent set of genotypic data (3.9 million SNPs) for 270 individuals from four world populations. These data give very high-resolution detail of the genetic architecture of these four populations and whet our appetite for more samples in these and additional populations!

- 2 Population Choice as a Consideration for Genetic Analysis Study Design, *11*
J.C. Stephens and M. Bamshad
- 3 Power Calculations, *19*
D.M. Evans and S. Purcell
- 4 Genetic Analysis: Moving between Linkage and Association, *31*
A.V. Smith
- 5 NCBI dbSNP Database: Content and Searching, *41*
M.L. Feolo and S.T. Sherry
- 6 Using the HapMap Web Site, *63*
A.V. Smith

SECTION 2: LABORATORY PROTOCOLS

The raw material of genetic research is DNA; individual-to-individual variation is the prism through which it is analyzed. This section deals with two critical aspects in the study of DNA variation: sample preparation and methods of common variation analysis. DNA of sufficient quality and quantity for the current generation of large-scale experimentation is often limiting. To address this, methods of DNA extraction, amplification, and cell culture are described. Parts 2 and 3 of this section turn to the generation of genotype data—beginning with the simplest and most abundant form of sequence variation, single-nucleotide polymorphism, and conclude with more complex DNA variation and the emerging study of DNA copy number variation. The specific topics are not meant to be an exhaustive description of all options for sample preparation and genotyping. Rather, the intent is to sample the wide range of sample type (cell culture and amplified DNA) and experimental scale (small-scale “home brew” genotyping vs. genome-wide standard products) that contribute to current genetic studies.

PART 1: PREPARATION OF DNA AND RNA

- 7 Isolation of Plant DNA for Genotyping Analysis, 87
N.M. Springer
 - Protocol 1: Isolation of Plant DNA for PCR and Genotyping Using Organic Extraction and CTAB, 82
N.M. Springer
 - Protocol 2: DNA Extraction from Freeze-dried Plant Tissue with CTAB in a 96-Well Format, 84
A.-P. Hsia, H.D. Chen, K. Ohtsu, and P.S. Schnable
 - Protocol 3: Purification of *Arabidopsis* DNA in 96-Well Plates Using the PUREGENE DNA Purification Kit, 87
Y. Li
- 8 Preparing RNA from Plant Tissues, 91
A.-P. Hsia, H.D. Chen, K. Ohtsu, and P.S. Schnable
 - Protocol 1: Maize Tissue Preparation and Extraction of RNA from Target Cells, 93
 - Protocol 2: T7-based RNA Amplification from Maize SAM, 97
- 9 Preparing DNA from Mammalian Sources, 103
A. Sahota, A.I. Brooks, and J.A. Tischfield
 - Protocol 1: Preparing DNA from Cell Pellets, 107
A. Sahota, A.I. Brooks, and J.A. Tischfield
 - Protocol 2: Preparing DNA from Fixed Tissue: Extraction and Whole-Genome Amplification, 110
C. Brueck, C. Brown, S. Michalik, D. Vassar-Nieto, E. Mueller, and G. Davis
 - Protocol 3: DNA Isolation from Rat Tail or Ear, 114
E. Cuppen
 - Protocol 4: Preparing DNA from Buccal Cells, 116
I.B. King
 - Protocol 5: Preparing Genomic DNA from Whole Blood: Small- and Mid-Scale Extraction, 120
A. Sahota, A.I. Brooks, J.A. Tischfield, and I.B. King
 - Protocol 6: Preparing DNA from Blood: Large-Scale Extraction, 124
A. Sahota, A.I. Brooks, and J.A. Tischfield
 - Protocol 7: Preparing DNA from Saliva, 129
A. Sahota, A.I. Brooks, and J.A. Tischfield
 - Protocol 8: Whole-Genome Amplification of Genomic DNA Using PCR, 131
C. Brueck, C. Brown, S. Michalik, D. Vassar-Nieto, E. Mueller, and G. Davis
 - Protocol 9: Single-Cell Whole-Genome Amplification, 136
C. Brueck, C. Brown, S. Michalik, D. Vassar-Nieto, E. Mueller, and G. Davis

Protocol 10: Whole-Genome Amplification using Φ 29 DNA Polymerase, 140

N.P. Burt

Protocol 11: DNA Extraction from Forensic Samples Using Chelex, 146

J.M. Butler

Protocol 12: Estimation of DNA Concentration in Forensic Samples Prior to Multiplex PCR Amplification, 148

J.M. Butler

PART 2: SNP VARIATION ANALYSIS

- 10 Intermediate-Throughput Laboratory-Scale Genotyping Solutions, 155

S.J. Macdonald

- 11 Intermediate-Throughput Laboratory-Scale Genotyping Protocols, 171

E. Cuppen, S.J. Macdonald, C. Ha, P.-Y. Kwok, W.B. Barbazuk, A.-P. Hsia, H.D. Chen, Y. Fu, K. Ohtsu, and P.S. Schnable

Protocol 1: Genotyping by Allele-specific Amplification, 172

E. Cuppen

Protocol 2: Genotyping by Dideoxy Resequencing, 174

E. Cuppen

Protocol 3: Oligonucleotide Ligation Assay, 176

S.J. Macdonald

Protocol 4: Template-directed Dye-Incorporation Assay with Fluorescence Polarization Detection, 182

C. Ha and P.-Y. Kwok

Protocol 5: Temperature Gradient Capillary Electrophoresis Assay, 192

A.-P. Hsia, H.D. Chen, and P.S. Schnable

Protocol 6: SNP Mining from Maize 454 EST Sequences, 195

W.B. Barbazuk, S. Emrich, and P.S. Schnable

- 12 Molecular Inversion Probes and Universal Tag Arrays: Application to Highplex Targeted SNP Genotyping, 199

G. Karlin-Neumann, M. Sedova, R. Sapolsky, J. Forman, Y. Wang, M. Moorhead, and M. Faham

Protocol: Targeted SNP Genotyping with MIP and Universal Tag Arrays, 203

- 13 Whole-Genome Genotyping, 213

S.B. Gabriel and M.P. Weiner

PART 3: COPY NUMBER AND COMPLEX VARIATION ANALYSIS

- 14 Comparative Genomic Hybridization to Detect Variation in the Copy Number of Large DNA Segments, 221

I.N. Holcomb and B.J. Trask

- 15 Representational Oligonucleotide Microarray Analysis Detection of Genetic Variation, 233

R. Lucito

- 16 Whole-Genome Sampling Analysis to Detect Copy Number Changes in FFPE Samples, 241

S. Jacobs

- 17 Molecular Inversion Probe Targeted Genotyping: Application to Copy Number Determination, 249

G. Karlin-Neumann, M. Sedova, R. Sapolsky, S. Lin, Y. Wang, M. Moorhead, and M. Faham

- 18 Microsatellite Markers for Linkage and Association Studies, 253

J. Gulcher

SECTION 3: DATA ANALYSIS

Following on the generation of multiple types of genetic data, Section 3 covers multiple approaches to data analysis. With rich reference data sets available, such as the Human HapMap, there are continually evolving approaches for data analysis to extract maximal information from genome variation. These include strategies for SNP selection for maximizing the "return on investment" of SNP genotyping by pinpointing the most informative SNPs to study; approaches to data visualization to simplify complex genotype data sets; methods of analysis to study association of genotype with phenotype; the use of genetic data sets to study evolutionary phenomena such as selection. While the density and breadth of genetic data that are routinely at hand for the investigator continue to expand, the specific approaches described here represent a set of issues and challenges amenable to the current generation of genome-wide SNP data for use in genetic association analysis.

- 19 Considerations for SNP Selection, 263
C. Carlson
- 20 Selection and Evaluation of tag-SNPs Using Tagger and HapMap, 283
P.I.W. de Bakker
- 21 Haploview: Visualization and Analysis of SNP Genotype Data, 289
J.C. Barrett
- 22 Considerations for Copy Number Analysis of FFPE Samples, 295
S. Jacobs
- 23 Assessing Significance in Genetic Association Studies, 303
M.J. Daly
- 24 Assessing Human Variation Data for Signatures of Natural Selection, 309
M. Bamshad and J.C. Stephens

SECTION 4: VARIATION STUDIES IN MODEL ORGANISMS

Section 4 includes an overview of eight different model species: three plant and five animal. The plant chapters include *Arabidopsis*, a traditional model for plant geneticists and biochemists, and two important crop plants (corn and rice). The animal chapters include rodent (mouse and rat), canine (domestic dog), feline (domestic cat), and primate (chimpanzee) species. For each species, the authors have included a description of the genome sequence as it currently exists (at the time of this writing), historical origin of the species, phenotypic variations, association studies, and Internet resources. Several of the chapters include information about the phylogeny of related strains and related species.

- 25 *Arabidopsis*, 317
Y. Li and J.O. Borevitz
- 26 Maize, 325
W.B. Barbazuk, A.-P. Hsia, H.D. Chen, Y. Fu, K. Ohtsu, and P.S. Schnable
- 27 Rice, 335
H. Leung, K.L. McNally, and D. Mackill
- 28 The Mouse, 353
C.M. Wade and M.J. Daly
- 29 The Rat, 361
E. Cuppen, N. Hübner, H.J. Jacob, and A.E. Kwitek
- 30 The Cat, 381
M.J. Lipinski, N. Billings, and L.A. Lyons

- 31 The Dog, 395
K. Lindblad-Toh and E.A. Ostrander
- 32 The Chimpanzee, 409
T.S. Mikkelsen, M.C. Zody, and K. Lindblad-Toh

SECTION 5: INSIGHTS INTO HUMAN VARIATION

Section 5 contains two diverse applications of genomic research to human beings. Chapter 33 illustrates the utility of sex-specific markers, those from mtDNA and the Y chromosome, in the elucidation of human migrations through Polynesia. Chapter 34 gives a thorough overview of the requirements and application of DNA testing to forensics. The section concludes with a third chapter that looks ahead to the future of genome research in humans and other species and to possible societal implications.

- 33 Genealogical Markers: mtDNA and the Y Chromosome, 421
M. Stoneking and M. Kayser
- 34 Forensic DNA Testing, 437
J.M. Butler
- 35 The Human Genome: What Lies Ahead, 453
M.P. Weiner and J.C. Stephens

APPENDIX: Cautions, 457

Index, 461

Foreword

The study of genetic variation has never been more exciting than it is at present, because of recent successes in linking specific variants with phenotypes and common diseases. In the last decade of the 20th century, the human genetics community generated linkage maps and applied these successfully to identify mutations in over 1000 Mendelian disorders. The applications to common diseases such as diabetes, asthma, inflammatory and infectious diseases, and cancer remained elusive until the advent of the human genome sequence that appeared quickly in the first years of the new millennium. This complete sequence made possible the large-scale characterization of human genome variation, culminating in an international consortium that generated the HapMap and new technologies that ranged from extensive analyses of single variants in hundreds of samples to parallel testing of hundreds of thousands of polymorphisms in thousands of samples. Parallel efforts in other organisms, from microbes to plants and animals, are proving just as exciting in linking biological variation with inherited variation. A high-level view across species reveals a richness of genetic diversity within species that is also diverse in its appearance: single-nucleotide differences (i.e., SNPs), nucleotide insertions/deletions, copy number variation, chromosomal rearrangements, and other structural variations.

In preparing a laboratory manual on genetic variation, the editors have assembled a remarkable team and generated a solid foundation for novices and experts alike. A striking feature of the manual is its breadth of scope, with only one of the five sections being focused on the traditional aspects of a laboratory manual, i.e., how to generate data (Section 2), which is rich in practical methods to be used in tomorrow's experiments. Whereas this section contains key information to be used by laboratory staff and trainees, Sections 1 and 3 provide global knowledge on major aspects of genetic research, spanning study design, utilization of relevant databases, bioinformatics tools, statistical analyses, and interpretation. Special consideration was provided to research communities interested in three plant species (*Arabidopsis*, corn, and rice) and five mammalian species (mouse, rat, cat, dog, and chimpanzee).

The tremendous growth of databases of human genetic variation in worldwide populations is boosting the field of population genetics and its use of new data in inferring the evolutionary history of man. Section 3 includes new methods that have been developed to shed light on where and how natural selection acted during human evolution to shape current human genome diversity and, in many cases, has resulted in common variants that affect intermediate biological pathways associated with disease susceptibility.

I challenge students and other readers to take the time to read the introduction, background, key concepts, and messages that are found in all sections of the manual, and not just to open the book to a specific protocol or reference on a particular analysis method. To understand genetic variation, one needs to merge biology, technology, informatics, mathematics, statistics, and evolution. Although today's techniques will surely continue to evolve as quickly

as they have in recent years, the founding principles that are outlined by world-class specialists will have lasting value and will help connect new facts and data sets into rational explanations.

Congratulations to the editors and authors for an outstanding work!

THOMAS J. HUDSON, M.D.
Ontario Institute for Cancer Research

Preface

Nearly two years ago, when we first embarked on this lab manual on genotyping, our first thought was, How long will such a manual be relevant? Would the technology being developed in the universities and commercial entities for the \$1000 genome sequence make genotyping, as we knew it then in late 2005, obsolete? With input from both our scientific colleagues and our editors at Cold Spring Harbor Laboratory Press, we decided to broaden the scope of the manual and place genotyping into the greater context of a complete genome variation study. Because of Cold Spring Harbor Laboratory's illustrious past, we wanted to be sure to include information that would be of interest to researchers in both the agricultural and animal fields.

In planning the sections of this manual, we thought about the kinds of information that might be useful to scientists at all levels of expertise. Therefore, we divided the manual into five sections. The first section describes study design and the tools the reader would need to understand before getting started. The sheer capacity for generating genomic variation data has an addictive effect: The more you see, the more you want. However, sheer volume of data cannot overcome poorly planned studies. In fact, good genotyping data in a bad study design, accompanied by inadequate attention to statistical issues, such as that of multiple comparisons, has the potential to yield seemingly valid results. The contributors to Section 1 have spelled out these issues and more.

Although it is easy to think that any genome study begins with choosing a method for genotyping, we hope that Section 1 shows the reader that a proper genetic study begins with the proper choice of a population and the questions to be addressed. Section 2 of this manual deals with the generation of genotyping and genome variation data. Divided into three parts, this section provides the traditional "recipe and protocol" section that one expects from the CSHL series of laboratory manuals.

In Section 2, Part 1, methods are described for the isolation and preparation of nucleic acids, including both RNA and DNA, from various samples. Chapters 7 and 8 focus on preparation of genetic materials from plants. RNA isolation is included because in many cases it is easier to deal with the less complex transcriptome. Chapter 9 includes 12 protocols for DNA isolation from mammalian sources and includes two procedures for whole-genome amplification, one using the polymerase chain reaction, and the second, rolling circle amplification. Many of these methods are available as kits from commercial suppliers. Where they can, the authors have included home recipes for the kit components and a troubleshooting guide.

Section 2, Part 2, contains methods for low- and intermediate-throughput SNP genotyping. This is the type of genotyping (hundreds to less than several thousand genotypes per day) that one would set up in an average-sized laboratory contemplating one or two medium-sized studies. The last decade has seen an explosion in the technology used for this throughput type of genotyping. Most of these methods can be thought of as having two components: assay and read-out. Example assays include single-base chain extension, TaqMan, oligonucleotide ligation,

tion, and allele-specific amplification. Read-out can be performed using gel and capillary electrophoresis, fluorescence polarization, flow cytometry, etc.

Like DNA sequencing and oligodeoxynucleotide primer synthesis, genotyping is fungible. It seems likely that in the future much of the average laboratory-based genotyping will be performed by outsourcing to separate core facilities, but an understanding of how the genotyping is generated is important in assessing its reliability.

An important aspect of genome variation that is becoming more appreciated as more information is revealed concerns the changes in copy number. The ability to discover and measure changes in chromosome copy number is becoming critical to our understanding of the differences between normal and disease states. We are fortunate to have several chapters included on this important topic.

Section 3 deals with analysis of the data generated and the tools that can be used to help with that analysis, including strategies for SNP selection, as well as genetic association analysis. In Section 4, we tried to give the reader an understanding of the current "state of the organism" for several plant and animal species that have their respective genomes sequenced and analyzed at various levels of completeness. We end with Section 5, describing some insights into human variation.

What's missing? We purposely did not include extensive information about ultra high-throughput (UHT) genotyping. These initiatives are expensive, and the technology is rapidly changing and beyond the scope of this current effort. At one point early in our discussions, we considered the inclusion of a section on one or more of the excellent examples where genetic studies have been used to increase our understanding of a genetic disease, such as cystic fibrosis or diabetes. We opted instead for the section on model organisms, which we felt would be of a broader interest. We debated extensively about including more (and different) model organisms, such as *Drosophila*, porcine, and bovine species; but in the end, we decided that the group we included was enough to give the reader an understanding of the kinds of information available.

We thank those at Cold Spring Harbor Laboratory Press who made this project possible. We appreciate the efforts of the editorial staff at CSHL Press: Inez Sialiano, our project coordinator who helped us to cope with—even avert—natural disasters; the freelance developmental editors Irene Pech, Laurie Goodman, and Martin Winer; and, especially, Kaaren Janssen, for her origination, developmental editing, and project management. We also thank Denise Weiss, production manager; Pat Barker, production editor; and Lauren Heller, desktop editor; for their expertise in production of this work, and John Inglis, Executive Director of the Press, and Jan Argentine, Book Development, Marketing and Sales Director, for their roles in supporting the project.

M.P.W. thanks some of the people who either directly or indirectly helped to inspire him and this initiative, including especially Camille Solbrig, Evan and Alan Weiner, Albert Wexler, Michael Fleming, Ron Yasbin, Stan Zahler, Ross McIntyre, Ed Buss, Ray Wu, Harold Scheraga, Joe Sorge Jr, Jay Short, Mike Luther, Dan Burns, Lee Babiss, Allen Roses, Phil Whitcome, and Jonathan Rothberg.

J.C.S. takes this opportunity to thank all his friends and colleagues who have encouraged and mentored him throughout his career. Because that is a very long list and spans 30 years, he wants to single out those mentors who strove to provide the "big picture," especially as it incorporates population genetics: Professors Wyatt W. Anderson, John Avise, Michael Clegg, Bruce S. Weir, Masatoshi Nei, Wen-Hsiung Li, Ranajit Chakraborty, Frank Ruddle, Ken Kidd, and Stephen J. O'Brien.

S.B.G. thanks family, friends, and colleagues at the Broad Institute for their continual encouragement and support during all endeavors.

Finally, we owe a huge debt of gratitude to our authors. Their timely responsiveness to our comments was greatly appreciated. Their dedication to this effort was probably best exempli-

fied by Dr. H. Leung, who managed to send us his excellent chapter on Rice (on time!) while in the midst of a typhoon that left his Institute bereft of electricity for a week. It was a pleasure to work with all of them.

From Mendel's peas to Darwin's finches to Benzer's bacteriophage to the deciphering of the human genome, there has been a continuum to the study of genetics. Both the understanding of our genetic heritage and the technology used to elicit that understanding have allowed a common lexicon to develop across all sciences. Since the unraveling of its structure more than 50 years ago, the double helix has become much more than a biological set of instructions; it has become a campfire around which all scientists can converse. We hope this manual allows this conversation to continue.

MICHAEL P. WEINER
STACEY B. GABRIEL
J. CLAIBORNE STEPHENS

9

Preparing DNA from Mammalian Sources

Amrik Sahota,¹ Andrew I. Brooks,^{1,2} and Jay A. Tischfield¹

¹*Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082;*

²*Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey—Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635*

INTRODUCTION

The availability of high-quality DNA from a large number of individuals is a prerequisite for the success of genetic variation studies. This requirement has spurred major technological advances in DNA extraction methods. Twenty years ago, large-scale manual extractions took more than 3 days to complete. Large-scale preparations can now be completed within 3 hours using automated extraction platforms—without the need for toxic reagents and often with higher yields than in the past. Our laboratories developed an inorganic, salt precipitation method (Madisen et al. 1987; Lahiri et al. 1992) for DNA extraction from moderate to large sample volumes (3–30 ml of whole blood or 1×10^7 to 2×10^8 cultured cells). This method eliminates the hazards of phenol exposure to laboratory personnel and the environment, and all reagents are prepared in-house. We used this procedure for over 15 years, but, for convenience and increased productivity and reliability, we switched to commercial reagents about 5 years ago. Comparative studies have shown that DNA obtained using the in-house manual extraction procedure is of equal quality to that obtained using com-

INTRODUCTION, 103

- Choice of Tissue, 104
- Choice of Anticoagulant, 104
- Summary of Extraction, 104
- DNA Quantity and Quality, 105
- Protein, RNA, and Other Contaminants, 105
- Digestion with Restriction Enzymes, 105

ACKNOWLEDGMENTS, 106

PROTOCOLS, 107

1. Preparing DNA from Cell Pellets, 107
2. Preparing DNA from Fixed Tissue: Extraction and Whole-Genome Amplification, 110
3. DNA Isolation from Rat Tail or Ear, 114
4. Preparing DNA from Buccal Cells, 116
5. Preparing Genomic DNA from Whole Blood: Small- and Mid-Scale Extraction, 120
6. Preparing DNA from Blood: Large-Scale Extraction, 124
7. Preparing DNA from Saliva, 129
8. Whole-Genome Amplification of Genomic DNA Using PCR, 131
9. Single-Cell Whole-Genome Amplification, 136
10. Whole-Genome Amplification Using Φ 29 DNA Polymerase, 140
11. DNA Extraction from Forensic Samples Using Chelex, 146
12. Estimation of DNA Concentration in Forensic Samples Prior to Multiplex PCR Amplification, 148

REFERENCES, 152

mercial reagents. The drawback of commercial reagents is that laboratory personnel know very little about the chemical composition of the reagents and kits they are using. We now use the in-house procedure as a backup, and we present it here for the benefit of laboratories that may find the purchase of commercial reagents or of automated equipment too costly.

Choice of Tissue

The amount of DNA that will eventually be required for a given project is unpredictable. Thus, we prefer to extract DNA primarily from lymphoblastoid cell lines (LCLs) and secondarily from blood. Human LCLs are established using Epstein-Barr virus (Neitze 1986; Caputo et al. 1991). DNA from these sources has been successfully assayed at multiple genotyping centers using a variety of analytical platforms (Alarcón et al. 2002; Holmans et al. 2004; Prasad et al. 2005; Suarez et al. 2006). With the recent development of two technologies, whole-genome amplification (WGA) and array-based analyses, it has been suggested that large amounts of genomic DNA may not always be necessary (Dean et al. 2002; Langmore 2002; Fathallah-Shaykh 2005; Dunphy 2006; Li et al. 2006; Quackenbush 2006; Reis-Filho et al. 2006; Tomlins et al. 2006).

There are many options for DNA extraction on a small scale (e.g., <0.5 ml of blood, dried blood spots, saliva, buccal cells, mouthwashes) (McEwen and Reilly 1994; Lum and Le Marchand 1998; Feigelson et al. 2001; Garcia-Closas et al. 2001; Heath et al. 2001; Steinberg et al. 2002). Specimens such as saliva, which can be collected using noninvasive procedures, may be viable alternatives to blood (Steinberg et al. 2002; Rylander-Rudqvist et al. 2006). Buccal swabs and mouthwashes have been used in several epidemiology studies, but the DNA yield from these sources, especially buccal swabs, is significantly lower than from saliva samples. Saliva preserved in Oragene solution (DNA Genotek, Ottawa, Ontario, Canada), as described in this chapter, is also easier to transport than mouthwash samples. Protocols for DNA extraction from buccal swabs or mouthwash samples have been documented in several publications (Lum and Le Marchand 1998; Feigelson et al. 2001; Garcia-Closas et al. 2001; Heath et al. 2001; Steinberg et al. 2002).

Choice of Anticoagulant

EDTA is the anticoagulant of choice for blood collection for DNA extractions, since it inhibits DNase activity and does not introduce volume changes. ACD tubes can be used for blood collection, but it is important to fill them to the mark to avoid causing changes in osmolarity, which can affect cell lysis. Heparin should be avoided, as it can bind to DNA during purification and can inhibit Taq polymerase used for PCR. Irrespective of the anticoagulant, the Vacutainer tube should be inverted several times to mix the blood. Blood can be shipped at ambient temperature, but if the delay between collection and extraction is more than 3 days, there will be some degradation of DNA and the yield will be lower than that from fresh blood.

Summary of Extraction

All DNA extraction procedures involve cell lysis, protein removal, and recovery of DNA. An RNA removal step may be included, depending on sample type. When DNA is isolated from whole blood, for example, nonnucleated red cells are first lysed to separate them from white cells. The white cells are then lysed using an anionic detergent, which solubilizes the cellular components. This is done in the presence of a DNA preservative that limits the activity of DNases. Cytoplasmic and nuclear proteins are removed by salt precipitation. Genomic DNA is precipitated with alcohol and dissolved in TE buffer. A portion of the DNA is diluted to a specified concentration, and the stock and diluted samples are stored in small aliquots at -70°C . This avoids repeated freezing and thawing of a large sample, which can lead to precipitate formation. Water is not used for storage because of the risk of acid hydrolysis of DNA. Any contaminants in the water may also cause DNA degradation.

DNA Quantity and Quality

For accurate DNA quantification, absorbance measurements should be taken in the range of 0.1 to 1.0. The sample dilution factor should be adjusted if the absorbance falls outside this range. Also, measurements should be taken in slightly alkaline buffer (10 mM Tris-HCl, 1 mM EDTA [pH 8.0]), since absorbance measurements and hence the $A_{260/280}$ ratio are pH dependent (Willinger et al. 1997). Lower pH can cause DNA degradation and can also result in a lower $A_{260/280}$ ratio and reduced sensitivity to protein contamination. Water should not be used for this purpose because of the wide variability in pH. In addition, the calculation factor ($1A_{260} = 50 \mu\text{g/ml DNA}$) is pH dependent and is not strictly valid when water is used for dilution.

DNA concentrations are routinely determined by UV spectroscopy. This method is cheap, rapid, robust, and semiautomated, but it measures total UV-absorbing material, which may include RNA in the case of DNA isolated from cell lines. (There is very little or no RNA in DNA isolated from blood.) Thus, in some cases, UV measurements may underestimate the amount of DNA. When the amount of DNA is limited, UV absorbance measurements can be made using the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, Delaware). This instrument requires 1 μl of sample, and measurements can be made without dilution.

Automated procedures, such as those based on fluorescence, can also be used for determination of DNA concentration. Fluorescence-based methods such as PicoGreen have much greater sensitivity and specificity than UV-based methods, but the method of choice for a particular sample depends on the concentration range for that sample. Thus, PicoGreen would not be suitable for measuring "stock" DNA concentrations, and UV would be unsuitable for measuring "working" DNA concentrations. Alternatively, quantitative PCR (qPCR) can be used for estimating DNA concentrations where the concentration of the sample is expected to be quite low, as in forensic studies or recovery of DNA from fixed tissue (see Protocol 12).

Electrophoresis in 1% agarose gels shows that DNA isolated using the methods described in this chapter is typically >23 kb in size and is thus suitable for Southern blotting and other molecular biological procedures, as well as PCR (Lahiri et al. 1992). DNA yield depends on the size, type, age, and quality of the starting material. A sample consisting of small cells, such as transformed lymphoblasts, will have a higher cell density and hence a higher DNA content than the same-sized sample consisting of larger cells.

Protein, RNA, and Other Contaminants

Good-quality DNA has a 260/280 ratio between 1.8 and 2.0. In the event a sample has a 260/280 ratio less than 1.6 (indicative of protein contamination), it should be re-precipitated to improve its purity (Manchester 1995, 1996; Glasel 1997). An $A_{260/280}$ ratio above 2.0 is indicative of RNA contamination, but this usually is not a problem with blood samples. RNase A treatment may be especially important when analyzing pooled DNA samples (e.g., for direct estimation of allele frequencies), to ensure that the pool contains equal amounts of each DNA. Any RNA contamination of the DNA will appear as a diffuse band near the dye front in gel electrophoresis. We have found that phenol-chloroform extraction is rarely required for further DNA purification and never since we began using the Gentra automated extraction procedure. PCR amplification shows that DNA extracted from saliva is contaminated with mycoplasma DNA (A. Sahota et al., unpubl.) and contains variable amounts of bacterial DNA, which may interfere with genotyping assays, especially those involving multiplex platforms. These contaminants have not been observed in blood or cell-line DNA.

Digestion with Restriction Enzymes

As a further check on quality, we digest each DNA sample with two different restriction enzymes and determine the extent of digestion with reference to a control DNA sample (cell line K562)

known to be completely digested with these enzymes. We use both a 6-base (EcoRI) and a 4-base (HaeIII) recognition sequence restriction enzyme. These digests give rise to fragment sizes spanning different molecular-weight ranges on agarose gels. Partial digestion with one or both enzymes is readily seen as higher-molecular-weight fragments outside the expected size ranges. Complete digestion with EcoRI also indicates that the input DNA is free of salt contamination, since this enzyme is highly susceptible to variations in salt concentration. Any DNA sample that does not show complete digestion is re-precipitated with high salt and isopropanol, and the digestion process is repeated.

This chapter presents a series of protocols for extraction of DNA from various sources including cell lines, tissues, blood, and saliva. The chapter also includes methods for whole-genome amplification and extraction of DNA for forensic studies. The mention of a product or company name does not imply endorsement by the authors; equally effective products may be available from other companies. The ultimate test of the quality of a DNA sample is analysis in multiple, independent laboratories. DNA prepared using the methods described here has been distributed to numerous laboratories worldwide for genotyping studies without any problems (Alarcon et al. 2002; Holmans et al. 2004; Prasad et al. 2005; Suarez et al. 2006).

ACKNOWLEDGMENTS

We thank the following colleagues for their scientific and technical contributions: Douglas A. Fugman, David A. Toke, Qi Wang, Laura Wilde, David Keller, and Hiep Tran. The work of the Rutgers University Cell and DNA Repository (RUCDR) is supported by grants and contracts from the National Institutes of Health and private organizations. RUCDR serves as the National Cell Repository for the Collaborative on the Genetics of Alcoholism project (NIAAA), Center for Collaborative Studies of Mental Disorders (NIMH), Center for Genetic Studies (NIDA), Genetic Repository (NIDDK), and the Cell and DNA Repository for several private organizations.

Protocol 2

Preparing DNA from Fixed Tissue: Extraction and Whole-Genome Amplification

Chad Brueck, Clyde Brown, Steve Michalik, Deborah Vassar-Nieto, Ernie Mueller, and Gary Davis

Sigma-Aldrich, St. Louis, Missouri 63103

Genomic DNA can be extracted from fixed tissues, although the inherent state of DNA degradation in these samples can present a significant problem in recovery. Whole-genome amplification (WGA) provides the means to amplify and effectively immortalize genomic DNA from fresh, frozen, archived, or chemically treated samples while maintaining genetic integrity. A significant advantage of this approach is its ability to amplify degraded or severely limited amounts of DNA.

One source of such nucleic acids is the potentially data-rich formalin-fixed paraffin-embedded (FFPE) tissues that are obtained during medical evaluation and are often archived for decades. These readily available clinical materials, which often have corresponding clinical and pathological data, are a notoriously difficult source from which to recover usable genomic DNA. There are

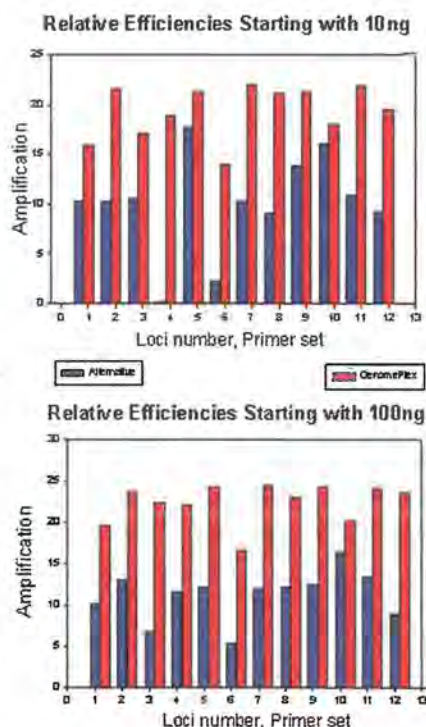


FIGURE 9-1. Results with FFPE samples. This experiment measures the relative amplification efficiencies of both PEP/DOP and MDA when amplifying damaged (FFPE) genomic DNA. The Y-axis is a measurement of 40 minus the Cycle Threshold (Ct) of a dozen gene-specific rat primers. A difference of 10 on this graph is approximately a thousand-fold loss in representation. These data demonstrate that GenomePlex WGA is better suited for DNA that has a high amount of degradation.

several hypotheses regarding DNA damage due to formalin fixation. The most common reasoning is that the DNA is progressively degraded through hydroxymethylation or irreversible cross-linkage of constituent bases. In addition, formaldehyde is known to cross-link proteins directly to DNA through peptide moieties, such as thiol, phenolic, and terminal amino groups. Another hypothesis goes further and suggests that the fixation method aggregates proteins, which subsequently bind DNA, rendering the template inaccessible to DNA polymerases and, hence, making the sequences unrecoverable by most amplification methods.

The GenomePlex WGA system is particularly well suited for these challenging degraded samples. (For a description of GenomePlex WGA strategy, see Protocol 9.) The Tissue WGA kit was formulated to amplify partially degraded genomic DNA from 1 mg of tissue. However, as little as 0.1 mg of tissue has successfully been used with this product. In this approach, FFPE tissue is lysed and genomic DNA is released in a 1-hour proteinase K incubation. During this time, the DNA is partially repaired using both chemical and enzymatic means. Then, an aliquot of lysate is used directly in the WGA reaction. The kit avoids tedious organic solvent extractions to remove excess paraffin, and produces on average 5–10 µg of WGA product. After purification, the WGA product can be analyzed in a manner similar to any other genomic or chromosomal DNA sample (see Fig. 9-1 for quantitative PCR [qPCR] analysis).

This protocol is used for extraction and subsequent amplification of the genomic contents of tissues.

MATERIALS

Reagents

Fresh, frozen, or FFPE tissue (block or slide)
 GenomePlex Tissue Whole Genome Amplification Kit (Sigma), which includes:
 Amplification Master Mix (10x)
 Cellytic Y Lysis Solution
 Library Preparation Buffer
 Library Preparation Enzyme
 Library Stabilization Buffer
 Nuclease-Free H₂O
 Proteinase K Solution
 WGA Polymerase

Equipment

Balance for measuring milligram quantities of tissue
 Pipettes and tips (dedicated)
 Razor blade
 Spectrophotometer
 Thermal cycler (PE 9700 or equivalent)
 Thin-Walled PCR Tubes (0.2 ml or 0.5 ml), or 96-well PCR plates

METHOD

Lysis and Fragmentation

1. Weigh out a 1-mg sample of tissue into a PCR-ready vessel. When working with FFPE tissue, remove excess paraffin with a razor blade.

As little as 0.1 mg of tissue has been successfully used with this kit. The minimum amount of tissue that can be processed should be determined empirically, starting with 1 mg of tissue.

2. To the sample, add 24 μl of CellLytic Y Lysis Solution and 6 μl of Proteinase K Solution.
3. Incubate reaction at 60°C for 1 hour, then heat to 99°C for 4 minutes.
The tissue may not be digested after proteinase K digestion. The GenomePlex Tissue Whole Genome Amplification Kit does not require the tissue to be completely digested. Sufficient DNA is released for WGA without completely digesting the tissue.
4. Immediately cool the sample on ice, and then centrifuge briefly to consolidate the contents.

Library Preparation

5. Combine 9 μl of nuclease-free H_2O and 1 μl of the tissue lysate from Step 4 into a fresh PCR-ready vessel.
6. Add 2 μl of 1x Library Preparation Buffer.
7. Add 1 μl of Library Stabilization Solution.
8. Mix thoroughly and place in thermal cycler at 95°C for 2 minutes.
9. Immediately cool the sample on ice, consolidate by centrifugation, and return to ice.
10. Add 1 μl of Library Preparation Enzyme, vortex thoroughly, and centrifuge briefly.
11. Place sample in a thermal cycler and incubate as follows:
 - i. 16°C for 20 minutes
 - ii. 24°C for 20 minutes
 - iii. 37°C for 20 minutes
 - iv. 75°C for 5 minutes
 - v. 4°C hold
12. Remove sample from thermal cycler and centrifuge briefly. Sample may be amplified immediately or stored at -20°C for 3 days.

Amplification

13. Add the following reagents to the entire 14- μl reaction from Step 12:
 - 7.5 μl of 10x Amplification Master Mix
 - 48.5 μl of Nuclease-Free H_2O
 - 5 μl of WGA DNA Polymerase
14. Vortex thoroughly, centrifuge briefly, and begin thermocycling. The following profile has been optimized for a PE 9700 or equivalent thermal cycler:

Cycle number	Denaturation	Annealing/Extension
1st cycle	2 minutes at 95°C	
20 cycles	15 seconds at 94°C	4 minutes at 65°C

15. After cycling is complete, maintain the reactions at 4°C or store at -20°C until ready for purification and subsequent analysis. Sigma's PCR Cleanup Kit is recommended for WGA DNA purification. The stability of WGA DNA is equivalent to genomic DNA stored under the same conditions. Sample purification is necessary prior to DNA quantitation in order to remove unincorporated primers that will otherwise interfere with some downstream applications.

TROUBLESHOOTING

Problem: The yield is low after cycling.

Solution:

1. The WGA reaction may be inhibited by contaminants in the tissue extract. Dilute inhibitors in the tissue lysate with H₂O by a factor of 1:10 or more prior to continuing with library preparation. Alternatively, too much tissue may have been used. In this case, use less tissue. To test for inhibition, include a DNA control and/or spike 10 ng of purified genomic DNA template into the proteinase K digestion.
2. Less than 0.1–1 mg of tissue may have been weighed. Remove excess paraffin with a razor blade prior to weighing tissue. This ensures that an adequate amount of tissue is used in the subsequent proteinase K digestion.
3. The template may have been of poor quality in FFPE tissues. Formalin fixation may irreversibly damage the DNA if the fixation process is carried out for excessive periods of time, or if the FFPE samples are improperly stored. Obtain properly fixed FFPE tissue samples.
4. Extraction may be insufficient. Incubate samples at 60°C for longer than 1 hour. Samples can be incubated for 2 hours to overnight, followed by incubation at 99°C for 4 minutes.
5. Post-reaction purification may be inappropriate. Use a method that retains single- and double-stranded DNA. Sigma's PCR Cleanup Kit is recommended.
6. The tissue may not be digested after proteinase K digestion. The GenomePlex Tissue Whole Genome Amplification Kit does not require the tissue to be completely digested. Sufficient DNA is released for WGA without completely digesting the tissue.

Problem: qPCR shows significant bias in WGA representation for the gene of interest.

Solution:

1. Controls may be inappropriate. Genomic DNA can only be compared to WGA DNA once the control DNA has been sheared. Use several pooled samples that have been subjected to the fragmentation protocol, or compare against DNA subjected to hydroshearing (Thorstenson et al. 1998).
2. The DNA sample may be limited or degraded. See the solution to the "low yield" problem.

Problem: The negative (no template) control yields a product.

Solution: One or more reagents may have been contaminated with DNA from an outside source. Replace the affected components. (Although this problem may not affect the results, a clean no-template control can only be achieved by replacing the affected components.)

Protocol 5

Preparing Genomic DNA from Whole Blood: Small- and Mid-Scale Extraction

Amrik Sahota,¹ Andrew I. Brooks,^{1,2} Jay A. Tischfield,¹ and Irena B. King³

¹Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082; ²Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635; ³Molecular Diagnostics Program, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

This protocol describes the extraction of genomic DNA from whole blood samples using either the QIAamp DNA Blood Mini Kits (for 200- μ l samples) or the QIAamp DNA Blood Midi Kits (for 0.5-ml samples). It is based on the use of affinity columns that are suitable for many different biological starting materials, with the initial steps being optimized for the DNA yields in the relevant tissues. The procedure is not high-throughput. Potential drawbacks of making it high-throughput are lower recoveries of DNA and possible variability in data quality. Such approaches may require additional quality assurance and control checks. DNA yields for this protocol vary with the number of nucleated cells in the blood sample and the quality of the frozen sample. Expected yields for the small-scale prep are typically 4–12 μ g per 200 μ l of blood and 25–50 μ g per 200 μ l of buffy coat (data from Qiagen). Expected yields from 0.5-ml blood samples from healthy donors are between 8 μ g and 35 μ g. Prior to performing the DNA extraction, it is advisable to become familiar with the QIAamp DNA Blood Mini or Midi Kit handbook. DNA banking must address confidentiality and regulatory issues as specified by institutional review boards, and all blood samples should be handled in accordance with procedures specified in the Hazard Awareness Management Manual (HAMM).

MATERIALS

Reagents

- Agarose gel
- Ethanol, absolute
- Phosphate-buffered saline (PBS, sterile)
- QIAamp DNA Blood Midi Kit solutions (Qiagen) (prepare according to QIAamp DNA Blood Midi Kit instruction handbook)
- Buffer AE (use as supplied)
 - Extra will be needed for DNA samples which are to be normalized (see Step 10iii).*
- Buffer AL (store at room temperature in the dark)
- Buffer AW1 (store at room temperature)
- Buffer AW2 (store at room temperature)
- QIAGEN Protease stock solution (store at 4°C; stable for 2–3 months)
- QIAamp DNA Blood Mini Kit solutions (for small-scale extraction)
- Buffer AE (use as supplied)

The following three buffers are stable for 1 year at room temperature:

Buffer AL (mix thoroughly by shaking before use)

Buffer AW1 (prepare by adding 25 ml of 96–100% ethanol)

Buffer AW2 (prepare by adding 30 ml of 96–100% ethanol)

Protease

Pipette 1.2 ml of the protease solvent (nuclease-free H₂O containing 0.04% sodium azide) into the vial containing lyophilized Qiagen Protease. Store the protease in 20- μ l aliquots at –20°C to prolong the life of the reagent.

Samples to be analyzed:

For small-scale extraction: Use 200 μ l of blood or buffy coat sample.

For mid-scale extraction: Prepare frozen whole blood aliquots from freshly obtained anticoagulated blood and store at –80°C until analysis. EDTA and heparin may interfere with PCR reactions (Garcia et al. 2002). Extract batches of 16–20 samples. Two or more batches can be extracted per day, depending on the schedule.

Equipment

Bar code reader (if using bar code labels) connected to a database

Centrifuge (ambient temperature) equipped to fit 15-ml conical tubes

Conical polypropylene tubes (15 ml, 2 per sample)

Cryoboxes (9 x 9), 81 vials

Equipment for agarose gel electrophoresis

Heating block or water bath preset to 56°C (for small-scale extraction)

Labels (5 per sample, white superior laser print) (Island Scientific)

Microcentrifuge tubes (1.5-ml, DNase-free, 4 per sample)

Micropipettors and tips (100 μ l, 500 μ l, 1 ml)

Plate reader (e.g., SpectraMax 250, Molecular Devices)

Quartz microplates (96 well) for UV reading

QIAamp DNA Blood Midi Kit (100) (for mid-scale extraction; Qiagen)

QIAamp DNA Blood Mini Kit (for small-scale extraction; Qiagen)

Racks for 15-ml and microcentrifuge tubes

Vortex mixer

Water bath, preset to 37°C and 70°C

METHOD 1

Small-Scale DNA Extraction (A. Sahota, A.I. Brooks, and J.A. Tischfield)

1. Dissolve any precipitate in Buffer AL by incubating at 56°C. Add 200 μ l of sample (blood or buffy coat) to a microcentrifuge tube containing 20 μ l of Qiagen Protease (or proteinase K). If the sample volume is less than 200 μ l, add enough PBS to bring the volume to 200 μ l. Mix by pipetting up and down.

Protease is the recommended enzyme for cell lysis, but proteinase K is preferred for tissue samples or when the lysis buffer contains high concentrations of EDTA or SDS.

2. Add 200 μ l of Buffer AL to the sample, vortex for 15 seconds, and incubate in a water bath or heating block at 56°C for 10 minutes. Pulse-centrifuge the tube after incubation to bring the sample to the bottom of the tube. Add 200 μ l of 100% ethanol to the sample and vortex for 15 seconds. Pulse-centrifuge again to remove droplets from inside the lid.

3. Add the mixture (620 μ l) to the QIAamp spin column without wetting the rim. Centrifuge at 8000 rpm (6000g) for 1 minute in a microcentrifuge. Discard the collection tube containing the filtrate into a bottle containing disinfectant (Lysol) and transfer the column to a new 2-ml collection tube. Add 500 μ l of Buffer AW1 without wetting the rim. Centrifuge as above, discard the collection tube containing the filtrate, and transfer the column to a new 2-ml collection tube.
4. Add 500 μ l of Buffer AW1 without wetting the rim. Centrifuge at 8000 rpm for 1 minute. Discard the collection tube containing the filtrate, and transfer the column to a new 2-ml collection tube. Add 500 μ l of Buffer AW2 without wetting the rim. Centrifuge at 14,000 rpm (20,000g) for 3 minutes. Discard the collection tube containing the filtrate and transfer the column to a new collection tube.
5. Centrifuge at 14,000 rpm for 1 minute to remove any buffer carryover. Place the QIAamp spin column into a clean 1.5-ml microcentrifuge tube and discard the collection tube containing any remaining filtrate. Elute the DNA by adding 200 μ l of Buffer AE to the column. Incubate at room temperature for 5 minutes, then centrifuge at 8000 rpm for 1 minute.
6. Quantify the DNA and analyze using PCR. Keep the DNA at 4°C while in use, or at -70°C for long-term storage.

Typical DNA yields are 4–12 μ g per 200 μ l of blood and 25–50 μ g per 200 μ l of buffy coat (data from Qiagen). Qiagen columns copurify DNA and RNA. This is not a problem with blood samples, which contain very little RNA. If RNA-free DNA is required, add 20 μ l of an RNase A solution (20 mg/ml) to the sample prior to the addition of buffer AL.

METHOD 2

Mid-Scale DNA Extraction (I.B. King)

1. To thaw frozen whole blood samples, incubate them for 15 minutes at 37°C, and then keep on ice during setup.
Anticoagulated whole blood aliquots must be thawed rapidly and then kept refrigerated or on ice.
2. Pipette 0.5 ml of whole blood into a 15-ml conical polypropylene centrifuge tube and add 0.5 ml of cold sterile PBS. Add 100 μ l of Qiagen Protease stock solution and 1.2 ml of Buffer AL to the sample. Cap and mix thoroughly by vortexing for 15 seconds. Incubate sample in a water bath at 70°C for 10 minutes.
3. Remove sample from the water bath and add 1 ml of ethanol. Mix again by vortexing for 15 seconds.
4. Carefully transfer all of the solution to the QIAamp Midi column placed in a 15-ml conical tube. Avoid spilling and do not moisten the rim of the QIAamp Midi column. Close the cap and centrifuge at 1850g (about 3000 rpm) for 3 minutes.
Wipe tubes with tissue at each centrifugation step to prevent any cross-contamination during centrifugation.
5. Remove the QIAamp Midi column, discard the filtrate, and place the QIAamp Midi column back into the 15-ml centrifuge tube. Carefully, without moistening the rim, add 2 ml of Buffer AW1 to the QIAamp Midi column. Close the cap and centrifuge at 4500g (about 5000 rpm) for 1 minute.
6. Carefully, without moistening the rim, add 2 ml of Buffer AW2 to the QIAamp Midi column. Close the cap and centrifuge at 4500g (about 5000 rpm) for 15 minutes.
7. Place the QIAamp Midi column in a clean 15-ml centrifugation tube, and discard the collection tube containing the filtrate. Add 100 μ l of Buffer AE, equilibrated to room temperature.

Pipette directly onto the membrane of the QIAamp Midi column and close the cap. Incubate at room temperature for 5 minutes. Centrifuge at 4500g (about 5000 rpm) for 5 minutes.

8. For maximum DNA yield, pipette 100 μ l of fresh Buffer AE, equilibrated to room temperature, onto the membrane of the QIAamp Midi column. Incubate at room temperature for 5 minutes. Close the cap and centrifuge at 4500g (about 5000 rpm) for 5 minutes. Discard the QIAamp Midi column and securely cap the tube.
9. Refrigerate extracted DNA until it is quantified and aliquoted.
10. Once 90 extractions (or 6 batches) are completed, batch the extracted DNA for microplate quantification. Analyze further as follows:
 - i. Perform microplate DNA quantification.
 - ii. Dedicate 50–100 ng of each sample for DNA quality evaluation on an agarose gel. Score the gel visually as follows: 0 = very good, 1 = good, 2 = poor.
 - iii. To prevent freeze/thaw complications, for each sample make 1 aliquot of the parent stock vial and 3 daughter aliquots of normalized vials. The original final volume of the stock vials is based on 150 μ l, and daughter aliquots are arbitrarily normalized to a DNA concentration of 30 ng/ μ l, containing approximately 250 ng of DNA total per aliquot. To do this, add calculated volumes of buffer to suitable concentrations of DNA samples and divide into aliquots.

When original yields are low, do not normalize; i.e., leave at original concentrations if less than 30 ng/ μ l.

11. Prepare the samples for freezing as follows:
 - i. Use preprinted bar code labels and affix them to microcentrifuge tubes prior to filling the tubes with samples. Give each sample a unique number. Cross-check labeled vials against tracking batch forms for errors. Use unique labels for stock solutions.
 - ii. Make sure all microcentrifuge vials are capped securely. Place the aliquots upright in 9 $\times$ 9 cryoboxes, starting at the upper-left corner of the box and placing one vial per box in consecutive order. Place stock and aliquot samples separately and mark clearly on the outside of the boxes. The naming convention for labels can be "Study Name, Type of Sample (whole blood DNA), Volume, and Box Number in Series." Verify vial order with the database.
 - iii. Freeze at -80°C for long-term storage.

DNA extraction, quantification, and normalization tracking forms should accompany the samples at all times, with well-defined quality control checks.

Protocol 6

Preparing DNA from Blood: Large-Scale Extraction

Amrik Sahota,¹ Andrew I. Brooks,^{1,2} and Jay A. Tischfield¹

¹Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082; ²Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635

This protocol describes methods for the large-scale extraction of DNA from blood using either in-house or commercial (PUREGENE) reagents. Also included is a method for DNA extraction from blood samples that are compromised or clotted.

MATERIALS

CAUTION: See Appendix for appropriate handling of materials marked with <!.>

Reagents

Agarose gel (1%)
Chloroform <!.>
Ethanol (70%)
Ethidium bromide <!.>
Isopropanol <!.>
NaCl (0.85%)
NaCl (6 M, saturated solution)
155 mM NH₄Cl <!.>/170 mM Tris (pH 7.7) (prewarmed to 37°C)
Proteinase K (41.7 mg/ml stock solution in H₂O) <!.>
PUREGENE reagents (Gentra Systems) (for commercial reagent method and for compromised or clotted blood sample method)
Cell Lysis Solution (D-50K2)
DNA Hydration Solution (D-50K4)
Glycogen Solution (20 mg/ml, R-5010) (for compromised or clotted blood sample method)
Protein Precipitation Solution (D-50K3)
RBC Lysis Solution (D-50K1)
RNase A (D-50K6) (Optional; see Method 2, Step 3)
Restriction enzymes: EcoRI and HaeIII
RNase A <!.> (DNase-free, 10 mg/ml) (Optional; see Method 1, Step 5)
SDS (20%) <!.>
TE buffer (1x) (10 mM Tris-HCl, 1 mM EDTA [pH 8.0])
TKM #2 lysis buffer (10 mM Tris-HCl, 10 mM KCl, 10 mM MgCl₂ hexahydrate, 2 mM EDTA, 0.4 M NaCl [pH 8.0])
Whole blood collected in EDTA
In-house reagent method: 1–3 tubes, typically 8 ml per tube
Commercial reagent and compromised or clotted blood sample methods: 5–10 ml

Equipment

Absorbent paper (for commercial reagent method)
 Allegra GS-6 centrifuge with rotor GH-3.8 (Beckman), or a similar tabletop centrifuge
 Biological waste bottle containing disinfectant (e.g., Lysol)
 Centrifuge tubes (50-ml conical)
 Cryogenic vials (1-ml and 2-ml, bar coded)
 Fishhook (made from a glass Pasteur pipette)
 Heating block preset to 56°C (optional)
 Ice (for compromised or clotted blood sample method)
 Incubator preset to 56°C
 Orbital shaker
 Rocking platform
 Spectrophotometer
 Transfer pipette (plastic)
 Vortex mixer
 Water bath preset to 37°C

METHOD 1

Large-Scale Extraction Using In-House Reagents

1. Isolate white blood cells (buffy coat) from 1–3 tubes of whole blood collected in EDTA. Centrifuge the tubes at 3000 rpm for 10 minutes in an Allegra GS-6 centrifuge with rotor GH-3.8, or a similar tabletop centrifuge.
2. Using a plastic transfer pipette, transfer the buffy coat layer into a 50-ml conical centrifuge tube. Save the top layer (containing plasma and platelets) at –70°C in case it is needed for (future) biochemical studies. Pool the bottom layer (containing red cells) into one of the blood tubes and store at –20°C or –70°C.

These samples contain sufficient numbers of immature (nucleated) red cells and white blood cells for isolation of small amounts of DNA. Use these to resolve any sample identity problems. The stored samples may be discarded once sample identity is no longer an issue.

3. Add 40 ml of 155 mM NH_4Cl /170 mM Tris (pH 7.7) (prewarmed to 37°C) to the buffy coat layer, mix slowly, and incubate for 7 minutes in a 37°C water bath.

This step lyses the red cells in the buffy coat and releases hemoglobin. The time and temperature are critical. Red cell lysis is not complete at lower temperatures or shorter times. White blood cell lysis can occur if the sample is left to incubate longer, and this can decrease the DNA yield.

4. Centrifuge the tube at 3000 rpm for 10 minutes and discard the supernatant into a biological waste bottle. Resuspend the white cell pellet in 40 ml of 0.85% NaCl, centrifuge again, and discard the supernatant.

This step removes any remaining hemoglobin. The pellet should be off-white at this stage; if it is not, the resulting DNA may need to be purified further after the extraction.

5. Resuspend the pellet in 6.4 ml of TKM #2 lysis buffer, followed by 200 μl of 20% SDS (0.6% final) and 50 μl of proteinase K (300 $\mu\text{g}/\text{ml}$ final). Incubate the tubes at 56°C for 1–2 hours in an orbital shaker. After this point, the samples can be left at room temperature overnight or over the weekend. RNase treatment is usually not necessary for DNA isolated from whole blood. If necessary, before the addition of proteinase K, add 25 μl of DNase-free RNase (10 mg/ml) to give a final concentration of 35 $\mu\text{g}/\text{ml}$, and incubate the sample at 37°C for 30 minutes.

These reagent volumes are for a white cell pellet of approximately 1×10^8 cells, which is typically obtained from about 20 ml of blood with a normal white cell count (4×10^6 to 8×10^6 per ml of blood). Reagent volumes may need to be increased or decreased depending on the pellet size.

6. Add 2.4 ml of 6 M saturated NaCl solution (1.6 M final) to the sample. Vortex the tube for 15 seconds and then allow it to stand at room temperature for 5 minutes to denature and precipitate proteins. Centrifuge at 3000 rpm for 10 minutes and transfer the supernatant to a new 50-ml conical centrifuge tube without disturbing the protein pellet.
7. Add an equal volume of room-temperature isopropanol to the supernatant and invert the tube several times to precipitate the DNA. Using a fishhook (made from a glass Pasteur pipette), pull the DNA strands out of isopropanol and rinse them twice with 70% ethanol. Transfer the tip of the fishhook with the attached DNA into a 2-ml bar-coded cryogenic vial and add one ml of 1x TE buffer. Then add one drop of chloroform to the vial and place it on a rocking platform until the DNA has completely dissolved (several days).
8. Determine the concentration of DNA by taking absorbance measurements at 260 nm and 280 nm in a spectrophotometer. Program the spectrophotometer to automatically calculate the DNA concentration, yield (based on original blood volume), and $A_{260/280}$ ratio, and upload these data into a DNA database.

The DNA yield is in the range of 15–35 µg/ml whole blood (depending on white cell count) and the yield from a 20-ml blood sample is around 480 µg. The $A_{260/280}$ ratio is around 1.9.

9. Check the DNA quality as follows:
 - i. Digest 1 µg of DNA with the restriction enzymes EcoRI and HaeIII (in separate reactions).
 - ii. Perform electrophoresis of the digests on a 1% agarose gel.
 - ii. Stain the gel with ethidium bromide.

There should be very little or no undigested DNA. Run an aliquot of the undigested DNA on the gel as well, to check that the DNA is of high molecular weight (>23 kb).
10. Aliquot the stock DNA into two or more 2-ml bar-coded cryogenic vials and store them at -70°C . To minimize freezing/thawing of stock DNA, prepare dilutions at 100 ng/µl (or another specified concentration) with 1x TE buffer and distribute 300 µl (30 µg of DNA) into each of six 1-ml bar-coded cryogenic vials. Store at -70°C for distribution to genotyping centers, in-house PCR analysis, etc.

METHOD 2

Large-Scale DNA Extraction Using Commercial Reagents

There are many commercial reagent kits for large-scale DNA extractions. After extensive validation studies, we adopted the PUREGENE procedure for large-scale DNA extractions from blood.

1. Transfer 5–10 ml of blood collected in EDTA tubes into a 50-ml conical centrifuge tube. Add RBC Lysis Solution (total volume of blood and lysis solution should equal 40 ml). Gently invert the tube several times to mix the contents. Incubate at room temperature for 7 minutes.
2. Centrifuge the samples in an Allegra GS-6 or similar centrifuge at 3600 rpm for 5 minutes to collect the white blood cells. Pour off the supernatant into a waste beaker. Resuspend the pellet in the residual liquid and add 10 ml of Cell Lysis Solution. Mix several times with a pipette to lyse the cells. Samples can be left in Cell Lysis Solution for at least 2 years.

3. If RNase treatment is required, add 50 μ l of RNase A (4 mg/ml), mix by inverting several times, and incubate the samples at 37°C for 15 minutes. Cool the samples to room temperature and add 3.33 ml of Protein Precipitation Solution. Vortex for 15 seconds to mix the Protein Precipitation Solution with the cell lysate.
4. Centrifuge samples at 3600 rpm for 5 minutes. The precipitated proteins should form a tight, dark pellet. Pour the DNA-containing supernatant into a 50-ml conical centrifuge tube and add 10 ml of isopropanol. Gently invert the tubes 20 times to precipitate the DNA. Centrifuge the samples at 3600 rpm for 5 minutes to collect the DNA. Pour the isopropanol supernatant into a waste container and invert the tube with the DNA pellet onto clean absorbent paper for 60 seconds.
5. Add 10 ml of 70% ethanol to the pellet and invert the tube several times to wash the pellet. Centrifuge the samples at 4200 rpm for 15 minutes to collect the DNA. Pour off the ethanol into a waste container, invert the tube onto clean absorbent paper for 60 seconds, and air-dry it for a few minutes.
6. Dispense 1 ml of DNA Hydration Solution to rehydrate the DNA pellet. Incubate in an orbital shaker for 1 hour at 56°C. Transfer the DNA into a 2-ml cryogenic vial and place on a rocking platform at room temperature until the DNA has completely dissolved (several days).
7. Determine DNA quantity and quality and prepare aliquots for long-term storage (see Steps 8–10 in Method 1).

The typical yield from 10 ml of blood is 300 μ g of DNA (range 150–500 μ g), depending on white cell count. The DNA is of high molecular weight (>23 kb) and is readily digested with a variety of restriction enzymes, including EcoRI, HaeIII, HinDIII, and PstI. The DNA is reported by Gentra to be stable for at least 10 years at –80°C.

METHOD 3

Compromised or Clotted Blood Samples

Compromised blood samples include the following: (1) samples that spent more than 3 days in transit; (2) samples that have been stored at 4°C for more than 1 week; (3) samples that have been stored at –20°C or –70°C; and (4) samples that were not mixed properly upon collection or were collected incorrectly in a red top (clot) tube. For such samples, use the following method, which is based on Gentra Systems protocols.

1. If the compromised sample is frozen, thaw it quickly at 37°C and transfer it to a 50-ml conical centrifuge tube. Make the volume up to 40 ml with RBC Lysis Solution. Mix well and incubate at room temperature for 5 minutes, mixing at least once during the incubation. Centrifuge at 3600 rpm for 5 minutes. Pipette off most of the supernatant, leaving about 4 ml of liquid in the tube. Vortex for 30 seconds to resuspend the cells in the liquid, add 10 ml of Cell Lysis Solution, and vortex again.
2. Incubate at 37°C for at least 2 hours or at room temperature overnight. The sample should be homogeneous after the incubation. Cool the samples on ice for 5 minutes, add 4.5 ml of Protein Precipitation Solution, and vortex for 20 seconds. Centrifuge at 3600 rpm for 10 minutes. Transfer the supernatant containing DNA into a clean 50-ml conical centrifuge tube and add 13.5 ml of isopropanol and 133 μ l of Glycogen Solution (20 mg/ml). Invert the tube several times to mix.
3. Centrifuge at 3600 rpm for 5 minutes to collect the DNA. Discard the isopropanol, drain the tube, and rinse the pellet with 70% ethanol. Centrifuge at 4200 rpm for 15 minutes, discard the ethanol, drain the tube, and air-dry the pellet. Rehydrate the DNA in 500 μ l of DNA Hydration Solution.

4. Use the following modifications of this method for clotted samples:

- i. After adding the RBC Lysis Solution, vortex the sample at high speed and then place it on a rotator at room temperature for 5 minutes to disperse as much of the clot as possible.
- ii. Centrifuge at 3600 rpm for 5 minutes, discard the supernatant, and resuspend the pellet in 10 ml of RBC Lysis Solution. Repeat the vortexing and incubation steps.
- iii. Centrifuge the sample, discard the supernatant, and resuspend the pellet in about 400 μ l of the residual liquid.
- iv. Add 10 ml of Cell Lysis Solution and 50 μ l of Proteinase K Solution (20 mg/ml), vortex, then incubate at 56°C for 2 hours to overnight, until the clot has dissolved completely.
- v. Cool the sample on ice for 5 minutes, add 3.3 ml of Protein Precipitation Solution, and vortex at high speed.
- vi. Centrifuge at 3600 rpm for 10 minutes, cool the sample on ice for 5 minutes, and transfer the supernatant containing the DNA into a clean 50-ml conical centrifuge tube.
- vii. Add 10 ml of isopropanol containing 20 μ l of Glycogen Solution (20 mg/ml).
- viii. Wash with 70% ethanol and rehydrate the DNA in 500 μ l of DNA Hydration Solution.

DNA yield from compromised blood samples is approximately half that from fresh blood, due to the action of DNases and lysis of white cells during freezing and thawing. The yield from clotted samples is even lower.

Comparative Genomic Hybridization to Detect Variation in the Copy Number of Large DNA Segments

Ilona N. Holcomb and Barbara J. Trask

*Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109;
Department of Genome Sciences, University of Washington, Seattle, Washington 98105*

INTRODUCTION

Genome-wide scans of the human genome reveal an extraordinary degree of variation in the copy number of large DNA segments (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005). The variations observed include insertions, deletions, and duplications. A single variant can encompass multiple genes and millions of base pairs of sequence. A segment of >1 kb that is present at a variable copy number in comparison with a reference genome is called a copy number variation (CNV). Classification of CNV frequency and location, on the scale accomplished for single-nucleotide polymorphisms (SNPs), is under way and will reveal the subset of CNVs that are present in >1% of the population (defined as copy number polymorphisms [CNPs] by Feuk et al. 2006). Ultimately, genomic catalogs of CNVs and analysis of their gene content will help us understand how this type of genomic variation contributes to normal phenotypic variation and disease susceptibility. A powerful tool that is capable of providing comprehensive analysis of large-scale variation is called array comparative genomic hybridization (array CGH).

Array CGH scans the genome at thousands of points for copy-number differences between a test and a normal reference DNA. Figure 14-1 shows an overview of array CGH. The test and reference genomes are differentially labeled with fluorescent dyes and compete for hybridization on an array of target spots, with each spot representing a given DNA sequence. The DNA in the target is either spotted or synthesized directly onto the array surface (typically a glass slide), usually in duplicate or triplicate. For each target spot, the ratio of test DNA fluorescence to reference DNA fluorescence is calculated and interpreted as a relative loss (ratio <1), gain (ratio >1), or no change (ratio ≈ 1). The strength of this whole-genome scan is that the exact genomic location in the draft sequence is known for all target spots. Consequently, we know the gene content of the locus rep-

INTRODUCTION, 221
ARRAY CGH PLATFORMS, 223
BAC Arrays, 223
Oligonucleotide Arrays, 225
SNP Arrays, 225
cDNA Arrays, 226
DNA PREPARATION, 226
CNVs IN THE REFERENCE GENOME, 227
HYBRIDIZATION, 227
DATA PROCESSING, 227
EXPLORING THE GENOME FOR CNVs, 228
SUMMARY, 230
REFERENCES, 231
WWW RESOURCES, 232

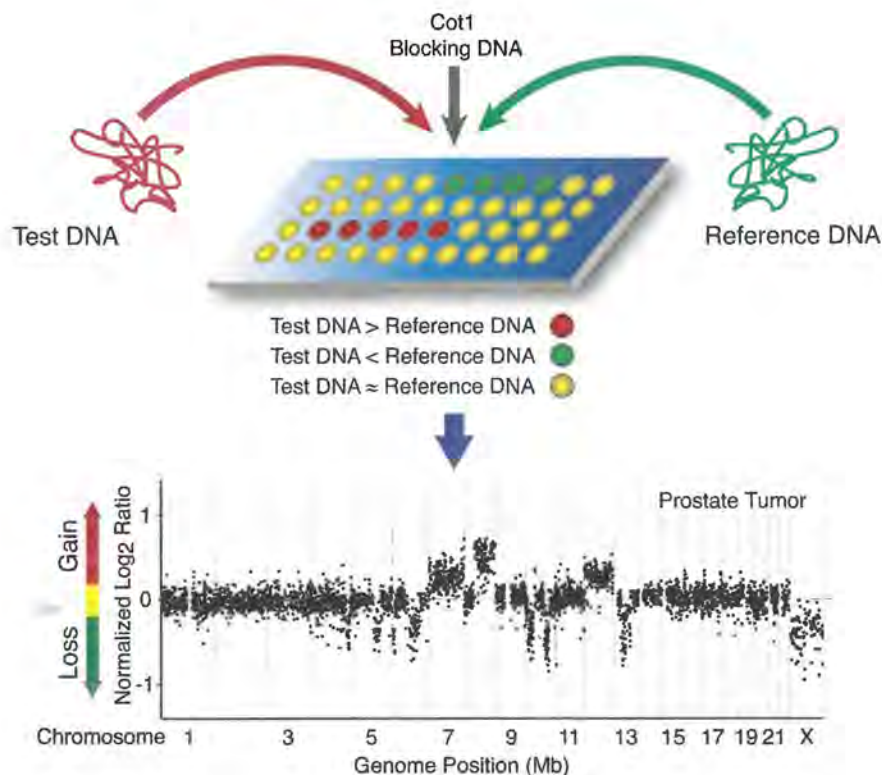


FIGURE 14-1. An overview of array CGH. The test DNA and reference DNA are differentially labeled with dNTP conjugated to fluorescent dyes. The labeled material is then cohybridized in the presence of Cot1 DNA to suppress the hybridization of repetitive elements found in either the labeled sequences or array elements. In this example showing the copy number changes for a prostate tumor, the ratio of fluorescence of test to reference DNA is calculated and log (base 2) transformed for each array element. For the plot shown, the Y-axis is the $\log_2$ ratios and X-axis is the genomic position in Mb separated by chromosome. Deviations significantly below a $\log_2$ ratio of 0 are classified as a relative reduction in the copy number of the test DNA; deviations significantly above a $\log_2$ ratio of 0 are classified as a relative increase in the copy number of the test DNA; and a $\log_2$ ratio close to 0 is classified as no change.

represented by a given target spot; i.e., the region between the previous target and the next target along the chromosome.

The concept of array CGH began with a technique called fluorescence in situ hybridization (FISH), a classic tool of cytogeneticists. In FISH, a fluorescent probe representing 5–200 kb of genomic sequence is hybridized to metaphase or interphase nuclei to ascertain the location(s) and copy number of that sequence. Conventional CGH is a whole-genome adaptation of FISH which employs the same principles that we now use for array CGH, except that the platform on which hybridization occurs in conventional CGH is metaphase chromosomes. Conventional CGH was originally developed to characterize gross copy number changes in the lesions of cancer patients (Kallioniemi et al. 1992). Although conventional CGH was a groundbreaking technology, it is time-consuming, requires the ability to accurately karyotype the human genome, and only crudely resolves the boundaries and locations of the chromosomal abnormalities.

A tremendous advance came with the advent of the first array CGH systems (Solinas-Toldo et al. 1997; Pinkel et al. 1998). These initial systems used the competitive principles of conventional CGH and an array platform composed of thousands of genomic clones. The resulting technology was high-throughput and had a resolution of approximately 1.5 Mb (versus 20 Mb for conventional CGH). Most importantly, every clone used was positioned on the draft of the human genome by a PCR-assayed sequence-tagged site (STS), an end sequence, or its full sequence. Array

CGH quickly became a staple of cancer genomics research and is increasingly used to detect abnormalities in congenital disorders. Now, array CGH is becoming an important tool in the effort to characterize the location, frequency, and gene content of CNVs.

Array CGH is an excellent tool to scan the genome for CNVs when used conscientiously. This chapter is intended to provide an understanding of the basic principles of array CGH and the different options available to the user to design their array CGH experiments. Specifically, the six subsections discuss the different array platforms available, test and reference DNA preparation, reference DNA choice, the basics of hybridization, data processing, and our current understanding of CNVs in the human genome.

ARRAY CGH PLATFORMS

Ideally, one would use the array platform with the highest resolution across the human genome (i.e., the highest number of targets per 3000 Mb) and the greatest sensitivity to copy number changes (e.g., the ability to report single-copy as well as multi-copy changes). However, these two desirable features are currently not united on the same array platform. The relationship between genomic resolution and copy number change is discussed for each of the four platforms covered in this section: bacterial artificial chromosome (BAC), oligonucleotide, SNP, and cDNA arrays. Arrays whose targets are composed of large genomic clones like BACs (100–200 kb of sequence) are very sensitive to genomic changes but typically have low genomic resolution (~1.5 Mb) (Albertson et al. 2000). A newer generation of arrays made up of oligonucleotides (20–100 bp of sequence) has a resolution down to the hundreds and even tens of kilobases, but has not been as sensitive to copy number differences as BAC arrays (Barrett et al. 2004; Carvalho et al. 2004; Sebat et al. 2004). Thus, the benefit of higher resolution is obtained possibly at the expense of sensitivity. Beyond these conceptual issues, practical considerations for platform choice include the minimum amount of test DNA required for array analysis and the relative price of each array. All of these topics for each array CGH platform are discussed in the following paragraphs.

BAC Arrays

Arrays of large genomic clones (mainly BAC clones) were the first array CGH platforms (Solinas-Toldo et al. 1997; Pinkel et al. 1998). A massive effort to sequence and map thousands of BAC clones made them optimal targets for site-specific copy number analysis (Cheung et al. 1999, 2001; Korenberg et al. 1999; Leversha et al. 1999; Kirsch and Ried 2000; Kirsch et al. 2000). Initially, thousands of BACs were sequenced-linked to the draft of the genome by BAC-end sequences or STSs, or because they had been chosen for complete sequencing as part of the Human Genome Project (Cheung et al. 2001; McPherson et al. 2001). Then, FISH mapping by the BAC resource consortium (Cheung et al. 2001) helped discriminate BACs that represented single or multiple locations (e.g., those containing recently duplicated sequences). This multitude of BACs is easily accessible from the BACPAC Resources Center (BPRC) (<http://bacpac.chori.org/>) and from a number of other distribution centers (<http://www.ncbi.nlm.nih.gov/genome/clone/distributors.html>). BAC clones contain 100–200 kb of genomic sequence, 10^4 times longer than the average sequence length of a target on an oligonucleotide array. These large genomic clones typically give more intense signals than shorter array targets, which theoretically result in better sensitivity to copy number changes. Generally, the BACs are spotted at approximately one clone per megabase across the genome. This translates to roughly 3000 array target spots.

The number and availability of well-mapped BACs have driven the production of higher-density arrays for the entire genome or specific regions. An example of the former is a newer generation of BAC arrays, called tiling arrays, which contain overlapping BACs collectively spanning the euchromatic portions of the genome (Ishkanian et al. 2004; Krzywinski et al. 2004; Li et al. 2004). Tiling

arrays are an excellent tool to achieve a high-resolution analysis of a test genome, but the demands of producing and maintaining the approximately 30,000 clones that are required for a full genomic tiling path may limit their widespread use. A tiling array made up of clones from a single chromosome is easier to generate and is a very effective way to get a representative analysis of the whole genome (Woodfine et al. 2004) or to specifically interrogate at high resolution a chromosome that is implicated in a particular disease (Ammérlaan et al. 2005). Another efficient method to interrogate particular regions of the genome that are not covered well by standard BAC arrays is to use a targeted array enriched for the regions of interest. A good example of a targeted array is the segmental duplication (SD) array used by Sharp et al. to study CNVs (Sharp et al. 2005). Most array platforms are designed to exclude SDs, which are large regions (>1 kb) of high sequence identity (>90%) in more than two locations in the genome. Several lines of evidence suggest that CNVs are enriched in regions of the genome containing SDs when compared to other genomic loci (see below, Exploring the Genome for CNVs; Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005). Thus, SD-poor arrays may result in observations of fewer CNVs than arrays that include sequences in or near SDs.

The three benefits of using a BAC array are that they are cost-effective, compatible with considerably less starting material than other platforms, and FISH-verifiable. First, BAC arrays are relatively inexpensive, at about one-third the total setup and processing cost of most high-resolution oligonucleotide-based arrays. Second, amplified material from as little as 10 ng of DNA can produce good data on BAC arrays (Guillaud-Bataille et al. 2004; Little et al. 2006). We have found that BAC arrays are compatible with amplified material from as few as 10 cells, which is 30,000 times less starting material than is needed for most other array platforms (I. Holcomb, unpubl.). Finally, an exceptional feature of BAC arrays is that the very same BACs reporting interesting changes can be used for FISH verification. Figure 14-2 shows an example of the FISH analyses used by Iafrate et al. to confirm copy number differences observed by BAC array (Iafrate et al. 2004). In this figure, they FISH-verify loss, gain, and no change for 3 individuals for the most prevalent CNV in their study (27/55 individuals). They and other workers have also used quantitative PCR to validate their findings.

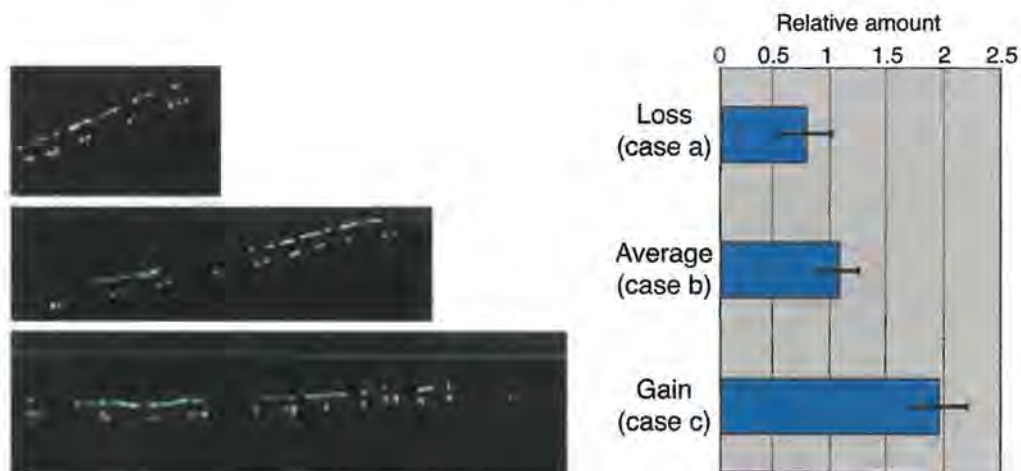


FIGURE 14-2. FISH verification by Iafrate et al. of the most common variant clone (BAC RP11-259N12) detected in their sample population. The copy number differences observed reflect differing numbers of tandem repeats in amylase genes contained within the clone. High-resolution fiber FISH was done on stretched DNA fibers using a Spectrum Green-labeled RP11-259N12 probe cohybridized with a 5' amylase gene probe (green) and a 3' amylase gene probe (red) to verify array CGH profiles (not shown) that indicated relative losses, normal ratios, or relative gains for the CNV represented by BAC RP11-259N12 in three unrelated healthy individuals. The three photos on the left show the FISH results for each case. The case with a relative loss showed 6 gene signals (a), the case with a normal ratio showed 9 gene signals (b), and the case with a relative gain showed 12 signals (c). Quantitative PCR results done on DNA from the same individuals (shown in the panel on the right) were consistent with the array CGH and FISH findings. (Modified, with permission, from Iafrate et al. 2004 [© Macmillan].)

Oligonucleotide Arrays

Oligonucleotide arrays are high-density arrays made up of short (20–100 bp) sequences. The exact location(s) of each sequence is known in the near-finished sequence of the human genome. These arrays are usually synthesized *in situ* directly onto the array surface—a procedure considerably less labor-intensive than maintaining and spotting cloned sequences. The production of an array with exceptionally high genomic resolution is possible due to the short length and ease of production of oligonucleotides. The genomic resolution of current oligonucleotide arrays is 30–50 kb, and efforts are ongoing to produce arrays with even higher resolution. In general, the oligonucleotide array targets are designed to exclude sequences that may decrease the ratio of signal to noise (e.g., SDs, short interspersed elements [SINES], long interspersed elements [LINES], and Alus). Unfortunately, the omission of SDs may decrease the observed numbers of CNVs (see above, BAC Arrays). However, the absence of the highly repetitive elements of the genome in the array targets will not only increase the signal-to-noise ratio, but may also reduce the required amount of the expensive repetitive blocker, Cot1 DNA (see below, Hybridization). To further increase signal over noise, Lucito et al. (2003) developed a method known as representational oligonucleotide microarray analysis (ROMA). ROMA is discussed in greater detail in the following chapter. In brief, ROMA involves a decrease of the complexity of test and reference DNA by BglII digestion and then amplification by ligation-mediated PCR. The array target sequences are designed to complement the amplified restriction fragments, thus increasing the specificity and enhancing the kinetics of the hybridization.

Two classes of oligonucleotide arrays, short (~20-mers) and long (~70-mers), are currently in use. The majority of short oligonucleotide arrays are SNP arrays. Longer oligonucleotide arrays were developed to optimize the balance between the specificity of the short sequences and the sensitivity of arrays with longer sequence targets. Long oligonucleotide arrays that are reported to be capable of detecting single-copy changes are commercially available from NimbleGen (Madison, Wisconsin) and Agilent (Palo Alto, California). Many companies that provide oligonucleotide arrays, including those that provide SNP chips, also make custom arrays that are targeted to specific regions of interest.

SNP Arrays

SNP arrays are a specialized class of short oligonucleotide arrays that were originally developed for genome-wide genotyping. More recently, SNP arrays have become suitable platforms for array CGH. Companies such as Affymetrix (Santa Clara, California) and Illumina (San Diego, California) have developed software that identifies copy gains and losses within the SNP array results. The hybridizations on SNP arrays are not true competitive experiments. To ascertain copy number differences, the hybridization intensities obtained from the test DNA are compared with average values derived from controls hybridized to separate arrays. However, the experimental design and the data obtained from SNP arrays are analogous to those of other array CGH platforms. Over time, the resolution of SNP chips has increased dramatically. The latest versions from Affymetrix (the 500K array) and from Illumina (the HumanHap500 array) have genomic resolutions of approximately 5 kb, 200 times the resolution of the average BAC array. Another advantage of using a SNP array is that this platform reports loss of heterozygosity (LOH) in addition to copy number changes. This information is generally important for cancer studies, because it allows one to distinguish between the loss of information via a deletion (the observation of both copy number loss and LOH) and segmental uniparental disomy (the observation of LOH only). Segmental uniparental disomy is the derivation of both copies of a chromosome segment from a single parent. Limiting factors for SNP arrays are that they require hundreds of nanograms of test DNA and can cost considerably more than other array platforms.

cDNA Arrays

cDNA arrays, long used for gene expression studies, have recently become amenable for array CGH use (Pollack et al. 1999). cDNA arrays are composed of clones representing whole or partial gene sequences obtained from cDNA libraries. They generally contain 10,000–30,000 array target spots. The clear benefit of these arrays is the ability to conduct CGH and expression studies on the same platform. However, one of the significant hurdles to using cDNA arrays for CGH is producing sufficient signal intensity from the array target spots to detect low-level copy number changes. A few laboratories have overcome this limitation (Pollack et al. 1999; Chen et al. 2004; Park et al. 2006), but we anticipate that the improvements in oligonucleotide array technology will obviate the need for cDNA arrays.

DNA PREPARATION

The test and reference DNA must be differentially labeled with fluorescent molecules (typically Cy3 and Cy5) for array analysis. This labeling is achieved by incorporating dye-conjugated dNTPs into sequences produced from the test or reference template by either a linear or exponential amplification. The size range of the labeled sequences must conform to the optimal size range required by the array platform being used. For most platforms, this range is from 100 bp to 1 kb, but users will need to verify the best size range for their platform and optimize their labeling method accordingly.

The quality and quantity of the test and reference DNA have profound effects on the quality of the array data. Starting with DNA that has undergone very little degradation and that is abundant (>200 ng) is the most effective way to reduce noise in the data set. DNA that meets these criteria is generally compatible with any labeling method. The simplest labeling techniques are linear amplifications that directly incorporate the dye-conjugated dNTPs: for example, random priming or nick translation. Random priming employs degenerate hexamers and a Klenow polymerase to amplify the template DNA. Efficient labeling with random priming requires that the template DNA be fully denatured. Thus, partial predigestion, especially of high-molecular-weight DNA, with a restriction enzyme that is a frequent cutter may improve the productivity of the reaction. The size range of the labeled sequences from random priming (100 bp–1 kb) is generally compatible with most array platforms. Nick translation does not use primers, but a DNase to create nick sites in the template DNA at which a polymerase initiates DNA synthesis. Due to the action of the DNase, the size range of the labeled sequences from a nick translation will decrease as the time of the reaction increases. The user may need to check the reaction products on an agarose gel at various time points until the appropriate size range is achieved.

Poor-quality DNA, such as heavily degraded DNA, is unlikely to work for array analysis. However, if the quantity of test DNA is limited (<200 ng), the sample may work with a pre-amplification step before labeling by one of the methods discussed above. Whole-genome PCR and multiple displacement amplification (MDA) methods can preamplify the template DNA. Two popular array-compatible PCR methods are degenerate oligonucleotide primed (DOP) (Telenius et al. 1992; Daigo et al. 2001) and ligation-mediated PCR (LMP) (Guillaud-Bataille et al. 2004). However, PCR-amplified DNA typically cannot be used with commercial oligonucleotide arrays, which usually require high-molecular-weight test DNA. Other labs prefer MDA, a relatively new, non-PCR-based method of amplification. MDA is reported to faithfully reproduce the genome as relatively long (~10-kb) sequence stretches that are compatible with both BAC and oligonucleotide arrays (Paez et al. 2004). MDA employs exonuclease-resistant random primers and DNA polymerase with high processivity (e.g., phi29) to amplify the template DNA. The user may need to test multiple techniques to assess which produces the best array data in his or her hands.

CNVs IN THE REFERENCE GENOME

With array CGH, copy number changes are observed relative to a reference DNA. Thus, CNVs in the reference may influence the frequency and detection of CNVs in the test population. Therefore, it is important to consider whether to use as a reference the genome of one individual or a pool of genomes from multiple individuals. The presence of a rare CNV in a single-genome reference is a potential problem, as it will show up in the majority of array results and lead to an overestimate of the frequency of the CNV. FISH can resolve the source of the CNV; or use of a pooled reference, by dilution of rare CNVs, can mitigate overcounting. The inability to detect CNVs shared between the test and reference genomes is a problem for both types of references. The same CNV in a test and single-genome reference will read as a ratio of 1, implying no change at the locus. The ratio value for a CNV shared by the test DNA and some individuals in the pooled reference will approach 1 (no change) as the number of reference individuals with that CNV increases. Overall, the possibility that the CNV is not detected increases as the number of individuals in the reference with the same CNV increases and as the sensitivity of the array decreases. Ideally, each genome should be tested against sufficiently many single-genome references to accurately predict population frequencies and optimize detection. Because the number of arrays needed to meet these criteria is likely to exceed the scope of most studies, an awareness of these reference effects is required.

HYBRIDIZATION

Test and reference DNA are prepared for hybridization by ridding the labeling reactions of unincorporated nucleotides and coprecipitating the labeled DNA with the repeat blocker, Cot1. Cot1 DNA is a critical component of the hybridization as it suppresses the repetitive content of the human genome (e.g., SINES, LINES, and Alus), which can dominate the hybridization and severely diminish the ability of the array target spots to respond to copy number change. Typically, 50–100 mg of Cot1 DNA is added.

A brief outline of the hybridization process is given here, but specific details are omitted, as the procedure is generally performed by an array facility. The precipitated sample is resuspended in a hybridization solution. This solution includes formamide, salt, and dextran sulfate. The formamide increases the effective hybridization temperature to reduce mismatches without the deleterious effects of actual high temperatures. The salt (e.g., saline sodium citrate [SSC]) stabilizes DNA hybrids by decreasing the effective hybridization temperature. Thus, the appropriate balance of formamide and salt is necessary to achieve proper (low mismatch) hybridization. The dextran sulfate increases the effective concentrations of the test and reference DNAs. Hybridizations are typically incubated overnight at 37°C in a humid chamber. Following hybridization, mismatched sequences are generally removed by washing the array slides in a solution of formamide and dilute salt (e.g., 2x SSC) at approximately 45°C. Lowering the salt concentration, increasing the percentage of formamide, and increasing the temperature increase the stringency of the wash. Finally, a fluorescence scanner is used to collect images of the arrays.

DATA PROCESSING

Processing array output has three steps: visual culling of bad array spots, data normalization, and copy number interpretation. First, scanned array images are inspected for flawed spots (e.g., spots that are overlapping, blank, or scratched). Commercial or publicly available software (e.g., GenePix Pro [Molecular Devices, Sunnyvale, California] and UCSF Spot and Sproc [Jain et al.

2002]) is available to support spot culling and to collate the data for further processing. The second processing step is normalization, a step used to correct for the multiple sources of systematic variations in array experiments. These variations arise from differences in labeling efficiencies, unequal amounts of total test and reference DNA, uneven hybridization across the array platform, and different intensities between the two lasers used in scanning the fluorescent output from the test and reference dyes. The end result of normalization is that the mean or median ratio is set to some standard value. In most cases, this standard value is a ratio of 1 or the log ratio of 1 (0). Commonly used normalization procedures are total intensity normalization, Lowess normalization, log centering, and ratio statistic (Quackenbush 2002). Self vs. self experiments can help identify the normalization scheme that works best for a given set of arrays.

Identifying sites of copy number change is the final step of data processing. The simplest approach is to use threshold values based on mean and standard deviation. However, several more elegant algorithms are available for copy number calling. All these methods attempt to infer the statistical significance of copy number differences and identify the boundaries of these differences. Because the array target sequences representing a particular chromosome and locus share a spatial relationship, most of these methods assess the status of each array target in the context of its neighbors. For example, the hidden Markov model (HMM) assigns spatially related target sequences to a particular state, unless the ratio value of a continuous subset of those target sequences is different enough to overcome a designated penalty value for changing states. HMM states are interpretable as losses, gains, or no change. Lai et al. (2005) provide a discussion of eleven different methods, including the HMMs, mixture models, maximum likelihood, regression, and wavelets. A list of the copy number tools assessed by Lai et al. and links to each algorithm Web site is available at <http://www.chip.org/~ppark/Supplements/Bioinformatics05b.html>.

EXPLORING THE GENOME FOR CNVs

From the number of recent publications, it is becoming clear that large and relatively complex forms of genetic variation are a significant part of the normal human genome. The three seminal papers in this field used competitive arrays with a panel of ethnically diverse individuals to identify a multitude of CNVs (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005). A summary of some of the findings discussed here is given in Table 14-1.

The hundreds of CNVs collectively identified by Sebat et al., Sharp et al., and Iafrate et al. suggest that we have just begun to touch on the large-spectrum CNVs in the human genome. A compilation of all CNVs identified in these three studies is shown in a genome-wide map in Figure 14-3. Sebat et al. (2004) evaluated 20 individuals using the ROMA system to identify 221 array targets with copy number differences representing 76 unique CNVs; this is an average of approximately 11 CNVs per person. Impressively, Sharp et al. (2005) found 53 of the 76 CNVs identified by Sebat et al. In total, the study by Sharp et al. identified 160 CNVs across 47 individuals using a

TABLE 14-1. Summary of Genome-wide Studies of CNVs

Study	Array platform	Sample size	Reference	Total CNVs	Average # per person	Median size (kb)	Number validated
Sebat	ROMA	20	individual	221	11.1	222	11/12
Iafrate	BAC	55	pooled	255	12.4	~150 ^a	18/18
Sharp	BAC	47	individual	160	3.4	~150 ^a	7/9

^aAdapted from Eichler (2006).

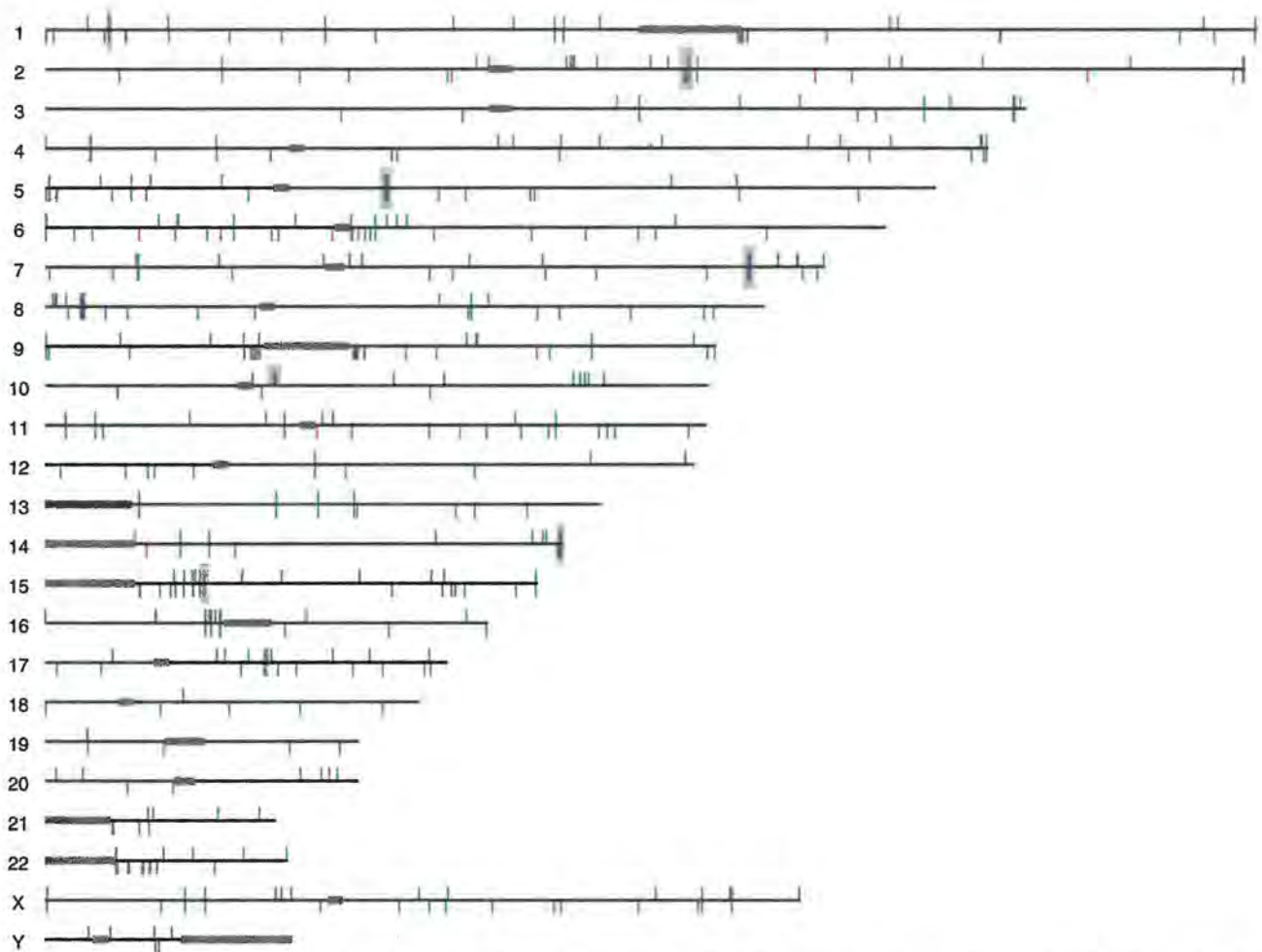


FIGURE 14-3. Genome-wide map of sites of copy number variation detected by Sebat et al., Iafrate et al., and Sharp et al. Colored bars on each chromosome indicate the location of CNVs (not to scale). The CNVs identified by Sebat et al. are in magenta, by Iafrate et al. are in green, and by Sharp et al. are in blue. Tick marks (not to scale) above the chromosome are gains, and those below are deletions. The seven sites that contain overlapping CNVs, irrespective of the type of change (i.e., loss or gain), from all three studies are highlighted by light gray boxes. Five of these seven sites contain CNVs that have concordant changes with at least one concordant CNV from each study. These five sites are on chromosomes 1, 2, 5, 7, and 10. Genome assembly gaps in pericentromeric and satellite regions are indicated by dark gray boxes. The CNV coordinates shown are from the July 2003 assembly and were obtained from <http://humanparalogy.gs.washington.edu/structuralvariation/>.

BAC array targeted toward SDs (3.4 CNVs per person). The large number of novel CNVs (107) identified by Sharp et al. is probably in part a result of their SD-enriched system (CNVs appear to be enriched in SD loci, see below). Surprisingly, however, Iafrate et al. (2004), who used a BAC array system enriched for non-duplicated clones, observed the greatest number of CNVs. In 55 individuals, they identified 255 CNVs (12.4 CNVs per person). Interestingly, 233 of the CNVs observed by Iafrate et al. were novel observations. Only 11 CNVs identified in Iafrate's study overlapped with the Sebat and Sharp studies, respectively. Furthermore, if the direction (i.e., loss or gain) of the variant is disregarded, only 7 sites were identified where CNVs overlapped for all three studies (see Fig. 3). These 7 sites include 5 sites for which the overlapping CNVs are changes in the same direction. The lack of concordance between the CNVs identified in the three studies may result from the diverse array platforms used or may point to valid differences in the CNV content of the individuals tested in the studies. Each study used independent methods to confirm

a subset of the CNVs they found. Verification rates were high. The rate was 78% (7/9) for Sharp et al., 92% (11/12) for Sebat et al., and 100% (18/18) for Iafrate et al.

The cumulative nonoverlapping CNVs identified in these three studies represent approximately 2.7% (88.35 Mb total) of the haploid genome. Sebat et al. (2004) observed an average CNV length of 465 kb and median length of 222 kb using an array composed of 85,000 oligonucleotides, which, of the three studies, is the array with the highest genomic resolution. The majority of CNVs identified by Sharp et al. (2005) and Iafrate et al. (2004) involved a single clone on their respective BAC arrays. A single clone variant can contain all or part of the intervening sequence between it and the neighboring clones, making it more difficult to estimate the actual size of the CNV. For example, the BAC clones on the array used by the Iafrate study are approximately 1 Mb apart. Thus, the gains or losses can involve as much as 2 Mb of DNA. However, an examination of genetic variation trends gives a median CNV size of approximately 150 kb for the Iafrate and Sharp studies (Eichler 2006). All the studies found an approximately equal number of losses and gains of CNVs and that the variants are distributed throughout the genome. However, the distribution of CNVs with respect to blocks of highly homologous SDs reveals an interesting relationship.

All three studies provided support for an enrichment of CNVs in regions that contain SDs, which are considered hot spots of chromosomal rearrangement (Bailey et al. 2002). Sebat et al. (2004) found 6-fold and 12-fold higher SD sequence content within deleted and duplicated CNVs, respectively, relative to the average SD content of the entire genome. Iafrate et al. (2004) found that approximately 25% of the BACs representing CNVs in multiple individuals mapped to regions that overlapped with SDs versus the 7.3% observed for all BAC clones on their array ($p < 0.0001$). Sharp et al. (2005) identified 11.5 times more CNVs by using the SD-enriched BAC array versus a non-SD targeted BAC array. As an increasing number of studies suggest that the duplication architecture of the human genome mediates normal variation (Bailey et al. 2002; Sharp et al. 2005; de Bustos et al. 2006; Goidts et al. 2006; Redon et al. 2006), it becomes interesting to postulate a causal relationship between SDs and CNVs. However, resolving causality is tricky. What can be resolved is the nature of any preferential association of CNVs with SDs by a thorough investigation of the regions of the genome rich in SDs. These regions include pericentromeric and subtelomeric zones, regions underrepresented in the current CNV surveys.

Importantly, a subset of CNVs contains genes that could influence normal phenotypic variation and disease susceptibility. In total, Sebat et al. observed copy number variation of 70 genes, Sharp et al. found 141 genes either completely or partially overlapping their variant BACs, and Iafrate et al. identified 67 BAC clones that encompassed one or more entire genes. The CNV-related genes found by Sharp et al. (2005) include some involved in immunity and metabolism, which may significantly contribute to phenotypic variations, such as resistance to pathogens and metabolic rates. Both Iafrate et al. (2004) and Sebat et al. (2004) identified a subset of CNVs containing cancer-related genes in apparently normal individuals. Seemingly, these genes are not the direct cause of disease, but the structural variants may initiate chromosomal rearrangements or influence gene expression, which could affect tumorigenesis. Overall, the genes found within or near CNVs may have profound effects on the biology of normal or disease processes.

SUMMARY

From our review of the current literature, it appears that CNVs are a prominent part of the normal variation of the human genome. Two extremely useful resources that summarize the results of CNV research studies are found on the Web at <http://humanparalogy.gs.washington.edu/structuralvariation/> and <http://projects.tcag.ca/variation/>. However, the process of cataloging the full spectrum of CNVs is in its infancy. We have yet to discover the exact size range of, the amount of

genomic sequence subject to, and gene content of CNVs. Although by definition CNVs can encompass as little as 1 kb, some may involve as much as 2 Mb of sequence (Iafrate et al. 2004). So far, potentially 88 Mb of genomic sequence is subject to copy number variation (Eichler 2006), but this value is likely to increase as additional studies emerge. Finally, we know that CNVs can involve multiple genes (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005), but we have yet to determine whether gene-containing variants or CNVs in general have an effect on normal phenotypic variation and disease susceptibility. To fully assess and curate human genetic diversity, we need to analyze CNVs on a scale comparable to that used to catalog the wealth of human SNPs, and array CGH is an essential tool in accomplishing that goal.

REFERENCES

- Albertson D.G., Ylstra B., Segraves R., Collins C., Dairkee S.H., Kowbel D., Kuo W.L., Gray J.W., and Pinkel D. 2000. Quantitative mapping of amplicon structure by array CGH identifies *CYP24* as a candidate oncogene. *Nat. Genet.* **25**: 144–146.
- Ammerlaan A.C., de Bustos C., Ararou A., Buckley P.G., Mantripragada K.K., Verstegen M.J., Hulsebos T.J., and Dumanski J.P. 2005. Localization of a putative low-penetrance ependymoma susceptibility locus to 22q11 using a chromosome 22 tiling-path genomic microarray. *Genes Chromosomes Cancer* **43**: 329–338.
- Bailey J.A., Gu Z., Clark R.A., Reinert K., Samonte R.V., Schwartz S., Adams M.D., Myers E.W., Li P.W., and Eichler E.E. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.
- Barrett M.T., Scheffer A., Ben-Dor A., Sampas N., Lipson D., Kincaid R., Tsang P., Curry B., Baird K., Meltzer P.S., et al. 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci.* **101**: 17765–17770.
- Carvalho B., Ouwerkerk E., Meijer G.A., and Ylstra B. 2004. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J. Clin. Pathol.* **57**: 644–646.
- Chen Q.R., Bilke S., Wei J.S., Whiteford C.C., Cenacchi N., Krasnoselsky A.L., Greer B.T., Son C.G., Westermann F., Berthold F., et al. 2004. cDNA array-CGH profiling identifies genomic alterations specific to stage and MYCN-amplification in neuroblastoma. *BMC Genomics* **5**: 70.
- Cheung V.G., Dalrymple H.L., Narasimhan S., Watts J., Schuler G., Raap A.K., Morley M., and Bruzel A. 1999. A resource of mapped human bacterial artificial chromosome clones. *Genome Res.* **9**: 989–993.
- Cheung V.G., Nowak N., Jang W., Kirsch I.R., Zhao S., Chen X.N., Furey T.S., Kim U.J., Kuo W.L., Olivier M., et al. (BAC Resource Consortium). 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**: 953–958.
- Daigo Y., Chin S.F., Gorringe K.L., Bobrow L.G., Ponder B.A., Pharoah P.D., and Caldas C. 2001. Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon profiling of breast cancers: A new approach for the molecular analysis of paraffin-embedded cancer tissue. *Am. J. Pathol.* **58**: 1623–1631.
- de Bustos C., Diaz de Stahl T., Piotrowski A., Mantripragada K.K., Buckley P.G., Darai E., Hansson C.M., Grigelionis G., Menzel U., and Dumanski J.P. 2006. Analysis of copy number variation in the normal human population within a region containing complex segmental duplications on 22q11 using high-resolution array-CGH. *Genomics* **88**: 152–162.
- Devries S., Nyante S., Korkola J., Segraves R., Nakao K., Moore D., Bae H., Wilhelm M., Hwang S., and Waldman F. 2005. Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors. *J. Mol. Diagn.* **7**: 65–71.
- Eichler E.E. 2006. Widening the spectrum of human genetic variation. *Nat. Genet.* **38**: 9–11.
- Feuk L., Carson A.R., and Scherer S.W. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* **7**: 85–97.
- Goidts V., Cooper D.N., Armengol L., Schempp W., Conroy J., Estivill X., Nowak N., Hameister H., and Kehrer-Sawatzki H. 2006. Complex patterns of copy number variation at sites of segmental duplications: An important category of structural variation in the human genome. *Hum. Genet.* **120**: 270–284.
- Guillaud-Bataille M., Valent A., Soularue P., Perot C., Inda M.M., Receveur A., Smaili S., Crollius H.R., Benard J., Bernheim A., et al. 2004. Detecting single DNA copy number variations in complex genomes using one nanogram of starting DNA and BAC-array CGH. *Nucleic Acids Res.* **32**: e112.
- Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., and Lee C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**: 949–951.
- Ishkanian A.S., Malloff C.A., Watson S.K., DeLeeuw R.J., Chi B., Coe B.P., Snijders A., Albertson D.G., Pinkel D., Marra M.A., et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* **36**: 299–303.
- Jain A.N., Tokuyasu T.A., Snijders A.M., Segraves R., Albertson D.G., and Pinkel D. 2002. Fully automatic quantification of microarray image data. *Genome Res.* **12**: 325–332.
- Kallioniemi A., Kallioniemi O.P., Sudar D., Rutovitz D., Gray J.W., Waldman F., and Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**: 818–821.
- Kirsch I.R. and Ried T. 2000. Integration of cytogenetic data with genome maps and available probes: Present status and future promise. *Semin. Hematol.* **37**: 420–428.
- Kirsch I.R., Green E.D., Yonescu R., Strausberg R., Carter N., Bentley D., Levensha M.A., Dunham I., Braden V.V., Hilgenfeld E., et al. 2000. A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nat. Genet.* **24**: 339–340.
- Korenberg J.R., Chen X.N., Sun Z., Shi Z.Y., Ma S., Vataru E., Yimlamai D., Weissenbach J.S., Shizuya H., Simon M.I., et al. 1999. Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. *Genome Res.* **9**: 994–1001.

- Krzywinski M., Bosdet I., Smailus D., Chiu R., Mathewson C., Wye N., Barber S., Brown-John M., Chan S., Chand S., et al. 2004. A set of BAC clones spanning the human genome. *Nucleic Acids Res.* **32**: 3651–3660.
- Lai W.R., Johnson M.D., Kucherlapati R., and Park P.J. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**: 3763–3770.
- Leversha M.A., Dunham I., and Carter N.P. 1999. A molecular cytogenetic clone resource for chromosome 22. *Chromosome Res.* **7**: 571–573.
- Li J., Jiang T., Mao J.H., Balmain A., Peterson L., Harris C., Rao P.H., Havlak P., Gibbs R., and Cai W.W. 2004. Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36**: 952–954.
- Little S.E., Vuononvirta R., Reis-Filho J.S., Natrajan R., Iravani M., Fenwick K., Mackay A., Ashworth A., Pritchard-Jones K., and Jones C. 2006. Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. *Genomics* **87**: 298–306.
- Lucito R., Healy J., Alexander J., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13**: 2291–2305.
- McPherson J.D., Marra M., Hillier L., Waterston R.H., Chinwalla A., Wallis J., Sekhon M., Wylie K., Mardis E.R., Wilson R.K., et al. (International Human Genome Mapping Consortium). 2001. A physical map of the human genome. *Nature* **409**: 934–941.
- Paez J.G., Lin M., Beroukhi R., Lee J.C., Zhao X., Richter D.J., Gabriel S., Herman P., Sasaki H., Altshuler D., et al. 2004. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**: e71.
- Park C.H., Jeong H.J., Choi Y.H., Kim S.C., Jeong H.C., Park K.H., Lee G.Y., Kim T.S., Yang S.W., Ahn S.W., et al. 2006. Systematic analysis of cDNA microarray-based CGH. *Int. J. Mol. Med.* **17**: 261–267.
- Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.L., Chen C., Zhai Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**: 207–211.
- Pollack J.R., Perou C.M., Alizadeh A.A., Eisen M.B., Pergamenschikov A., Williams C.F., Jeffrey S.S., Botstein D., and Brown P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**: 41–46.
- Quackenbush J. 2002. Microarray data normalization and transformation. *Nat. Genet. (suppl.)* **32**: 496–501.
- Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shaperro M.H., Carson A.R., Chen W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.
- Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Mánér S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.
- Sharp A.J., Locke D.P., McGrath S.D., Cheng Z., Bailey J.A., Vallente R.U., Pertz L.M., Clark R.A., Schwartz S., Segraves R., et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**: 78–88.
- Solinas-Toldo S., Lampel S., Stilgenbauer S., Nickolenko J., Benner A., Dohner H., Cremer T., and Lichter P. 1997. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes, Chromosomes & Cancer* **20**: 399–407.
- Telenius H., Carter N.P., Bebb C.E., Nordenskjöld M., Ponder B.A., and Tunnacliffe A. 1992. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**: 718–725.
- Woodfine K., Fiegler H., Beare D.M., Collins J.E., McCann O.T., Young B.D., Debernardi S., Mott R., Dunham I., and Carter N.P. 2004. Replication timing of the human genome. *Hum. Mol. Genet.* **13**: 191–202.

WWW RESOURCES

- <http://bacpac.chori.org/> BACPAC Resources Center (BPRC), Children's Hospital Oakland Research Institute, California.
- <http://www.ncbi.nlm.nih.gov/genome/clone/distributors.html> Clone Registry, National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland.
- <http://www.chip.org/~ppark/Supplements/Bioinformatics05b.html> Lai et al. 2005. Supplementary material. *Bioinformatics*, 2005. Computational Genomics (PI: Peter J. Park).
- <http://humanparalogy.gs.washington.edu/structuralvariation/> Human Structural Variation Database, Department of Genome Sciences, University of Washington, Seattle.
- <http://projects.tcag.ca/variation/>

Representational Oligonucleotide Microarray Analysis Detection of Genetic Variation

Rob Lucito

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724

INTRODUCTION

Historically, copy number variations (CNVs) were first observed by comparing karyotypes of normal and diseased individuals. Studying the karyotypes became clinically important for the diagnosis of certain syndromes, ranging from Down syndrome to cancer (Benson 1961; Stimson 1967). Generally, it is much easier to identify CNVs in cancer, since they can be so numerous, and they can vary widely in the number of copies for a specific region. With

knowledge of the region, methods such as Southern blotting or fluorescence in situ hybridization (FISH) can be performed. However, if a scan of the entire genome is needed or there is no prior knowledge of a region of interest, other techniques are more appropriate. Early generation methods for CNV detection such as comparative genomic hybridization (CGH) (Thompson and Gray 1993) were useful but suffered from low resolution. With the mapping information of the genome project and the development of microarray technologies, CGH was adapted to an array platform (Pinkel et al. 1998; Pollack et al. 1999), and the completion of the human genome sequence paved the way for the development of array CGH methods with exquisite resolution (Lucito et al. 2003; Barrett et al. 2004; Selzer et al. 2005). With this increased resolution, it became apparent that CNVs in the normal population are frequent but had been undetected previously (Lucito et al. 2003; Sebat et al. 2004).

Together with Dr. Michael Wigler, we developed an array-based CGH method to detect CNVs in cancer, although we quickly learned that it was very accurate at identifying small CNVs in the normal population. This method, representational oligonucleotide microarray analysis (ROMA) (Lucito et al. 2003), was an advancement of representational difference analysis (RDA), a technique developed by Dr. Wigler and Dr. Nikolai Lisitsyn (Lisitsyn et al. 1993). Central to both methods is the practice of making a representation of the genome. A representation is reproducible sampling of the genome, prepared by cleavage of the genome with a restriction endonuclease, ligation of PCR adapters, and amplification with a PCR-adequate polymerase such as Taq polymerase. Generally, when the template is a population of mixed fragment sizes,

INTRODUCTION, 233

Representations, 234

FACTORS AFFECTING ROMA, 234

ARRAY PROBE DESIGN, 236

POST-HYBRIDIZATION, 236

SELECTION AND ANALYSIS OF LARGE
SAMPLE SETS, 238

Analysis of Large Data Sets, 238

REFERENCES, 239

an enzyme such as Taq polymerase will preferentially amplify the smaller fragments, with resulting representations in the size range of 100–1200 bp. The lack of amplification of the larger fragments results in a decrease in complexity in the sample. In comparison to total genome hybridization, the decrease in complexity of a representation allows for an increased signal to noise, due to increased hybridization efficiency and increased labeling efficiency for fragments. The increased labeling efficiency is a result of the array being well matched to the content of the representation. In other words, the nucleotide content of the array more closely matches the representation's content than the entire genome.

Representations

The complexity of the representation can be shaped by the restriction endonuclease used for the initial digestion of the genomic DNA. For example, when a restriction endonuclease with a six-nucleotide cleavage site is used, the complexity decreases to 2–5% of the original genome. When a restriction endonuclease with a four-nucleotide cleavage site is used, the complexity decreases to 60–80%, again depending on the cleavage site of the endonuclease. The complexity of the representation can be further shaped by depletion of the representation. In this process, a second restriction endonuclease cleaves the ligation of the genomic DNA prior to PCR amplification. This allows limitless possibilities to achieve the desired complexity.

As a consequence of the required PCR amplification, representations can be easily prepared from as little as 50 ng of genomic DNA. This has made it easier to analyze samples that cannot be renewed, such as tumor specimens, where yield is often in the range of approximately 100 ng. Overamplification can occur: With extreme numbers of cycles, the measured ratios decrease. In most cases, this is not an issue. However, detection of small CNVs (covering regions containing few probes on the array) that are either hemizygous or duplicated can be problematic. We have compared representations prepared with varying number of cycles and have found no change in the level of detection of CNVs, using as many as 25 cycles. Presently, we do not recommend PCR amplification with more than 25 cycles, under the template conditions discussed above.

FACTORS AFFECTING ROMA

Several factors affect the ability to detect CNVs in a particular population, including the array itself, the representation, species, and experimental design. Representations are remarkably accurate at reflecting the copy number of the genome in relation to a reference sample. To accurately detect CNVs within a genome, all sources of system noise that can be controlled must be suppressed. The amplification of samples by PCR adds a possible source of noise that can be easily addressed. Comparable amounts of template genomic DNA should be used, and the test and reference samples should be prepared in parallel from the start of the restriction endonuclease digestion to the PCR amplification. In theory, it is possible to perform ROMA on a single sample without reference, and this is currently under investigation.

The design of the array is equally important when one is attempting to detect CNVs in normal samples. Many CNVs are found in regions that are duplicated within the genome. If these regions are removed, much less variation will be detected. Of course, the addition of regions in the genome that are found more than once must be approached carefully. As with any hybridization-based method, the more times a target is present in the genome, the less specific the hybridization will be. By analogy, repetitive regions in the genome are so numerous that they are useless for copy number detection with hybridization-based methods.

The probes on the array should perform roughly equally. If probe performance is sporadic, a CNV covered by several probes where internal probes are poor reporters could be interpreted as a

cluster of single-nucleotide polymorphisms (SNPs). In this case, CNV detection and, more specifically, automated forms of detection will tend to miss such CNVs. Probes with similar performance should be used for the array, or probe performance parameters should be utilized to interpret the output data. Methods have been developed to predict probe performance, and the present ROMA array utilizes probes of similar performance.

Of equal importance is resolution on the array. If there are not enough probes for a given region, CNVs will not be detected. Many regions are fairly small (~100–200 kb) and arrays with fewer than 40,000 probes will miss many of these regions. The placement of probes could also be an issue. There are untranslated RNAs that are critical to the function of a cell and that can act in an oncogenic manner. Thus, gene-centric CGH arrays are not an optimal design. The initial ROMA arrays were roughly 82,000 probes (although much higher densities can be achieved), randomly scattered throughout the genome so that all regions (with the exception of highly repetitive regions) were equally represented on the array.

In most cases, when ROMA is used to identify genetic variation of normal samples, reference samples will be from a different individual. Since representations use restriction endonucleases to shape complexity, the method uncovers restriction fragment length polymorphisms (RFLPs). If a SNP is present in the test sample and the initial restriction endonuclease cleavage site is destroyed, in comparison to the reference sample an RFLP is created. Usually the new fragment is too large to amplify efficiently and thus is no longer in the representation. Probes recognizing such a fragment would have high-intensity measurements in the reference but very low in the test sample, producing high ratios that cause these probes to scatter from the majority of probe measurements. When these probes are numerous, it can be distracting to the human eye, but the quality of the data is not affected. In fact, since these measurements are the result of SNP detection, they can be used to advantage for informatic identification of CNVs using segmentation algorithms, which is discussed below.

The detection of SNPs is somewhat exaggerated in laboratory species such as the mouse. In contrast to the fairly heterogeneous human genome, the laboratory mouse is extremely homogeneous, except for more recent germ-cell genetic alterations that presumably occur. The mouse ROMA array was based on the first mouse genomic sequence available, which was almost entirely from C57BL6/J. Therefore, if any other species is compared to C57BL6 reference on the array, a large number of SNP events are detected. These events are clustered into patches across the genome, so closely in some regions that at first glance the clusters appear to be large CNVs. Interestingly, the genomic position of SNP clusters varies from one strain to another in comparison to C57BL6/J. Variation between different strains means that experiments must be designed carefully, since the wrong choice of reference could lead to uninterpretable results. The simplest solution is to use material from the same mouse as a reference whenever possible. A sib would be the next best choice, although there are some minor genetic differences between sibs (I. Hall, unpubl.). If a catalog is made of various strains, lineage can be delineated, using the genomic profile of a particular mouse. Comparison of germ-cell to somatic DNA allows one to identify germ-cell DNA alterations in restriction endonuclease cleavage sites over many generations. This in turn enables one to determine the rate of mutation in the animals and to identify any possible hot spots of mutation.

When human ROMA is used to analyze normal copy number variation, the reference is almost always from a different individual, unless somatic and germ-cell DNA are compared. All CNVs that are detected must be interpreted in comparison to the reference. (This is generally not an issue with SNPs.) Therefore, it is preferable to have as few reference DNAs as possible, perhaps one male and one female. Using the same reference samples makes it somewhat easier to identify the copy number for a specific CNV. In addition, whenever possible, it is useful to analyze parents to trace the inheritance of CNVs. When looking for *de novo* alterations, this is essential.

ARRAY PROBE DESIGN

All probes on the array are designed to be complementary to representational fragments, i.e., small restriction fragments. Probe design begins by making an *in silico* representation based on the available genome sequence. This is performed simply by locating and recording all instances of the restriction site of interest and size selecting to 100–1200 bp. Separately, the entire genome is annotated into N-mers of differing length, most commonly, 15- and 21-nucleotide lengths. As a precursor to probe selection, all identified representational fragments are broken down into all possible overlapping 50-mers (the length of the final probes for the array), compared to the library of genome N-mers, and annotated with N-mer frequencies representing the number of times a fragment N-mer is found in the genome library of N-mers (for more detail, see Healy et al. 2003). This is done for N-mers of length $N = 21$ and $N = 15$, using a Burrows-Wheeler transformation (Burrows and Wheeler 1994). For each fragment within the *in silico* digest, we quantitatively characterize every possible 50-mer. The 21-mer frequencies are used as a measure of overall uniqueness in the genome. Our most stringent criterion is that all constituent 21-mers must be unique. However, this is too stringent when analyzing normal variation. For those probes that pass the 21-mer uniqueness measures, 15-mer frequencies are minimized, since they act as a predictive measure of the likelihood of cross-hybridization.

Probes are selected in descending order of aggregate N-mer frequencies, essentially a filtration process. This limits the otherwise immense search space of candidate probes to those that are most likely to give us accurate ratio measurements and least likely to encounter cross-hybridization events or to fail the synthesis process. The probes are then tested empirically. We typically begin by designing and arraying tenfold more probes than can fit on one array. The probes are arrayed on multiple arrays with some overlap for comparison measures. Each array is used in an “enzyme-depletion” experiment where a representation is compared with an identical representation that has been further subjected to digestion with a second enzyme. Since the entire sequence of the genome is known, we have accurate predictions of which fragments should be eliminated by the second digestion step. We select those probes that have the highest signal intensities in the nondigested channel, as we have determined this to be a good measure of performance. The placements of probes on the array surface should be randomized. In this way, no hybridization artifact on the array surface will appear as a consistent genomic alteration.

POST-HYBRIDIZATION

After hybridization, the arrays are scanned on a standard scanner such as the GenePix 4000B, and intensity acquisition is performed with various software packages. The GenePix scanner has a maximal resolution of 5 μm . It is important to average pixel intensities; averaging less than 9 pixels is not advisable. With the maximal resolution, it is difficult to analyze features smaller than 15 $\times$ 15 μm . The arrays currently in use have features that are 17 $\times$ 17 μm . The Axon scanner is still applicable for analysis. For higher-resolution arrays which use smaller features, other analysis software must be used. After data acquisition, the array measurements can be normalized by several different methods. We presently use a Lowess curve fitting algorithm adapted from methods described by Dr. Terry Speed and colleagues (Yang et al. 2002), followed by an algorithm that corrects for local hybridization artifacts. The copy number data have variation in the measurements between adjacent probes, caused by several different factors, including noise in the array hybridization, labeling efficiency, washing, and the representational process. Some of this variation is removed by performing duplicate experiments. For assessing genetic variation in normals where the variation is relatively small (in number of bp) and the alterations often vary by only one copy, duplicate experiments (if not more) are strongly suggested. The hybridization can simply be

repeated, but for two-color hybridizations (such as with cy3 and cy5), it is best to perform a color reversal (or dye swap), where the samples are labeled with the alternative dye as was used for the first hybridization. After normalization, the probe ratios can be averaged, or other methods can be used to remove noise. An example would be probe outliers caused by unincorporated dye. These outliers can be removed by comparing duplicate hybridizations. For a specific feature, if the ratio in one experiment is above one standard deviation, the lower ratio probe is used. This may remove some information content, but we have found that, in general, the quality of the experimental data is improved.

Even with the averaging of experiments, the data are inherently noisy. A number of different algorithms can be employed to remove this noise. The simplest is a moving average (Pollack et al. 1999), which moves in genomic order and averages adjacent probes. The number of probes being averaged can be adjusted. This improves the smoothing process but decreases the measured resolution, since averaging of probes past the breakpoint brings down the average for the probes on the end of the breakpoint. Pollack et al. improved the method by performing normal-normal comparisons which allowed the calculation of a threshold to call false discoveries. This method is quite useful for copy number detection in cancer, but it has limitations for the measurement of normal variation, since small lesions may be missed. Other methods are more involved and model the data (Hodgson et al. 2001; Autio et al. 2003; Snijders et al. 2003). We have used an algorithm developed by Olshen et al. (2004) and a second one developed by Bud Mishra for ROMA data which can be easily used for other CGH data (Daruwala et al. 2004). A third algorithm developed at Cold Spring Harbor Laboratory by Lakshmi Muthuswamy and Michael Wigler expressly for the analysis of ROMA data can also be used for analysis of other CGH platforms (Lakshmi et al. 2006). Each of these three algorithms is model-based. The method developed by Olshen and colleagues is based on binary segmentation but uses a reference distribution of the data to account for noise in the system. Mishra and colleagues use a Bayesian model that can be viewed as an optimization process minimizing a score function. Both of these methods accurately identify copy number alterations in cancer as well as normal genetic variation, but they require significant compute time to analyze an experiment as compared to the method developed by Muthuswamy and Wigler.

The method developed by Muthuswamy and Wigler is an alternative which partitions the log ratios of the intensities into segments of uniform distributions. The ratios are arranged in genome order and are divided into blocks of 100 data points with arbitrary boundaries. Then, the boundaries are iteratively moved by minimizing the variance, and further refined by using a two-distribution Komogrov-Smirnov (KS) null hypothesis test (Conover 1980). Only boundaries that give a p -value of less than 10^{-5} are accepted. Due to the nature of the KS test, no segment smaller than 3 probes in length is considered. This method is excellent for the analysis of lesions comprising 3 or more probes. A more accurate method of segmentation is being developed which utilizes a hidden Markov model. It uses the presence of SNPs to define the states of alteration between the two genomes; i.e., zero copies, one copy, three copies, or more copies. Once this is determined, the state of other alterations can be identified based on initial calculations.

After processing, the data can be visualized to identify CNVs and any SNPs present between the normal and reference samples. Since the alterations identified are in comparison to the reference, it is possible that a CNV could be present as three copies in both the test and reference sample, and the difference would cancel out. This is unlikely for rare CNVs, but much more possible for common CNVs. If one sample has a CNV that is slightly larger than that found in the other sample, the result would appear as two CNVs, since the central region would be cancelled out. These possibilities must be taken into account when analyzing data.

Large variations can be validated by other methods. Clearly, the most accurate method to validate CNV variation is fluorescent in situ hybridization (FISH). This method can assess the true copy number of an allele present in a sample, whereas the CGH data can only determine the level in comparison to the reference sample. However, FISH is somewhat slow in throughput and requires intact cells. An alternative is quantitative PCR (QPCR). This method does not require

whole cells, can be performed with mid-throughput, and requires little DNA template. We have used this technique routinely to determine the relative copy number of CNVs identified in large ROMA data sets. The copy number is always calculated in comparison to a reference sample. If the reference is known to possess two copies of the region being studied, the true copy number of the test sample can accurately be determined. The most accurate determination requires a second control probe whose copy number is known for the reference as well as the test sample.

SELECTION AND ANALYSIS OF LARGE SAMPLE SETS

Although CNVs identified in a single sample can be validated, copy number detection is often used to analyze a large sample set. This may be performed on a large set of normal individuals to determine the variation in the normal population of humans (or mice). Alternatively, one can analyze a large set of individuals that has, or is suspected to have, a syndrome or disease, and the identified CNVs can be coalesced to determine whether any are associated with the particular syndrome or disease. Some history should be known about the individuals whose samples are to be analyzed. For example, to study familial susceptibility to pancreatic cancer, it would be necessary to know that the individual and at least one more member in the immediate family had pancreatic cancer. Because pancreatic cancer affects a relatively small number of individuals worldwide, two individuals within the same family would increase the probability that the cancer had a familial component (provided that environmental factors were removed). The disease or syndrome studies should be as homogeneous as possible. If a syndrome encompasses more than one true disease, it will be more difficult to determine the candidate regions for any one subsyndrome. Even in the case of familial susceptibility to pancreatic cancer, more than one gene may be involved. Careful selection of the test sample set is imperative for creating a valuable set of CNVs. Since primary cells are not available as a source of DNA, sample cells are often transformed by EBV infection so that large amounts can be grown. Such material should be approached with caution, since the transformation process and the extended period of time in culture can themselves result in genomic anomalies.

Analysis of Large Data Sets

To determine which identified CNVs are potentially associated with the disease being studied, as many normal CNVs as possible must be removed or filtered from the data set. One needs to amass a set of normals to make a catalog of CNVs and use the catalog to remove normal CNVs from the test data set. The definition of a normal CNV is somewhat arbitrary and depends on the test set; i.e., the normal set will change depending on the syndrome that is being studied. For example, if the test set is for patients who are susceptible to one or more cancers, all normals in the catalog who are from cancer-susceptible families should be removed. Thus, it is imperative that accurate family data are obtained for the individuals in the normal catalog. Selecting the catalog for comparison to a sample set with a rare cancer such as pancreatic cancer is relatively easy. Strict rules can be used to remove any "normals" who have had pancreatic cancer or have had any immediate family members who have had pancreatic cancer. However, selection is more difficult with more common cancers that have a much higher rate of sporadic occurrence or with syndromes that have low penetrance. In addition, one of the catalog members could be an affected individual who is asymptomatic. If such an individual enters the normal data set, meaningful regions could be lost during the filtering process. We address one possible way of rescuing such data below.

The size of the normal set for CNV filtering should be the largest possible, to allow the removal of common CNVs within the test set of samples. Of course, there will be rare CNVs that cannot be filtered from the population, no matter how large the filter set becomes. If the normal sample set is too small, the test data set will contain a large number of relatively common normal CNVs.

Once the selection of the normal sample set is completed, ROMA can be performed. In addition, normal data sets from other labs that are publicly available can also be used to filter common CNVs. This should only be done if accurate patient information is available.

Current ROMA arrays have 85,000 probes, but we have begun using 390,000-probe arrays. In many cases, the normal sample filter data set, as well as the test data set, can become very large. The data set, if based on all probes, would require a large data frame to represent all samples and would slow down computing time. Much of the data from the genome reports nonevents or vast regions without CNVs. By only representing the identified CNVs, the data can be drastically reduced in size without decreasing the information content. Once this transformation is performed for all samples, a master file can be produced for the normal set that can be used to filter the test samples.

For filtration, all CNVs of the normal set that are found in the test set are subtracted. Often, fragments of CNVs are left as remnants when a CNV found in the test set is larger in length (covered by more probes) than the CNV found in the normal filter set. In the worst case, the CNV in the test set would be larger on both ends, compared to the CNV found in the normal set. After the filtration, it would appear that there are two CNVs in close proximity. The entire CNV can be removed, but this must be done with caution, because valuable data could be lost. At the end of this process, the frequency that a CNV is found in the test set population can be calculated. Removal of CNVs found in the normal population is a strict constraint, especially if there is the unintentional inclusion of an asymptomatic individual. An alternative to simply removing regions that are found in the normal population is to allow some flexibility in the level of detection. If a CNV is found in the test set and in only one individual of the normal set, it will not be removed. This level of detection or flexibility can be increased up to whatever point is deemed informative for the data set. The drawback of such flexibility is that it will increase the number of rare CNVs that happen to be in both data sets as well as some common CNVs. Therefore, it is useful to analyze the filtered data by several different criteria.

Once the regions of CNVs in the test set are identified, other analyses can commence. As an example, the coordinates of the regions could be compared to results from other mapping studies using alternative technologies, such as SNP mapping and quantitative trait locus (QTL) analysis. Informatic analyses can be used to identify all genes within the regions to determine whether there are any interesting candidate genes based on gene function and the syndrome or disease being studied. It is unlikely that all CNVs will be validated, since this process is time consuming and labor intensive. However, it is important to validate a number of CNVs (possibly those of specific interest) with another method. FISH is the most accurate method for validation, but QPCR is a legitimate alternative. It is somewhat difficult and requires a large number of replicas, since the difference expected can be only one cycle. Nonetheless, it is a viable option if whole cells are not available.

REFERENCES

- Autio R., Hautaniemi S., Kauraniemi P., Yli-Harja O., Astola J., Wolf M., and Kallioniemi A. 2003. CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics* **19**: 1714–1715.
- Barrett M.T., Scheffer A., Ben-Dor A., Sampas N., Lipson D., Kincaid R., Tsang P., Curry B., Baird K., Meltzer P.S., et al. 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci.* **101**: 17765–17770.
- Benson E.S. 1961. Leukemia and the Philadelphia chromosome. *Postgrad. Med.* **30**: A22–A28.
- Burrows M. and Wheeler D.J. 1994. A block sorting lossless data compression algorithm. HP technical report SRC-RR-124, Hewlett-Packard, Palo Alto, California.
- Conover W.J. 1980. *Practical nonparametric statistics*, 2nd edition. Wiley, New York.
- Daruwala R.S., Rudra A., Ostrer H., Lucito R., Wigler M., and Mishra B. 2004. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci.* **101**: 16292–16297.
- Healy J., Thomas E.E., Schwartz J.T., and Wigler M. 2003. Annotating large genomes with exact word matches. *Genome Res.* **13**: 2306–2315.
- Hodgson G., Hager J.H., Volik S., Hariono S., Wernick M., Moore D., Nowak N., Albertson D.G., Pinkel D., Collins C., et al. 2001. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.* **29**: 459–464.

- Lakshmi B., Hall I.M., Egan C., Alexander J., Leotta A., Healy J., Zender L., Spector M.S., Xue W., Lowe S.W., et al. 2006. Mouse genomic representational oligonucleotide microarray analysis: Detection of copy number variations in normal and tumor specimens. *Proc. Natl. Acad. Sci.* **103**: 11234–11239.
- Lisitsyn N., Lisitsyn N., and Wigler M. 1993. Cloning the differences between two complex genomes. *Science* **259**: 946–951.
- Lucito R., Healy J., Alexander J., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13**: 2291–2305.
- Olshen A.B., Venkatraman E.S., Lucito R., and Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.
- Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.L., Chen C., Zhai Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**: 207–211.
- Pollack J.R., Perou C.M., Alizadeh A.A., Eisen M.B., Pergamenschikov A., Williams C.F., Jeffrey S.S., Botstein D., and Brown P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**: 41–46.
- Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.
- Selzer R.R., Richmond T.A., Pofahl N.J., Green R.D., Eis P.S., Nair P., Brothman A.R., and Stallings R.L. 2005. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* **44**: 305–319.
- Snijders A.M., Pinkel D., and Albertson D.G. 2003. Current status and future prospects of array-based comparative genomic hybridisation. *Brief Funct. Genomics Proteomics* **2**: 37–45.
- Stimson C.W. 1967. Possible causes of mongolism (Down's syndrome) and other chromosomal aneuploidies. *Mich. Med.* **66**: 436–441.
- Thompson C.T. and Gray J.W. 1993. Cytogenetic profiling using fluorescence in situ hybridization (FISH) and comparative genomic hybridization (CGH). *J. Cell. Biochem. Suppl.* **17G**: 139–143.
- Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., and Speed T.P. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**: e15.

16

Whole-Genome Sampling Analysis to Detect Copy Number Changes in FFPE Samples

Sharoni Jacobs

Affymetrix, Santa Clara, California 95051

INTRODUCTION

WGSa and Copy Number Detection

Human populations and individuals can differ widely in both nucleotide sequence and copy number (CN) of DNA segments. Sequence differences commonly occur in the form of single-nucleotide polymorphisms (SNPs), of which millions have been identified (International HapMap Consortium 2005). Recent studies have also begun to highlight CN polymorphism frequencies (Iafate et al. 2004; Sebat et al. 2004; Redon et al. 2006), suggesting that a comprehensive study of genetic differences between samples should combine detection of both sequence and CN variation.

The Affymetrix GeneChip® Mapping Arrays comprise a class of SNP oligonucleotide microarrays that can be used to determine both genotype and CN at specific SNP locations. The original Mapping array, known as the Mapping 10K Array and, later, the Mapping 10K 2.0 Array, interrogates over 10K SNPs on a single array. The next generation of Mapping arrays, the Mapping 100K Set, includes two arrays (the Mapping 50K Xba and the Mapping 50K Hind Arrays) that together analyze 116,204 SNPs. The third generation of these arrays, the Mapping 500K Set, which includes the Mapping 250K Nsp and the Mapping 250K Sty Arrays, examines 500,568 SNPs with a median and mean inter-SNP distance of 2.5 kb and 5.8 kb, respectively.

Genomic DNA applied to these arrays is processed using the Whole Genome Sampling Analysis (WGSa) assay (Fig. 16-1) (Kennedy et al. 2003). Importantly, this assay involves a complexity reduction step, so that instead of applying the entire genome to these arrays, a targeted subset of the genome is selectively labeled and hybridized. Reduction of DNA target assembly is accomplished as follows: During WGSa, 250 ng of genomic DNA is digested by a restriction enzyme (NspI or StyI when using the Mapping 500K Arrays), and adapters are ligated to the ends of the digested DNA. A total of 500 ng of DNA is required for application when both arrays in the

INTRODUCTION, 241

WGSa and Copy Number Detection, 241

FFPE Samples for CN Detection, 242

DNA PREPARATION, 243

FFPE DNA QUALITY ASSESSMENT, 244

APPLYING FFPE DNA TO THE WGSa ASSAY, 245

DNA Quantitation, 245

Digestion and Ligation, 245

PCR, 245

Purifying and Pooling Amplification Reactions, 246

Fragmentation and Labeling, 246

Data Analysis, 247

PERFORMANCE METRICS, 247

SUMMARY, 247

REFERENCES, 248

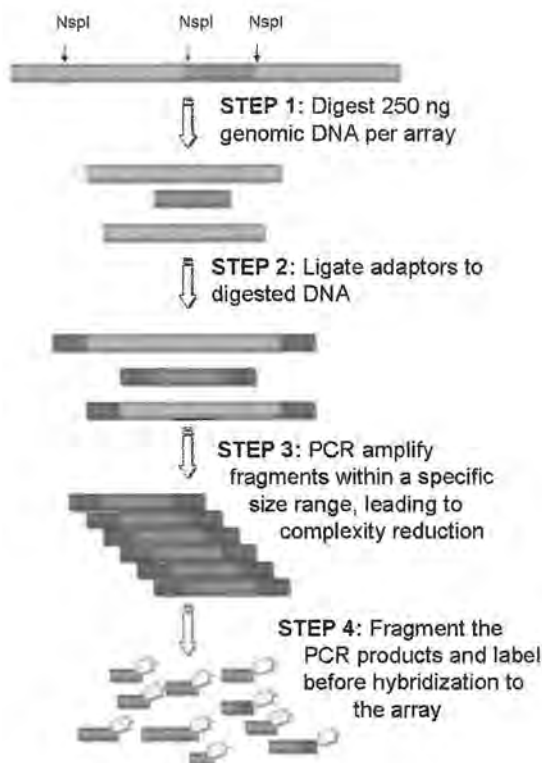


FIGURE 16-1. The Whole Genome Sampling Analysis (WGSA) assay prepares DNA for the Mapping 500K Arrays.

Mapping 500K Set are used. The digested DNA is amplified by PCR using a universal primer set that recognizes the adapters. The PCR conditions and reagents are optimized to maximize amplification of products within a specific size range. This size range is approximately 100 bp to 1100 bp for the Mapping 10K and 500K Arrays. For the Mapping 100K Arrays, the amplicons range from about 250 bp to 2 kb.

PCR products are then fragmented with DNase I, labeled with biotin, and hybridized to the array. Hybridization efficiencies to probes on the array that recognize alternate alleles of each SNP are detected by staining the biotin-labeled DNA fragments with fluorescent-labeled streptavidin.

Differential intensities for the probes that recognize the major and minor allele are used to determine SNP genotype (Matsuzaki et al. 2004; Di et al. 2005). These probe intensities can also be used to determine genomic CN at the location of each SNP (Bignell et al. 2004; Huang 2004; Lin et al. 2004; Nannya et al. 2005). To determine CN, a test sample is compared to a reference consisting of one or many samples. The reference is assumed to represent a diploid state with $CN = 2$, and the ratio of probe intensities between the test and the reference is calculated in order to identify the number of copies of a DNA fragment in the test sample.

FFPE Samples for CN Detection

The WGSA assay in combination with the Mapping arrays is used widely for CN analysis of high-quality DNA samples; i.e., samples that have been collected from blood, fresh or frozen tissue, or cell lines. Formalin-fixed, paraffin-embedded (FFPE) samples, however, represent the most prevalent form of archived clinical samples (Fig. 16-2), but they provide additional challenges for molecular assays. FFPE processing is advantageous primarily to the pathologist, who desires a rapid means to fix, stain, and view the samples for a speedy and efficient diagnosis. Unfortunately,



FIGURE 16-2. Example of a paraffin block containing a formalin-fixed kidney sample, and 10- μ m sections from the FFPE sample contained in the eppendorf tubes.

the method usually results in the degradation of FFPE DNA and in the contamination and chemical modification of these DNA samples.

Because of these issues, FFPE DNA is not suitable for all molecular assays designed for high-quality DNA samples. A protocol recommended for processing FFPE DNA samples through WGS and to the Mapping arrays is described below. Note that poorer quality of FFPE samples may still result in a decrease in the number of SNPs providing informative data, and therefore overall genome coverage. Guidelines on data analysis of FFPE samples are outlined in Chapter 22. The important modifications to this WGS protocol specific to FFPE-derived samples include:

1. inclusion of a pre-WGS PCR-based quality assessment of FFPE DNA samples
2. pooling of >3 amplification reactions to attain 90 μ g of PCR product
3. quality control (QC) steps to monitor performance (expect different performance from FFPE and non-FFPE DNA samples)

DNA PREPARATION

Several high-quality commercial kits are currently on the market for extracting DNA from FFPE samples. The quality of DNA that is extracted from a given sample will vary depending on the extraction protocol, so that the same FFPE block might yield relatively poor or relatively intact DNA using alternate kits. A recommendation for extraction is the use of the Qiagen DNeasy Blood & Tissue kit (www.qiagen.com) with modifications (Wu et al. 2002). Suggested modifications to the protocol that is outlined in the Qiagen DNeasy Blood & Tissue Handbook include:

1. 95°C, 15-minute treatment prior to Proteinase K digestion, in ATL buffer
2. 3 days of Proteinase K treatment at 56°C, with daily additions of the same amount of PK used on the first day
3. NH_4OAc + EtOH cleanup of extracted DNA, as described in the "Genomic DNA Preparation" section of the GeneChip® Mapping 500K Assay Manual (www.affymetrix.com)

Another method that has demonstrated success in providing FFPE DNA of sufficient quality for the Mapping arrays is Argylla Technologies PrepParticles (beta version) (<http://argylla.com/>).

There are many extraction methods available that may also be used successfully. Regardless of the method, the resulting DNA sample should be assessed for quality, since not all samples will be suitable for the Mapping arrays. To pass the quality test, the DNA sample must be void of extensive contamination and degradation. Refer to the next section for more details on QC testing of FFPE DNA samples.

FFPE DNA QUALITY ASSESSMENT

FFPE DNA sample quality will differ due to a number of variables; for example, differences in the fixation protocol and the extraction protocol, and the years of storage. Some FFPE samples will fail to produce nucleic acids of sufficient quality and quantity for downstream analysis. Therefore, each FFPE DNA sample should be quality tested, and samples that fail this QC test should not be applied to the Mapping arrays. The proportion of samples that fail this QC test will vary based on the source of FFPE samples, extraction methodology, and other factors.

In WGS, PCR is the limiting step that determines whether a given FFPE DNA sample will provide informative data on the Mapping array. Therefore, the FFPE DNA quality assessment test should include PCR. The three PCR-based QC tests in the table below are able to determine the suitability of an FFPE sample for these arrays. Any one of these three tests may be used for FFPE DNA quality assessment.

The purpose of these QC tests is to identify the largest producible amplicon size for a given FFPE sample. Importantly, these QC examples all include the amplification of a variety of fragment sizes during PCR, including larger fragment sizes up to 800 bp. Because the Mapping 500K Arrays interrogate SNPs on fragments up to 1100 bp, an FFPE sample that qualifies for array analysis should be able to provide large amplicon sizes. Although it is not necessary to amplify all fragments on the array (in silico, the SNPs on the largest fragment sizes can be excluded during analysis), it is important to amplify a significant percentage of these SNPs. Table 16-1 gives an indication of the number of SNPs present on the various fragment sizes. We do not recommend proceeding with samples that fail to produce PCR fragments greater than 300 bp.

Method	Description	Advantages and disadvantages	References
Multiplex PCR	• Simultaneous amplification of several fragments ranging in size from ~100 to 800 bp. • Each lab may design primers and reactions to amplify any region of the genome. • It is essential that the reactions are designed to amplify not only small but also large (600, 700 bp) fragments.	(+) Requires a small amount of DNA. (-) Targeted PCR amplification of specific sites may identify locus-specific quality instead of genome-wide DNA quality.	van Beers et al. (2006) This article describes multiplex PCR of fragments up to 400 bp only.
RAPD PCR	• Uses a single primer set (10-mers) to amplify a ladder of fragments. • Intact DNA will produce amplicons reaching ~2 kb. • The amplicon ladder produced from FFPE samples may reach a smaller maximum size. • This is not a kit.	(+) Requires a small amount of DNA. (+) Provides non-targeted amplification across the genome. (-) Bacterial DNA in the Taq Polymerase may be amplified to provide high background when using an enzyme preparation that is not highly purified. (-) Some variability can be seen between replicates and dilutions.	Siwoski et al. (2002), with the suggested modifications: • Visualize bands with EtBr on an agarose gel • Use a highly purified Taq polymerase, such as Qiagen HotStar HiFidelity DNA Polymerase
First steps of WGS	• 250 ng of DNA are digested, ligated to adapters, and PCR-amplified using a universal primer set. • See Figure 16-3A for an example.	(+) Provides best correlation to Mapping array performance because it is part of the Mapping assay. (+) Products that pass QC inspection may directly continue through the WGS assay and hybridize to the array. (-) Requires the most DNA, time, and money per sample.	GeneChip® Mapping 500K Assay Manual

TABLE 16-1. The Number of SNPs Retained on the Mapping 500K Arrays When Applying Various Fragment Size Filters

Fragment size filter	Mapping 250K Sty Array	Mapping 250K Nsp Array	Combined 500K Set	% of SNPs on 500K Set
≤ 200 bp	1,260	1,813	3,073	0.6
≤ 300 bp	15,831	13,650	29,481	5.9
≤ 400 bp	45,459	39,506	84,965	17.0
≤ 500 bp	82,085	74,387	156,472	31.3
≤ 600 bp	120,011	113,702	233,713	46.7
≤ 700 bp	155,576	153,213	308,789	61.7
≤ 800 bp	187,689	190,900	378,589	75.6
≤ 900 bp	213,302	222,317	435,619	87.0
≤ 1 kb	230,509	244,665	475,174	94.9
No filter	238,300	262,258	500,568	100.0

APPLYING FFPE DNA TO THE WGS ASSAY

The Mapping 500K Set, or the individual Mapping 500K Arrays, should be used in preference to the other Mapping arrays when processing FFPE DNA. The Mapping 100K Arrays are not appropriate for degraded samples due to their requirement for larger amplicons during PCR. The Mapping 10K Array can be used for FFPE DNA analysis, as it does share the same distribution of PCR amplicon sizes as the Mapping 500K Arrays, but this array will provide significantly decreased resolution.

The WGS assay is described in detail in the GeneChip® Mapping 500K Assay Manual provided by Affymetrix (www.affymetrix.com). Here, we highlight the specific steps that require special attention or that have been modified for application of FFPE DNA samples.

DNA Quantitation

For accurate quantitation of double-stranded FFPE DNA, use a method that specifically detects dsDNA, such as PicoGreen® (www.invitrogen.com). Other methods of quantitation, such as UV spectroscopy, may provide inaccurate measurements of DNA due to the higher amounts of contaminants in FFPE samples.

Digestion and Ligation

No changes are required during the digestion and ligation steps of WGS. Begin with 250 ng of DNA, and digest with NspI for samples destined for the Mapping 250K Nsp Array, and StyI for samples directed toward the Mapping 250K Sty Array. The ligation reaction will result in 100 µl of product.

PCR

The established protocol requires three amplification reactions per DNA sample, using 10 µl of ligation product per PCR reaction. Therefore, a total of 30 µl is taken from the 100-µl ligation product. After PCR, the three reactions are pooled together, and 90 µg of PCR product is carried over to the next step (Fragmentation). Because degraded FFPE DNA may not produce larger amplicons during PCR (Fig. 16-3A), the yield of PCR products is often lower for FFPE samples, and it is necessary to pool additional PCRs to reach the suggested 90 µg. In anticipation of the need to pool additional PCRs, set up six amplification reactions per sample (using 60 µl from the 100-µl ligation product). For severely degraded samples, nine reactions may be required. Importantly, this step does not affect the input amount of DNA, which remains at 250 ng per array.

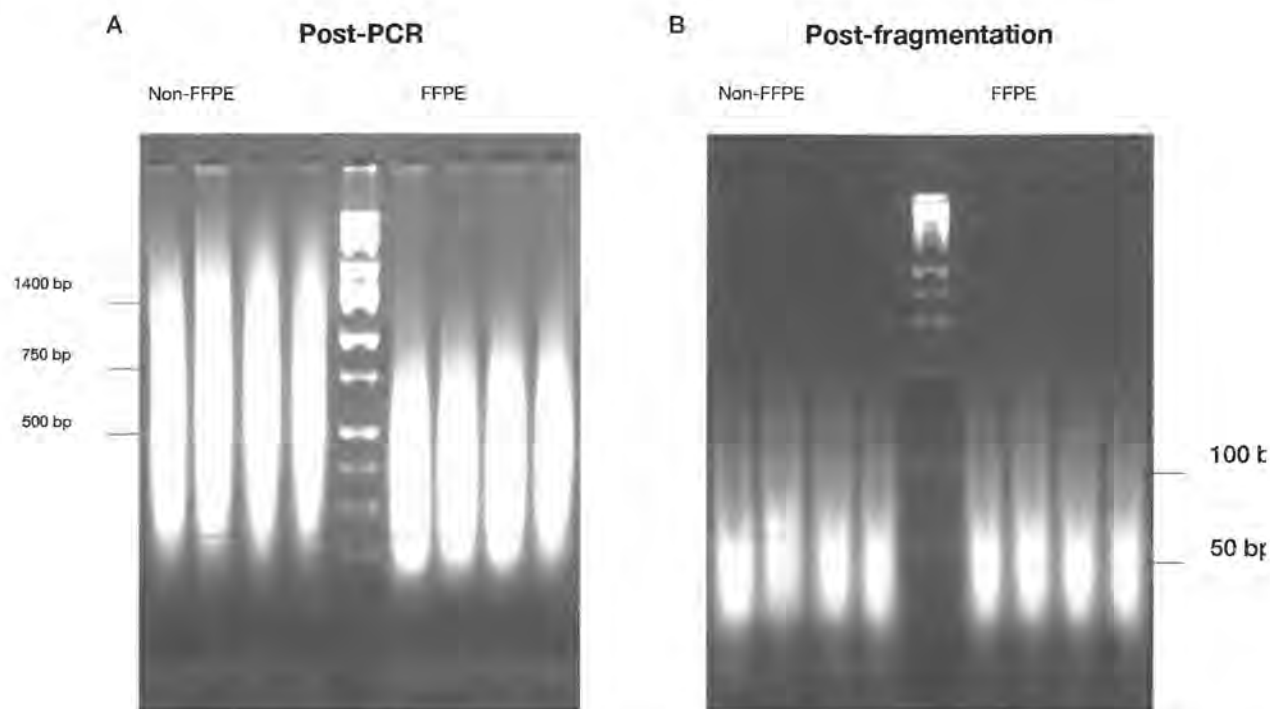


FIGURE 16-3. QC inspections of sample performance during WGA. (A) PCR products are visualized on an agarose gel for both non-FFPE and FFPE samples. The larger fragment sizes failed to be amplified from the FFPE samples. Nonetheless, there is ample PCR product for the smaller fragments, indicating that these samples are suitable for proceeding with WGA and hybridizing to the arrays. (B) Fragmentation products are visualized on a gel for both non-FFPE and FFPE samples. Products look essentially equivalent and are suitable for labeling and hybridization to the arrays.

Combining only three PCRs from an FFPE sample is acceptable even when this provides less than 90 μ g of product, but pooling additional samples to increase the amount of DNA applied to the array generally increases the quality of the resulting data.

Purifying and Pooling Amplification Reactions

Clontech® 96-well plates are used to pool and purify PCRs for the Mapping 500K assay. Up to three reactions can be pooled together in a single well. When more than three PCRs are performed for a single DNA sample, multiple wells should be used for purification. Thus, if there are six reactions from one DNA source, there will be two wells used for PCR purification for this sample. The DNA from these wells can be combined during the DNA elution step as follows: After washing both wells three times with water and allowing the DNA to completely dry on the plate (as per instructions in the GeneChip® Mapping 500K Assay Manual), elute the first well with 45 μ l of RB buffer while leaving the second well dry. Next, use the same RB buffer that now contains the DNA eluted in the first well to elute the DNA in the second well. In this way, all of the PCR products from the six reactions will be eluted into the same 45 μ l of RB. If nine amplification reactions were performed for a single DNA source, use three wells on the Clontech® plate for this sample, and so on.

Fragmentation and Labeling

Fragment 90 μ g of PCR product when possible. Expect decreased performance on the array if a smaller amount of DNA is used as input to fragmentation.

Data Analysis

Refer to Chapter 22 for details on how to analyze Mapping array data resulting from application of FFPE samples.

PERFORMANCE METRICS

Performance of the DNA sample can be monitored during and after the WGS assay (see table below). It is useful to include a positive control in a batch of FFPE samples, such as the standard high-quality DNA sample provided with the Affymetrix kit, Ref103.

SUMMARY

FFPE samples represent the largest resource for archived clinical samples, but they provide special challenges for molecular assays due to DNA degradation, chemical modification, and sample contamination. Special consideration should be given when choosing an extraction method to obtain the best-quality DNA possible from a given FFPE block. All FFPE samples should be assessed for quality using a PCR-based test. Those that provide amplicons of adequate size can be applied to genome-wide analysis of genotype and CN using the Mapping 500K Arrays. Genome coverage for an FFPE sample compared to a high-quality DNA sample may be reduced in cases where PCR fails to amplify SNPs on the larger fragment sizes.

QC Inspection	Details
Image the PCR products on an agarose gel (Fig. 16-3A).	Non-FFPE samples: The PCR products should provide a smear from approximately 100 bp to >1 kb. FFPE samples: The larger PCR products will likely be missing, but there should be robust amplification of smaller PCR products. Do not proceed with samples that do not provide strong bands greater than 300 bp.
Measure the PCR yield using UV spectroscopy.	Non-FFPE samples: These should provide >90 µg yield from a pool of three PCR reactions. FFPE samples: These may have significantly reduced PCR yields. Ideally, samples should provide 90 µg when six or nine PCRs are pooled. If 90 µg is not attained with nine PCRs, expect minimal performance from this sample. Generally, higher yields of PCR products indicate larger numbers of SNPs that can be included during data analysis.
Image the fragmentation products on a gel (Fig. 16-3B).	Fragmentation products should be approximately 50 to 100 bp, for both FFPE and non-FFPE samples.
Check call rate.	Non-FFPE samples: The best data are found when the call rate (the percentage of SNPs assigned genotypes) is $\geq 93\%$* when using the DM algorithm and $\geq 96\%$** when using BRLMM (both algorithms are applied in the genotyping software GTYPE 4.1). FFPE samples: These often have reduced call rates. Nonetheless, call rates are usually above 70% for successful application. Values lower than this suggest poor data quality.

*With a p -value cutoff of 0.33.

**With a p -value cutoff of 0.33 and a BRLMM score threshold of 0.5.

There are a few modifications recommended for application of FFPE samples to the Mapping array. During WGS, increasing the number of amplification reactions to obtain 90 µg of PCR product before fragmentation generally improves data quality. Modifications during data analysis are also necessary, and these are discussed in Chapter 22 of this manual. Whole-genome amplification of precious DNA samples can be paired with the WGS assay when using high-quality DNA samples (Wong et al. 2004; Zhou et al. 2005); but, at this time, there are no recommendations available for performing whole-genome amplification on degraded samples before application to the Mapping assay.

REFERENCES

- Bignell G.R., Huang J., Greshock J., Watt S., Butler A., West S., Grigorova M., Jones K.W., Wei W., Stratton M.R., et al. 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* 14: 287–295.
- Di X., Matsuzaki H., Webster T.A., Hubbell E., Liu G., Dong S., Bartell D., Huang J., Chiles R., Yang G., et al. 2005. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* 21: 1958–1963.
- Huang J. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* 1: 287–299.
- Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., and Lee C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* 36: 949–951.
- International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437: 1299–1320.
- Kennedy G.C., Matsuzaki H., Dong S., Liu W.M., Huang J., Liu G., Su X., Cao M., Chen W., Zhang J., et al. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.* 21: 1233–1237.
- Lin M., Wei L.J., Sellers W.R., Lieberfarb M., Wong W.H., and Li C. 2004. dChipSNP: Significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* 20: 1233–1240.
- Matsuzaki H., Dong S., Loi H., Di X., Liu G., Hubbell E., Law J., Berntsen T., Chadha M., Hui H., et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* 1: 109–111.
- Nannya Y., Sanada M., Nakazaki K., Hosoya N., Wang L., Hangaishi A., Kurokawa M., Chiba S., Bailey D.K., Kennedy G.C., and Ogawa S. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* 65: 6071–6079.
- Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W., et al. 2006. Global variation in copy number in the human genome. *Nature* 444: 444–454.
- Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.
- Siwoski A., Ishkanian A., Garnis C., Zhang L., Rosin M., and Lam W.L. 2002. An efficient method for the assessment of DNA quality of archival microdissected specimens. *Mod. Pathol.* 15: 889–892.
- van Beers E.H., Joosse S.A., Ligtenberg M.J., Fles R., Hogervorst E.B., Verhoef S., and Nederlof P.M. 2006. A multiplex PCR predictor for aCGH success of FFPE samples. *Br. J. Cancer* 94: 333–337.
- Wong K.K., Tsang Y.T., Shen J., Cheng R.S., Chang Y.M., Man T.K., and Lau C.C. 2004. Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic Acids Res.* 32: e69.
- Wu L., Patten N., Yamashiro C.T., and Chui B. 2002. Extraction and amplification of DNA from formalin-fixed, paraffin-embedded tissues. *Appl. Immunohistochem. Mol. Morphol.* 10: 269–274.
- Zhou X., Temam S., Chen Z., Ye H., Mao L., and Wong D.T. 2005. Allelic imbalance analysis of oral tongue squamous cell carcinoma by high-density single nucleotide polymorphism arrays using whole-genome amplified DNA. *Hum. Genet.* 118: 504–507.

Molecular Inversion Probe Targeted Genotyping: Application to Copy Number Determination

George Karlin-Neumann, Marina Sedova, Ronald Sapolsky, Steven Lin, Yuker Wang, Martin Moorhead, and Malek Faham

Affymetrix, South San Francisco, California 94080

INTRODUCTION

The ability to discover and accurately measure chromosome copy number (CN) amplifications, deletions, and loss of heterozygosity (LOH) in cancer samples has great clinical potential for diagnosis, prognosis, and treatment decisions (Slamon et al. 2001). A primary challenge in copy number determination is the ability to

make quantitative as well as qualitative genotyping measurements. Cancer samples present the additional challenges of sample purity (mixtures of cancer and normal cells in the assayed tissue); low DNA sample amount; and in the case of archived formalin-fixed, paraffin-embedded (FFPE) samples, physically and chemically degraded DNA which does not amplify well and is difficult to assay with existing methods (van Beers et al. 2006). Several array-based techniques have been developed to assess DNA copy number. The ideal copy number assay would provide in a highly multiplexed fashion genotype and precise copy number estimation not only on abundant, high-quality DNA samples, but also on low amounts of degraded DNA. Chapter 12 describes the use and many advantages of molecular inversion probes (MIPs) in targeted SNP genotyping. The advantage of utilizing MIPs in the study of copy number stems from its exquisite specificity due to the requirement of the interaction between two homology sequences on the same molecule as described in Chapter 12. In addition, the small target sequence requirement (~40 bp of homology with the probe's terminal detection sequences) makes it suitable for the study of degraded samples. These features make the MIP-genotyping assay well suited to the challenge of copy number determination even in low amounts of unamplified, degraded sample DNA. The quantitative capabilities of the MIP genotyping assay are illustrated below in a modified version of the assay, not yet commercially available. This version requires <75 ng of unamplified genomic DNA and was tested with a 50,000-plex MIP genome-wide human panel on DNA derived both from normal and cancer cell lines and from FFPE samples, and it performed very well on each. To determine the copy number for each allele of a given marker in an unknown sample, each allele's signal intensity is first calibrated against copy number in a set of reference samples.

INTRODUCTION, 249

RESULTS OF CN ESTIMATES FROM MIP
GENOTYPING ASSAY, 250

SUMMARY AND CONCLUSIONS, 251

ACKNOWLEDGMENTS, 252

REFERENCES, 252

RESULTS OF CN ESTIMATES FROM MIP GENOTYPING ASSAY

To assess the accuracy and precision of MIP CN estimates, the assay was run on cell lines having various numbers of copies of the X chromosome, from 1X males to 3X, 4X, and 5X lines (cell lines were obtained from the Coriell Institute, Camden, New Jersey). The 50K-plex panel contains >900 scorable markers on the X chromosome. Plots of marker CN versus chromosome are shown in Figure 17-1 for 1X (male), 3X, 4X, and 5X cell lines. The markers on each chromosome are distinctly colored from their neighbors, and autosomes are ordered from left to right starting with Chromosome 1 on the left (red) through Chromosome 22 on the right (pink), with the X-chromosome markers on the extreme right (peach). Each marker CN value is unsmoothed but, nonetheless, shows high precision; smoothing of only two markers increases this precision still further. The CN estimates for X-chromosome markers in 1X male lines was 1.06 with a relative standard deviation (rsd) of 0.12; for the 3X cell line, CN = 3.09 with rsd = 0.10; for the 4X cell line, CN = 3.98 with rsd = 0.10; and for the 5X cell line, CN = 4.96 with rsd = 0.10. Copy number estimates are both accurate and precise within this range. It is noteworthy that the assay has been able to identify some putative regions of CN aberration in these cell lines, e.g., the apparent amplification seen in Chromosome 14 (purple) in the 3X line. Additional data with spikes of known concentrations show that the assay is linear to >50 copies (data not shown).

The ability of the assay to identify deleted regions and amplified regions within this range is further illustrated in Figure 17-2. The figure depicts the chromosome CN plot of the cancer cell line UACC812 (cell line obtained from the American Type Culture Collection [ATCC]). The chromosomes are ordered from left to right, with Chromosome 1 on the far left and X chromosome on the far right. Striking increase in copy number is evident for several markers. The levels of several of these amplifications have been confirmed also by TaqMan measurements (Wang et al. 2005).

In a third example of this application, the ability of this modified MIP assay to perform well on DNA derived from both normal and tumor FFPE samples has been investigated. In this case, DNA was purified from several 10- μ m slices of all sample blocks using a standard QIAmp DNA

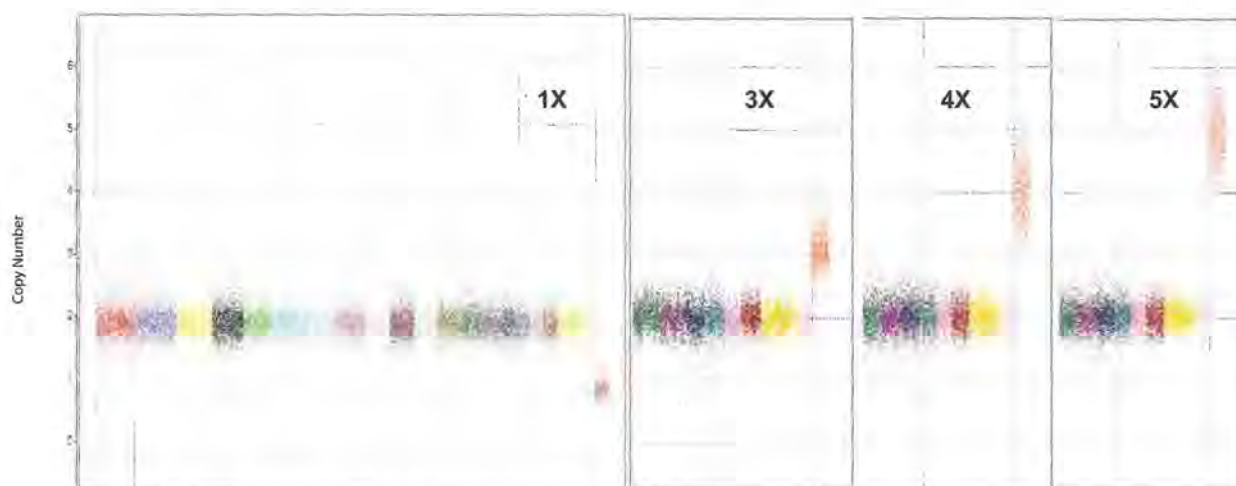


FIGURE 17-1. Accuracy and precision of CN measurements in X-chromosome titration cell lines. Four plots are shown (left to right): for 1X (male), and for 3X, 4X, and 5X cell lines. In each plot, the autosomes are ordered from left to right, with Chromosome 1 on the left (red) through Chromosome 22 on the right (pink), and with the X-chromosome markers (peach) on the extreme right. All points represent the CN sum of both alleles and are derived from unsmoothed CN measurements. Details of the results are discussed in the text.

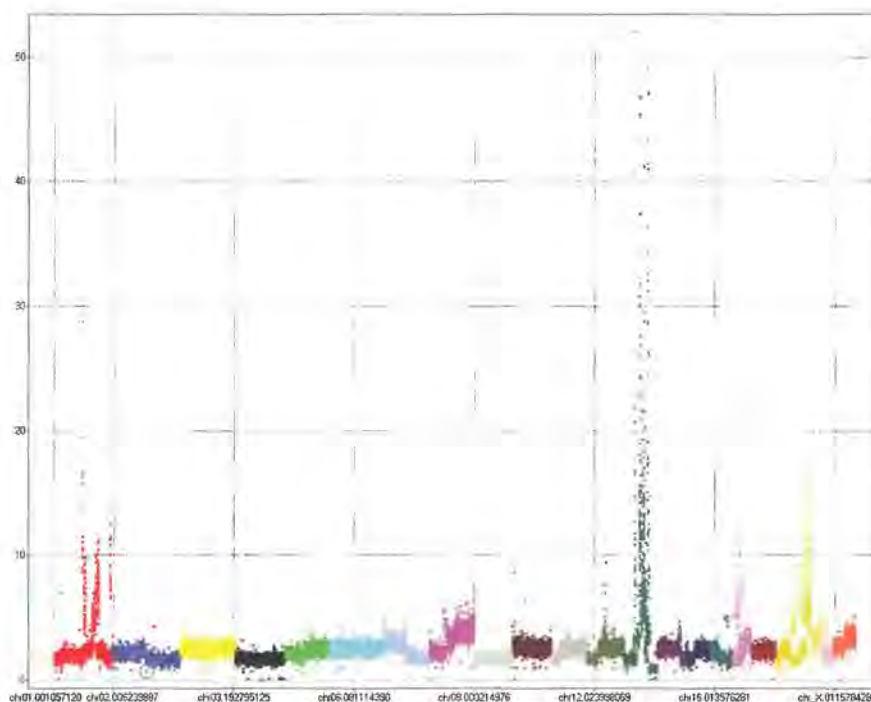


FIGURE 17-2. UACC812 cancer cell line amplifications and deletions. In this plot, also, the chromosomes in this cancer line are ordered from left to right, with a series of colored markers representing Chromosomes 1 through 22. The colored tracings depict amplification or deletion of the corresponding markers. All points represent the CN sum of both alleles and are derived from unsmoothed CN measurements.

purification procedure (Qiagen) and assayed under the same conditions as those described above for the cell line DNAs. A CN-versus-chromosome plot from a representative normal, male FFPE sample is shown in Figure 17-3. As shown in the cell line plots in Figure 17-1, each point represents a CN measurement for a single, unsmoothed marker. In Figure 17-3, the plot shows both accurate quantitation of the known 1-copy markers on the X chromosome (peach-colored, extreme right side of plot) and very good discrimination between these and the 2-copy autosomal markers (remaining series of colored markers representing Chromosomes 1 through 22). Thus, the MIP genotyping assay can provide high-quality CN measurements on DNA from both normal and FFPE samples.

SUMMARY AND CONCLUSIONS

Beyond affording highly accurate genotypes, the fundamental properties of the MIP assay system enable accurate and precise copy number measurements of highly multiplexed marker sets. Due to the assay's high specificity and large dynamic range, accurate measurement of deletions and amplifications over nearly 2 orders of magnitude can be made; the absence of cross-talk among alleles permits accurate allele-specific copy number determination without smoothing of adjacent markers. The small footprint of MIP probes on their genomic target (~40 bp) makes them well suited for copy number measurements in degraded DNA from FFPE samples. This has been demonstrated, here, at the 50,000-plex level and thus allows this technology to be applied to genome-wide assessments of deletions, amplifications, and LOH.

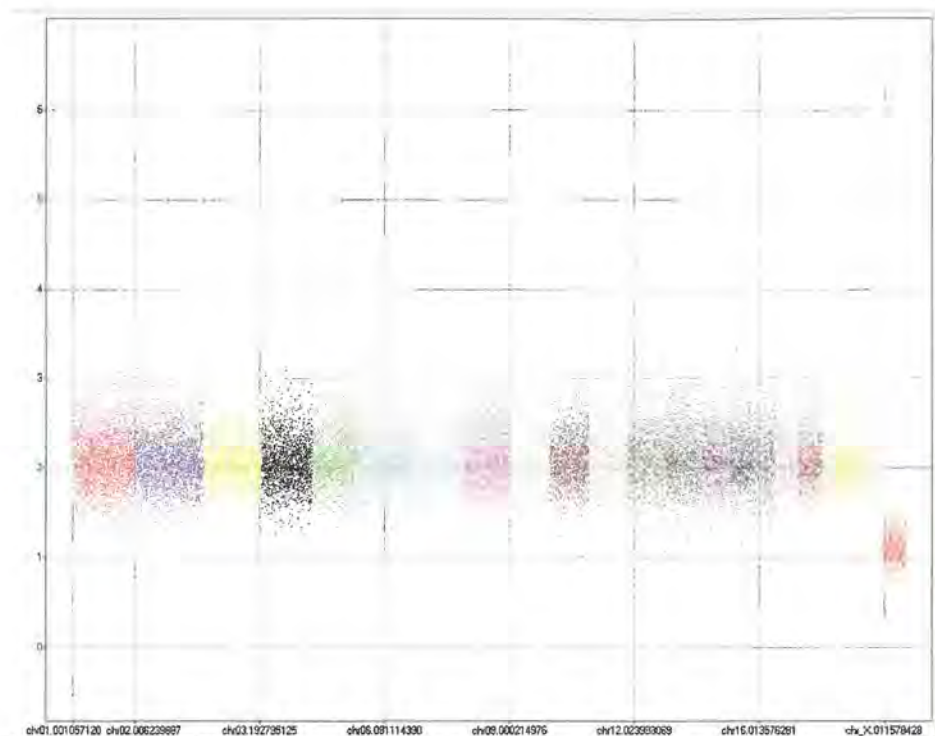


FIGURE 17-3. CN plot for normal male FFPE tissue (1X vs. 2X). DNA was purified from several 10- μ m slices of sample FFPE tissue blocks using a standard QIAmp DNA purification procedure (Qiagen). The organization of chromosomes is as described in Fig. 17-1 (2X autosomes from left to right and 1X chromosome on far right), and the CN genotyping assays were performed under the same conditions as described in Fig. 17-1 for the cell line DNAs. All points represent the CN sum of both alleles and are derived from unsmoothed CN measurements. Details of the results are discussed in the text.

ACKNOWLEDGMENTS

We thank the many people from ParAllele and Affymetrix who have labored devotedly in developing the MIP technology, with special acknowledgment to the algorithm and software developers who have been instrumental in developing the copy number analysis.

REFERENCES

- Slamon D.J., Leyland-Jones B., Shak S., Fuchs H., Paton V., Bajamonde A., Fleming T., Eiermann W., Wolter J., Pegram M., et al. 2001. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**: 783–792.
- van Beers E.H., Joosse S.A., Ligtenberg M.J., Fles R., Hogervorst F.B., Verhoef S., and Nederlof P.M. 2006. A multiplex PCR predictor for aCGH success of FFPE samples. *Br. J. Cancer* **94**: 333–337.
- Wang Y., Moorhead M., Karlin-Neumann G., Falkowski M., Chen C., Siddiqui F., Davis R.W., Willis T.D., and Faham M. 2005. Allele quantification using molecular inversion probes (MIP). *Nucleic Acid Res.* **33**: e183.

18

Microsatellite Markers for Linkage and Association Studies

Jeffrey Gulcher

deCODE Genetics, Reykjavik, Iceland, and Woodridge, Illinois

INTRODUCTION

During the 1990s and the first several years of this century, microsatellites or short tandem repeats (STRs) were the workhorse genetic markers for hypothesis-independent studies in human genetics, facilitating genome-wide linkage studies and allelic imbalance studies. However, with the rise of higher-throughput and cost-effective single-nucleotide polymorphism (SNP) platforms, we are currently

in the era of the SNP for genome scans. Nevertheless, it is important to note that microsatellites remain highly informative and useful measures of genomic variation for linkage and association studies. Their continued advantage in complementing SNPs lies in their greater allelic diversity than biallelic SNPs as well as in their population history, in which single-step expansion or contraction of the tandem repeat on the background of ancestral SNP haplotypes can break up common haplotypes, leading to greater haplotype diversity within the linkage disequilibrium (LD) block of interest. In fact, microsatellites have recently starred in association studies leading to widely replicated discoveries of type 2 diabetes (TCF7L2) and prostate cancer genes (the 8q21 region) (Amundadottir et al. 2006; Grant et al. 2006). At the end of the day, it will be important to catalog all variation, including SNPs, microsatellites, copy number variations, and polymorphic inversions in human genetic studies. This chapter describes the utilities of microsatellites and experimental approaches in their use.

INTRODUCTION, 253

Use of Microsatellites in Linkage Studies, 253

Use of Microsatellites in Association Studies, 255

Experimental Approaches to the Use of
Microsatellites in Case/Control Studies, 258

SUMMARY AND CONCLUSIONS, 260

ACKNOWLEDGMENTS, 260

REFERENCES, 260

Use of Microsatellites in Linkage Studies

Microsatellites have long been the variant of choice in linkage studies for several reasons. First, they have great diversity of alleles, averaging five to ten alleles. This leads to a heterozygosity rate of commonly used microsatellites from 50% to 80%; in contrast, the best possible heterozygosity rate of biallelic markers such as SNPs is 50%, but in practice the heterozygosities of SNPs average 30% or lower. This increase in heterozygosity leads to much higher information content when using microsatellites than when using the same number of SNPs. Advocates of SNPs for linkage have suggested that SNPs can be much more informative if one is doing multipoint linkage and using much larger numbers of SNPs (five- to tenfold more SNPs than microsatellites). That is certainly true if

the SNPs can be reliably typed and quality controlled and are not at too high a density to result in LD between SNP markers. Failure to ensure that linkage markers are not in LD results in very impressive artifactual linkage peaks, since linkage programs assume that all markers are in linkage equilibrium.

Second, microsatellites were the first variant that could be measured systematically from small amounts of DNA using PCR tied to high-throughput radioactive fragment separation by polyacrylamide electrophoresis, replacing the unwieldy Southern-blot-dependent restriction fragment length polymorphisms (RFLPs) by the early 1990s. Further scale-up of throughput by an order of magnitude came by replacing radioactive labeling with fluorescent labels of PCR fragments and automatic real-time reading of fragments separated by Applied Biosystem sequencers (Davies et al. 1994). The fluorescent labeling allowed for multiplexing markers labeled with different dyes and of different lengths in the same lane, as well as an internal molecular-weight standard within each lane on the slab gel, greatly increasing the accuracy of length assessment. Robotic mixing of DNA with PCR reagents and allele-calling software systems became necessary to keep up with this scale-up in throughput (Palsson et al. 1991). By the end of the 1990s, this high throughput was greatly expanded with the release of the capillary electrophoresis methods, which allowed automatic loading and prevented cross-contamination or cross-bleaching of lanes that sometimes plagued gel electrophoresis. One ABI 3730 can currently generate at least 30,000 genotypes per day. This sounds very modest compared to the current SNP chips. However, it is important to realize that microsatellite genotyping remains very cost-competitive with custom-designed SNP typing when one takes into account the increased information content per marker.

Third, microsatellites are easy to quality control. As with SNPs, microsatellites can occasionally suffer from allelic dropout (i.e., failure to measure one of the two alleles sometimes leading incorrectly to a homozygote genotype) due to a SNP within the primer site. The rate of this occurrence is similar between microsatellites and SNPs. However, when allelic dropout occurs in a particular SNP, it is much more difficult to detect than when it occurs with a microsatellite, given the large difference between heterozygosities. About three times as many genotyping errors due to allelic dropout will be missed for SNPs as for an equal number of microsatellites. If the number of SNPs required to generate good information content (90%) for a given set of pedigree structures is 5–10 times the number of microsatellites needed to yield the same level of information content, the number of SNP markers plagued by undetected genotyping errors may be 15–30 times that from an equivalent microsatellite scan. Therefore, quality control is even more important for accurate SNP linkage scans than for microsatellite scans.

Fourth, there is a high-resolution genetic map created with microsatellites but not for SNPs. The deCODE Genetics High-Resolution Genetic Map (HRGM) published in 2002 (Kong et al. 2002) was created with 5136 curated microsatellite markers using 1257 meiotic events (average resolution, 0.6 centiMorgan [cM]). The Marshfield map, which was created using more than 8000 microsatellite markers, has much lower resolution (3 cM) because the number of meiotic events within the pedigrees was only 188. Illumina created a map using 4000 SNP markers but used families with a similar number of meiotic events as used for the Marshfield map (Murray et al. 2004). Some have tried to interpolate genetic distance by overlaying the SNPs by physical position over deCODE's microsatellite markers. However, because physical distance is almost never proportional to genetic distance, significant errors may occur in estimation of genetic distance, which can lead to both false-positive and false-negative linkage signals (Gretarsdottir et al. 2002).

Microsatellite genotyping is not straightforward, and not many labs can efficiently validate and use microsatellite assays. Fortunately, there are several high-throughput laboratories that can competently and cost-effectively carry out microsatellite genotyping with very little error (less than 0.3–1%). Marshfield Clinic, led by the key developer of microsatellite genotyping, James Weber, was the pioneer in high-throughput microsatellite genotyping services (Weber and Broman 2001). Weber has recently moved to a private company (www.preventiongenetics.com). CIDR is another source for genotyping services, although it is focused on 10-cM framework marker scans and not

custom microsatellites to follow up initial linkage signals (www.cidr.jhmi.edu). The largest microsatellite genotyping laboratory offering contract services is the author's deCODE Genetics, which provides high-density genome scans using 500, 1000, 2000, and 5000 microsatellite markers (www.decode.com/genotyping). It also offers region-specific high-density panels for selectively increasing information content within initial linkage peaks, as well as efficient custom-design microsatellite marker discovery and testing for association studies (see below).

In summary, microsatellites have been the most widely used variant for linkage studies and remain so, even in this era of high-density SNP genotyping. Methods for their use in linkage are beyond the scope of this chapter, but may be found in Gulcher and Stefansson (2006).

Use of Microsatellites in Association Studies

It has become clear that most functional variants in genes contributing to common diseases are not nonsynonymous SNPs found in coding exons, but instead, appear to affect gene expression and RNA splicing. The wealth of data from the HapMap project allows more careful selection of a subset of nonredundant SNPs that tag independent haplotypes (tag-SNPs) for each LD block to act as surrogates for functional variants. To date, Illumina is the only vendor of genome-wide high-density SNP chips that has taken advantage of the HapMap data, using it to define and select tag-SNPs for its 300K, 550K, 650K, and 1M SNP bead chips. This contrasts with the alternative approach of random SNP selection, as practiced by vendors such as Affymetrix.

tag-SNP selection is covered in detail in Chapter 19. The end result of a whole-genome scan using tag-SNPs is that LD blocks are defined, within which SNPs are highly correlated and haplotype diversity is typically low. A potential disadvantage of the limited diversity in the haplotypes that define the haplotype space within the LD block can be overcome with microsatellites. We note that many groups report using a few tag-SNPs within each LD block and find that those SNPs define at least 90% of haplotypes in the LD block. However, it is very important to realize that haplotype coverage does not equal haplotype diversity. Therefore, many studies will be underpowered to find a correlation between the given LD block in the corresponding gene and the disease of interest if the frequency of the marker allele used is much greater than the frequency of the functional variant for which it is a surrogate (Ke et al. 2004).

In practice, one is attempting to use tag-SNPs and haplotypes as surrogate markers for underlying functional variants within the same LD block. The power of that surrogate is proportional to the degree of correlation between the surrogate and the functional variant as measured by R^2 (Ke et al. 2004). Unfortunately, if the frequencies of the surrogate and functional variants are not similar, there is a substantial loss in power for case/control association studies. For example, if one used a common haplotype or tag-SNP with a population allelic frequency of 30% to detect a functional variant with a frequency of 10%, there would be a 10-fold drop in power. Thus, for a case/control association study looking at 300 patients and 300 controls, the power would be equivalent to studying 30 patients and 30 controls, essentially wasting the expensive recruitment and phenotyping effort.

Therefore, it may be prudent to define not only haplotype coverage but also haplotype diversity within each LD block in the study. It is helpful to create a measure of haplotype diversity. If our goal is to have efficient surrogates of functional variants, and if we assume that we want to be powered to find variants as low as 5–10% allelic frequency in the general population, the ideal set of haplotypes would number ten, and each would have an equal allelic frequency of 10%. Therefore, we set an arbitrary measure of haplotype diversity as the sum of the frequencies of haplotypes that are each less than or equal to 10%. For example, if there are three common haplotypes each of 20% frequency and four other haplotypes each of 10%, haplotype diversity would equal 40% and we would say the remaining 60% of haplotype space has inadequate diversity.

Within the average LD block of 60 kb, six to ten common tag-SNPs may be selected to account for much of the information of the 50–60 HapMap SNPs. This can lead to substantial savings in

TABLE 18-1. Previous Estimates of the Microsatellite Mutation Rate per Generation

Mutation rate per generation	Reference
0.000194	Huang et al. (2002).
0.00045	Whittaker et al. (2003).
0.00056	deCODE experience

genotyping costs. Assuming no historical recombinations between the markers, there are expected to be no more than $N + 1$ independent haplotypes tagged by N SNPs within an LD block. Thus, only six tag-SNPs within a block of strong LD would define, at most, seven haplotypes (one variant per haplotype). However, seven independent haplotypes would have an average frequency of 14%, which would not come close to satisfying the arbitrary ideal criterion of having no haplotype more frequent than 10%. One solution to this problem is to attempt to divide common tag-SNP-defined haplotypes further by adding one or two microsatellites to each LD block of interest.

Microsatellites complement SNPs as useful markers for allelic association to disease, even though they are not as stable as SNPs during population history. However, their average mutation rate is lower than originally reported by Weber (see Table 18-1) (Beckman and Weber 1992). Our estimate, based on tens of thousands of microsatellite markers first qualified as Mendelian by a plate of 30 triads, is about 5×10^{-4} or 1 mutation per 2000 generations. This is consistent with two other studies that looked at a large number of microsatellites (Huang et al. 2002; Whittaker et al. 2003). This mutation rate is not ideal for maintaining strong LD between the surrogate marker and a high-frequency underlying functional disease variant. However, this mutation rate is actually ideal for common functional variants of intermediate frequency (e.g., 2–40%). As an example, microsatellite alleles can tag HLA extended haplotypes of intermediate frequency (Vorechovsky et al. 2001). Because one microsatellite can have 5–20 variants, it may effectively split up haplotype space within a given LD block. Furthermore, microsatellites can break up common ancestral SNP haplotypes to less common microsatellite/SNP haplotypes. This provides an efficient method to greatly increase haplotype diversity in any given LD block. For a real example, Table 18-2A shows 10 independent haplotypes defined by 6 tag-SNPs within a single LD block of about 100,000 bases. About 70% of haplotype space is not adequately covered by haplotypes of

TABLE 18-2. Haplotype Diversity within a Given LD Block

A. tag-SNPs only									B. Addition of 1 microsatellite									
RS1438	RS7627220	RS7651931	RS4894808	RS2124546	RS6787115	#Markers	#	Freq.	RS1438	RS7627220	D3S3725	RS7651931	RS4894808	RS2124546	RS6787115	#Markers	#	Freq.
3	2	1	2	4	3	6	60	0.321987	3	2	12	1	2	4	3	7	60	0.090376
1	1	3	3	2	1	6	60	0.161569	3	2	14	1	2	4	3	7	60	0.077723
1	1	3	3	4	3	6	60	0.130702	1	1	0	3	3	4	3	7	60	0.068971
3	2	1	3	4	3	6	60	0.108933	3	1	4	1	3	4	3	7	60	0.06465
3	1	1	3	4	3	6	60	0.064655	1	1	6	3	3	4	3	7	60	0.063353
3	2	1	3	2	1	6	60	0.054627	3	2	12	1	3	4	3	7	60	0.062264
3	2	1	2	2	1	6	60	0.035271	3	2	10	1	2	4	3	7	60	0.05929
3	2	1	3	4	1	6	60	0.034676	1	1	6	3	3	2	1	7	60	0.047222
3	1	3	3	2	1	6	60	0.034146	3	2	8	1	2	4	3	7	60	0.044808
1	1	1	3	4	3	6	60	0.009132	1	1	0	3	3	2	1	7	60	0.042155

(A) tag-SNPs only. (B) Addition of 1 microsatellite.

10% frequency or lower. Addition of just one microsatellite in Table 18-2B leads to diversification of the common haplotypes. Now, no SNP/microsatellite haplotype has a frequency greater than 10%. For example, the first SNP haplotype in Table 18-2A has a high frequency and is broken up into at least two haplotypes of lower frequency by the addition of the microsatellite (Table 18-2B). Note that the microsatellite can be too variable, breaking up haplotypes into too many low-frequency haplotypes that may be below the frequency of the functional variant. However, the haplotype association analysis can be done with two runs—SNPs with microsatellites, and SNPs without microsatellites with appropriate correction of *p*-values for twice the number of tests.

In practice, microsatellites have been quite useful for detecting significant associations to common diseases. For example, the widely replicated gene for type 2 diabetes, TCF7L2, was found by positional cloning through a genome-wide linkage scan of Icelandic families (Reynisdottir et al. 2003) with 1200 microsatellites. The follow-up locus-wide case/case association study of 1185 Icelandic patients and 931 controls applied 228 microsatellite markers over the two-LOD (log of the odds) drop interval (10.5 Mb) of the suggestive linkage peak on chromosome 10 (Grant et al. 2006). The average spacing of the microsatellites for the association study was 46 kb, resulting in an average of 1 or 2 microsatellites per LD block. Single-marker association using the Fisher exact test found a single microsatellite marker, DG10S478, within intron 3 of the TCF7L2 gene with significant association to the most common allele with a relative risk (RR) of 0.67 (uncorrected *p*-value of 2.1×10^{-9}). We noted that the other alleles with greater than 2% frequency showed a RR of 1.21 to 1.53 (see Table 18-3i). We merged all of the less common alleles into one common composite allele, called X, which gave a significant RR of 1.5 and a population allelic frequency of 0.276. The microsatellite association was replicated in a Danish cohort and a U.S. cohort (see Table 18-3ii). Note that the allelic frequencies and RR are quite similar across these three Caucasian populations. The overall RR is 1.56 with a corrected *p*-value of 7.8×10^{-15} and RR of 2.4 for homozygotes in a multiplicative model (Grant et al. 2006).

After we replicated this marker, we looked for HapMap SNPs in the same LD block within TCF7L2 that might show equivalent or even better risk. There were five SNPs that showed strong LD with the common microsatellite allele, and all showed significant association to type 2 diabetes, but none demonstrated as strong association or RR as the microsatellite (Grant et al. 2006). The best SNP, rs7903146, has now been replicated in about 20 populations, including Caucasians (e.g., Damcott et al. 2006; Florez et al. 2006; Groves et al. 2006) and Africans (Helgason et al. 2007). Our recent genome-wide association (GWA) scan using genotypes from a 317K Illumina bead chip array run on 1491 Icelandic T2DM patients and 4712 controls found the SNP with the best single-marker association to be the same SNP in TCF7L2, rs7903146, found through the positional cloning described above. The association had an uncorrected *p*-value of 3.7×10^{-13} , assuming a multiplicative model, which far exceeds the threshold of 1.3×10^{-7} .

TABLE 18-3. Microsatellite Association Studies

	Allele	Affected freq.	Control freq.	RR	Two-sided <i>P</i>
(i) All alleles of DG10S478					
Iceland (1185 / 931)	0	0.636	0.724	0.67	2.1×10^{-9}
	4	0.005	0.002	2.36	0.12
	8	0.093	0.078	1.21	0.090
	12	0.242	0.178	1.48	4.6×10^{-7}
	16	0.022	0.015	1.53	0.076
	20	0.001	0.003	0.39	0.17
(ii) Allele X of DG10S478					
Iceland (1185 / 931)	x	0.364	0.276	1.50 [1.31, 1.71]	2.1×10^{-9}
Denmark (228 / 539)	x	0.331	0.260	1.41 [1.11, 1.79]	0.0048
USA (361 / 530)	x	0.385	0.253	1.85 [1.51, 2.27]	3.3×10^{-9}
Combined	x	—	—	1.56 [1.41, 1.73]	4.7×10^{-18}

A second example of the utility of microsatellites for association studies is the widely replicated association of markers on chromosome 8q to prostate cancer. deCODE performed linkage studies using its genome-wide set of 1200 microsatellite markers to map prostate cancer to chromosome 8q (Amundadottir et al. 2006). Over a 10-Mb segment, 358 microsatellite assays were designed within the linkage peak at a density of 28 kb and used to genotype a case/control cohort with 869 unrelated Icelandic patients and 596 controls. Association was found to the common allele 0 of DG8S737 with a RR of 1.79 and was significant even when accounting for the number of tests. This microsatellite association was replicated in Swedish and U.S. Caucasian cohorts of prostate cancer. However, the most interesting aspect was that this microsatellite marker showed even stronger association to prostate cancer in African-Americans than in Caucasians (population-attributable risk of 16% versus 8%). This shows that microsatellites can be useful for confirmation studies across populations and even across ethnicities. The same microsatellite marker has been replicated in other populations (Freedman et al. 2006). Our recent GWA study, using the 317K SNP bead chip from Illumina, once again shows the strongest significant association to a SNP in strong LD to this microsatellite. Depending on one's point of view, these diabetes and prostate cancer studies validate one or more of the following four approaches: (1) Linkage is useful to map common disease genes, (2) GWA is useful to find common disease genes, (3) microsatellites are useful for discovery of common disease genes and their replication, and (4) the Illumina 317K bead chip produces data that appear to be unaffected by yield disparity or how alleles are called between cases and controls (which can lead to dramatic false-positive signals that drown out the true positives).

Experimental Approaches to the Use of Microsatellites in Case/Control Studies

This section gives more details on the use of microsatellites in practice for case/control association studies within candidate regions or genes. For candidate gene studies, microsatellite markers are useful complements to tag-SNPs that are selected for each LD block. A typical study selects about five tag-SNP markers in each LD block overlapping the candidate gene, using SNP selection programs such as Tagger (de Bakker et al. 2006). One or two microsatellite markers can also be added to each LD block to diversify the haplotype space.

Microsatellite markers may be found using the tandem repeat finder (TRF) program from Benson (1999). Primers are designed using adequate criteria for PCR assays, and primers that include repeats or known SNPs in the sequence are excluded. We estimate the likelihood of a successful microsatellite (i.e., a readable and polymorphic marker) by associating the characteristics of the microsatellite, including monomer length, microsatellite length, and repeat structure, with experimental success in a set of more than 25,000 microsatellite markers previously designed and tested at deCODE. We use this information to rank the putative microsatellites according to expected quality, and we have identified more than 40,000 putative microsatellites in the genome that we classify as rank 1. In our experience, more than 70% of these turn out to be usable. However, a good portion of the predicted microsatellites in the lower ranks also result in validated assays such that densities of 10–40 kb are usually achievable in candidate regions.

We have found microsatellite markers useful to ultrafine-map linkage peak intervals during the follow-up case/control association study to isolate the disease gene. Although one may choose to select only those linkage peaks that meet criteria for significant genome-wide linkage, we have found that even suggestive linkage peaks are useful for picking regions of the genome for ultrafine mapping. In fact, in the type 2 diabetes and prostate cancer studies described above which resulted in the most widely replicated genes in those diseases to date, the linkage peaks were only modest. However, it is important to point out that we apply affecteds-only allele-sharing linkage analysis without specification of the inheritance model. This greatly decreases the number of tests, since a parametric linkage study may look at more than 100 models that can lead to high rates of false-positive linkage.

We also find microsatellite markers useful to follow up on the strongest association signals in high-density GWA studies. Most of the strongest associations in GWA studies are due to very common alleles with very modest risk, such as 1.2–1.4. It is useful to follow up on such signals in the discovery population using extra markers within the same LD block (as well as other LD blocks in the same gene), including additional tag-SNPs and one or two microsatellites. This may lead to better capture of the original signal but with a higher RR, or it may lead to the discovery of other disease-associated variants in the same LD block or in other parts of the same gene. If one finds even stronger signals than the original tag-SNP on the chip, this may improve the chances of replication if it is a true disease gene.

Data from genome-wide linkage studies using microsatellites are still useful in the current GWA era, especially when one must correct the p -values by 317,000 or 550,000 SNP marker tests. One can prespecify linkage intervals that are used in the primary analysis in the GWA study. For example, if one selects the top five or six suggestive linkage peaks covering only 3% of the genome, then the correction may be for only 10,000 SNP tests within those peaks rather than for all 317,000 tests. The secondary analysis would run the rest of the genome.

For our strongest linkage peaks, the high-density SNP genotypes from the bead chip represent at the least very efficient ways to ultrafine-map the linkage intervals for the follow-up case/control association. However, it is useful to complement such high-density data with microsatellites for increased haplotype diversity, since there are only six to nine tag-SNPs per LD block on average.

Defining haplotype diversity is one method for determining adequacy of coverage of each candidate gene or interval LD block. As described above, one goal is to use enough SNP and microsatellite markers to leave no haplotype in a given LD block with a frequency greater than 10%. Inadequate haplotype diversity in an LD block may lead to the situation in which the disease-associated functional variant is on the background of a very common ancestral haplotype. If the variant is only on a portion of the common haplotype, that haplotype may not show tight enough correlation to the disease, and the gene association will be missed. In this case, the haplotype was a poor surrogate for the underlying functional variant. It is straightforward to integrate microsatellite markers into the LD patterns. The standard definitions of D' and R^2 are extended to include microsatellite markers by determining the average of the values for all possible allele combinations of the two markers weighted by the marginal allele probabilities. Single point and haplotype associations comparing the patient group with controls are carried out as described in other chapters in this manual. Although there are no accepted criteria to define statistical significance of a haplotype or SNP association, it is clear that p -values should be corrected by the number of tests performed (including number of markers [alleles-1], phenotype runs, and statistical methods used). It is also clear that replication in other independent cohorts of the disease is necessary.

For any significant association results, one must check for population stratification. It is important to confirm the self-reported ethnicity in each group and to ensure that the genetic backgrounds of the patient group match those of the control groups. Microsatellite marker sets are especially informative in distinguishing major ethnic groups. To evaluate genetically estimated ancestry of the study cohorts, we use a panel of 75 unlinked microsatellite markers from about 2000 microsatellites genotyped in a multiethnic cohort of 35 European-Americans from Baltimore; 88 African-Americans from Pittsburgh, Baltimore, and North Carolina; 34 Chinese (Canton); and 29 Indian-Americans (Zapotec) (Helgadóttir et al. 2006). Of the 2000 microsatellite markers, the selected set showed the most significant differences between the European-Americans, African-Americans, and Asians, and implicitly Indians, and also had good quality and yield. Another study by Tang et al. (2005) also used 31 of these markers.

Using genotype data from these markers on the ethnic plates (i.e., the Caucasian, West African, and Asian HapMap plates, each of 96 samples), Structure software will estimate the genetic background of test individuals (Pritchard et al. 2000; Falush et al. 2003). Structure infers the allele fre-

quencies of K ancestral populations on the basis of multilocus genotypes from a set of individuals and a user-specified value of K, and assigns a proportion of ancestry from each of the inferred K populations to each individual. One can use the ethnicity panel in two ways. First, one can include in the analysis only those Caucasians who have at least 90% European ancestry in patients and controls. Second, we have modified the NEMO haplotype association algorithm to adjust for differences in genetic background between cases and controls so that one can include all Caucasians according to self-report (Helgadóttir et al. 2006).

SUMMARY AND CONCLUSIONS

In summary, microsatellites have demonstrated utility in both linkage and association studies. The fact that they have higher mutation rates than single-base substitutions allows them to break up apparently immutable common SNP haplotypes into lower frequencies that may better match functional variants of intermediate frequency or the rare functional variants they are meant to detect.

ACKNOWLEDGMENTS

The author thanks Struan Grant for creating the haplotype diversity data in Table 18-2 while he was working at deCODE.

REFERENCES

- Amundadóttir L.T., Sulem P., Gudmundsson J., Helgason A., Baker A., Agnarsson B.A., Sigurdsson A., Benediktsson K.R., Cazier J.B., Sainz J., et al. 2006. A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**: 652–658.
- Beckman J.S. and Weber J.L. 1992. Survey of human and rat microsatellites. *Genomics*. **12**: 627–631.
- Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**: 573–580.
- Damcott C.M., Pollin T.L., Reinhart L.J., Ott S.H., Shen H., Silver K.D., Mitchell B.D., and Shuldiner A.R. 2006. Polymorphisms in the transcription factor 7-like 2 (*TCF7L2*) gene are associated with type 2 diabetes in the Amish: Replication and evidence for a role in both insulin secretion and insulin resistance. *Diabetes* **55**: 2654–2659.
- Davies J.L., Kawaguchi Y., Bennett S.T., Copeman J.B., Cordell H.J., Pritchard L.E., Reed P.W., Gough S.C., Jenkins S.C., Palmer S.M., et al. 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**: 130–136.
- de Bakker P.I., Burt N.P., Graham R.R., Guiducci C., Yelensky R., Drake J.A., Bersaglieri T., Penney K.L., Butler J., Young S., et al. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**: 1298–1303.
- Falush D., Stephens M., and Pritchard J.K. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- Florez J.C., Jablonski K.A., Bayley N., Pollin T.L., de Bakker P.I., Shuldiner A.R., Knowler W.C., Nathan D.M., and Altshuler D. (Diabetes Prevention Program Research Group). 2006. *TCF7L2* polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N. Engl. J. Med.* **355**: 241–250.
- Freedman M.L., Haiman C.A., Patterson N., McDonald G.J., Tandon A., Waliszewska A., Penney K., Steen R.G., Ardlie K., John E.M., et al. 2006. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci.* **103**: 14068–14073.
- Grant S.F., Thorleifsson G., Reynisdóttir L., Benediktsson R., Manolescu A., Sainz J., Helgason A., Stefansson H., Emilsson V., Helgadóttir A., et al. 2006. Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**: 320–323.
- Gretarsdóttir S., Sveinbjornsdóttir S., Jonsson H.H., Jakobsson F., Einarsson E., Agnarsson U., Shkolny D., Einarsson G., Gudjonsdóttir H.M., Valdimarsson E.M., et al. 2002. Localization of a susceptibility gene for common forms of stroke to 5q12. *Am. J. Hum. Genet.* **70**: 593–603.
- Groves C.J., Zeggini E., Minton J., Frayling T.M., Weedon M.N., Rayner N.W., Hitman G.A., Walker M., Wiltshire S., Hattersley A.T., and McCarthy M.I. 2006. Association analysis of 6,736 U.K. subjects provides replication and confirms *TCF7L2* as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55**: 2640–2644.
- Gulcher J. and Stefansson K. 2006. Positional cloning: Complex cardiovascular traits. *Methods Mol. Med.* **128**: 137–152.
- Helgadóttir A., Manolescu A., Helgason A., Thorleifsson G., Thorsteinsdóttir U., Gudbjartsson D.E., Gretarsdóttir S., Magnusson K.P., Gudmundsson G., Hicks A., et al. 2006. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat. Genet.* **38**: 68–74.
- Helgason A., Palsson S., Thorleifsson G., Grant S.F., Emilsson V., Gunnarsdóttir S., Adeyemo A., Chen Y., Chen G., Reynisdóttir I., et al. 2007. Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39**: 218–225.
- Huang Q.Y., Xu F.H., Shen H., Deng Y.J., Liu Y.J., Liu Y.Z., Li J.L., Recker R.R., and Deng H.W. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* **70**: 625–634.

- Ke X., Durrant C., Morris A.P., Hunt S., Bentley D.R., Deloukas P., and Cardon L.R. 2004. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13**: 2557–2565.
- Kong A., Gudbjartsson D.F., Sainz J., Jonsdottir G.M., Gudjonsson S.A., Richardsson B., Sigurdardottir S., Barnard J., Hallbeck B., Masson G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31**: 241–247.
- Murray S.S., Oliphant A., Shen R., McBride C., Stecke R.J., Shannon S.G., Rubano T., Kermani B.G., Fan J.B., Chee M.S., and Hansen M.S. 2004. A highly informative SNP linkage panel for human genetic studies. *Nat. Methods* **1**: 113–117.
- Palsson B., Palsson E., Perlin M., Gudbjartsson H., Stefansson K., and Gulcher J. 1991. Using quality measures to facilitate allele calling in high-throughput genotyping. *Genome Res.* **9**: 1002–1012.
- Pritchard J.K., Stephens M., and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- Reynisdottir I., Thorleifsson G., Benediktsson R., Sigurdsson G., Emilsson V., Einarsdottir A.S., Hjorleifsdottir E.E., Orlygsdottir G.T., Bjornsdottir G.T., Saemundsdottir J., et al. 2003. Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. *Am. J. Hum. Genet.* **73**: 323–335.
- Tang H., Quertermous T., Rodriguez B., Kardis S.L., Zhu X., Brown A., Pankow J.S., Province M.A., Hunt S.C., Boerwinkle E., Schork N.J., and Risch N.J. 2005. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am. J. Hum. Genet.* **76**: 268–275.
- Vorechovsky I., Kralovicova J., Laycock M.D., Webster A.D., Marsh S.G., Madrigal A., and Hammarstrom L. 2001. Short tandem repeat (STR) haplotypes in HLA: An integrated 50-kb STR/linkage disequilibrium/gene map between the RING3 and HLA-B genes and identification of STR haplotype diversification in the class III region. *Eur. J. Hum. Genet.* **9**: 590–598.
- Weber J.L. and Broman K.W. 2001. Genotyping for human whole-genome scans: Past, present, and future. *Adv. Genet.* **42**: 77–96.
- Whittaker J.C., Harbord R.M., Boxall N., Mackay I., Dawson G., and Sibby R.M. 2003. Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164**: 781–787.