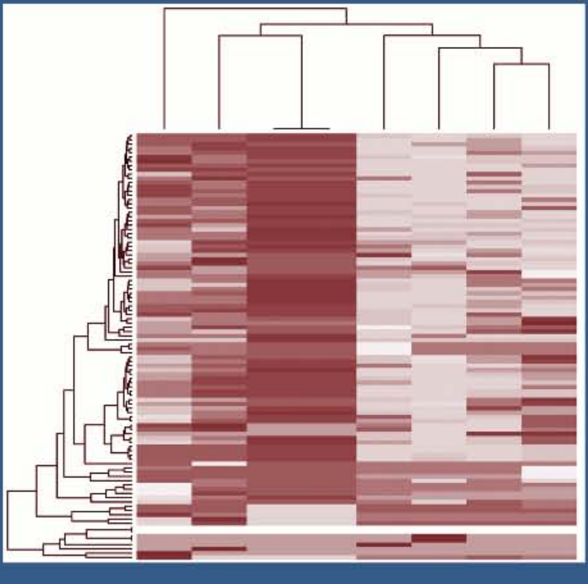
DNA MICROARRAYS AND RELATED GENOMICS TECHNIQUES

Design, Analysis, and Interpretation of Experiments



EDITED BY

DAVID B. ALLISON
GRIER P. PAGE
T. MARK BEASLEY
JODE W. EDWARDS

Chapman & Hall/CRC
Taylor & Francis Group

DNA MICROARRAYS AND RELATED GENOMICS TECHNIQUES

Design, Analysis, and Interpretation of Experiments

Biostatistics: A Series of References and Textbooks

Series Editor
Shein-Chung Chow
Duke Clinical Research Institute
Duke University
Durham, NC, USA

1. *Design and Analysis of Animal Studies in Pharmaceutical Development,* Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications,* James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded,* Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy,* Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective,* Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick
6. *Difference Equations with Public Health Applications,* Lemuel A. Moyé and Asha Seth Kapadia
7. *Medical Biostatistics,* Abhaya Indrayan and Sanjeev B. Sarmukaddam
8. *Statistical Methods for Clinical Trials,* Mark X. Norleans
9. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation,* Mikel Aickin
10. *Statistics in Drug Research: Methodologies and Recent Developments,* Shein-Chung Chow and Jun Shao
11. *Sample Size Calculations in Clinical Research,* Shein-Chung Chow, Jun Shao, and Hansheng Wang
12. *Applied Statistical Design for the Researcher,* Daryl S. Paulson
13. *Advances in Clinical Trial Biostatistics,* Nancy L. Geller
14. *Statistics in the Pharmaceutical Industry, 3rd Edition,* Ralph Buncher and Jia-Yeong Tsay
15. *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments,* David B. Allsion, Grier P. Page, T. Mark Beasley, and Jode W. Edwards

DNA MICROARRAYS AND RELATED GENOMICS TECHNIQUES

Design, Analysis, and Interpretation of Experiments

EDITED BY

DAVID B. ALLISON
GRIER P. PAGE
T. MARK BEASLEY
JODE W. EDWARDS

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Published in 2006 by
Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8247-5461-1 (Hardcover)
International Standard Book Number-13: 978-0-8247-5461-7 (Hardcover)
Library of Congress Card Number 2005050488

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

DNA microarrays and related genomics techniques : design, analysis, and interpretation of experiments / editors, David B. Allison ... [et al.].
p. cm. -- (Biostatistics)
Includes bibliographical references and index.
ISBN 0-8247-5461-1 (alk. paper)
1. DNA microarrays. I. Allison, David B. (David Bradley), 1963- II. Biostatistics (New York, N.Y.)

QP624.5.D726.D636 2005
572.8'636--dc22 2005050488

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Series Introduction

The primary objectives of the *Biostatistics Book Series* are to provide useful reference books for researchers and scientists in academia, industry, and government, and also to offer textbooks for undergraduate and graduate courses in the area of biostatistics and bioinformatics. This book series will provide comprehensive and unified presentations of statistical designs and analyses of important applications in biostatistics and bioinformatics, such as those in biological and biomedical research. It gives a well-balanced summary of current and recently developed statistical methods and interpretations for both statisticians and researchers/scientists with minimal statistical knowledge who are engaged in the field of applied biostatistics and bioinformatics. The series is committed to providing easy-to-understand, state-of-the-art references and textbooks. In each volume, statistical concepts and methodologies will be illustrated through real world examples whenever possible.

In recent years, the screening of thousands of genes using the technique of expression microarrays has become a very popular topic in biological and biomedical research. The purpose is to identify those genes that may have an impact on clinical outcomes of a subject who receives a test treatment under investigation and consequently establish a medical predictive model. Under a well-established predictive model, we will be able not only to identify subjects with certain genes who are most likely to respond to the test treatment, but also to identify subjects with certain genes who are most likely to experience (serious) adverse events. This concept plays an important role in the so-called personalized medicine research. This volume summarizes various useful experimental designs and statistical methods that are commonly employed in microarray studies. It covers important topics in DNA microarrays and related genomics research such as normalization of microarray data, microarray quality control, statistical methods for screening of high-dimensional biology, and power and sample size calculation. In addition, this volume provides useful approaches to microarray studies such as clustering approaches to gene microarray data, parametric linear models, nonparametric procedures, and Bayesian analysis of microarray data. It would be beneficial to biostatisticians, biological and biomedical researchers, and pharmaceutical scientists who are engaged in the areas of DNA microarrays and related genomics research.

Shein-Chung Chow

Preface

WHAT ARE MICROARRAYS?

Microarrays have become a central tool used in modern biological and biomedical research. This book concerns expression microarrays, which for the remainder of the book, we simply refer to as microarrays. They are tools that permit quantification of the amount of all mRNA transcripts within a particular biological specimen. There are several different technologies for producing microarrays that have different strengths and weaknesses. These platforms and alternatives are discussed in Chapter 1 by Gaffney et al.

Viewed as "hot" and highly exotic tools as recently as the late 1990s, they are now ubiquitous in biological research and the modern biological researcher can no more be unaware and unexposed to microarray research and its results than one can remain ignorant of clinical trials, questionnaire studies, genome scans, animal models, or any of the other tools that have become standard parts of our armamentarium. Although much development in microarray research methodology is still needed, it is clear that microarrays are here to stay.

WHY THIS BOOK?

In one sense, microarrays are simply measurement assays. Just as one can measure, for example, the amount of insulin (which is the product of a gene) in blood, we can measure the products of genes with microarrays in any tissue. What distinguishes microarrays from traditional approaches is their "omic" nature. That is, they have capacity to measure *all* gene transcripts at once. This ushered in the subfield of *transcriptomics*. A particular challenge is that because of the expense of microarray research and the fact that it is often directed at basic discovery and hypothesis generation/exploration missions, the number of variables (transcripts) available in microarray studies tends to exceed the number of cases (subjects) by several orders of magnitude. Traditional statistical approaches to design and analysis were not developed in the context of such high dimensional and small sample problems. We and many others now find that our training in traditional statistical methods is not especially well-suited to such situations.

We (the editors) were first introduced to the analysis of microarray data ca. 1999. At that time, there were almost no statistical papers providing approaches to analyze microarray data or design microarray studies from a statistical perspective. By 2003, this situation had changed dramatically and we estimate that there were hundreds of papers thereon (Mehta et al., 2004). This overwhelming deluge of methods from these papers is quite daunting to either the applied investigator looking for methodologies to utilize or the methodologist trying to keep up with the field.

As part of the research efforts funded by the National Science Foundation, we have hosted an annual retreat for scientists interested in analytic methods for microarray research for the last five years. The impetus for this book came in part from discussions held at those retreats. We felt there was a need for a book that consolidated many of the existing methodologic advances and compiled many of the issues and methods into a single volume. This book is aimed at both the investigator who will conduct analyses of microarray data and at the methodologists who will evaluate existing and develop future methodologies.

WHAT IS HERE?

We have structured this book in a manner that we believe parallels the steps that an investigator or an analyst will go through while conducting and analyzing a microarray experiment from conception to interpretation. We begin with the most foundational issues: ensuring the quality and integrity of the data and assessing the validity of the statistical methods we employ. We then move on to the often neglected, but critical aspects of designing a microarray experiment. Gadbury et al. (Chapter 5) address issues such as power and sample size, where only very recently have developments allowed such calculations in a high dimensional context. The third section of the book is the largest, addressing issues of the analysis of microarray data. The size of this section reflects both the variety of topics and the amount of effort investigators have devoted to developing new methodologies. Finally, we move on to the intellectual frontier — interpretation of microarray data. New methods for facilitating and affecting formalization of the interpretation process are discussed. The movement to make large high dimensional datasets public for further analysis and methods for doing so are also addressed.

WHAT IS NOT HERE?

This book is not a detailed exposition of software packages (although some are mentioned in specific chapters), biochemistry, or the mechanics of the physical production of microarrays or biological specimens for analysis via microarrays. Interested readers should consult other more topical books in these areas (Jordan, 2001; Grigorenko, 2002; Ye and Day, 2003) Many closely related disciplines such as proteomics and metabolomics are not discussed in any depth although the astute reader will readily see the commonalities among the statistical and design approaches that can be applied to such data.

THE FUTURE

There is no question that this field will continue to advance rapidly and some of the specific methodologies we discuss herein will be replaced by new advances in the near future. Nevertheless, we believe the field is now at a point where a foundation of key categories of methods has been laid and begun to settle. Although the details may change, we believe that the majority of the key principles described herein and the foundational categories are likely to stand the test of time and serve as a useful guide to the reader. We look forward to new biological knowledge that we anticipate will emerge from the evermore sophisticated technologies and analysis as well as the exciting new statistical advances sure to come.

REFERENCES

Girgorenko E.V. (2002) *DNA Arrays: Technologies and Experimental Strategies*. CRC Press, Boca Raton, FL.

Jordan B.R. (ed.) (2001) *DNA Microarray: Gene Expression Applications*. Springer-Verlag, Berlin.

Mehta T., Tanik M., and Allison D.B. (2004) Towards sound epistemological foundation of statistical methods for high-dimensional biology. *Nature Genetics* 36: 943–947.

Ye S. and Day I.N.M. (2003) *Microarray and Microplates*. Bios Press, Oxford.

Editors

David B. Allison received his Ph.D. from Hofstra University in 1990. He then completed a postdoctoral fellowship at the Johns Hopkins University School of Medicine and a second postdoctoral fellowship at the NIH-funded New York Obesity Research Center at St. Luke's/Roosevelt Hospital Center. He was a research scientist at the New York Obesity Research Center and Associate Professor of Medical Psychology at Columbia University College of Physicians and Surgeons until 2001. In 2001, he joined the faculty of the University of Alabama at Birmingham where he is currently Professor of Biostatistics, Head of the Section on Statistical Genetics, and Director of the NIH-funded Clinical Nutrition Research Center. He has authored over 300 scientific publications and edited three books. He has won several awards, including the 2002 Lilly Scientific Achievement Award from the North American Association for the Study of Obesity and the 2002 Andre Mayer Award from the International Association for the Study of Obesity, holds several NIH and NSF grants, served on the Council of the North American Association for the Study of Obesity from 1995 to 2001, and has been a member of the Board of Trustees for the International Life Science Institute, North America, since January 2002. He serves on the editorial boards of *Obesity Reviews; Nutrition Today; Public Library of Science (PLOS) Genetics; International Journal of Obesity; Behavior Genetics; Computational Statistics and Data Analysis; and Human Heredity*.

Dr. Allison's research interests include obesity, quantitative genetics, clinical trials, and statistical and research methodology.

Grier P. Page, Ph.D. was born in Cleveland, Ohio in 1970. He received his B.S. in Zoology and Molecular Biology from the University of Texas, Austin. Then he received his M.S. and Ph.D. in Biomedical Sciences from the University of Texas–Health Sciences Center—Houston under the mentorship of Drs. Eric Boerwinkle and Christopher Amos. Dr. Page has been involved in the use and analysis of microarrays since 1998 for expression, genomics, and genotyping. He is very active in the development of new methods for the analysis of microarray data as well as methods and techniques for the generation on the highest quality microarray data. He uses microarrays in his research in the mechanisms of cancer development, nutrient production, and nutrient gene interactions especially in cancer and plants. He is currently a member of the Section on Statistical Genetics, Department of Biostatistics the University of Alabama, Birmingham.

T. Mark Beasley, Ph.D. is Associate Professor of Biostatics and a member of the Section on Statistical Genetics at the University of Alabama at Birmingham. He is the leader of the measurement and inferences teams for a funded National Science Foundation (NSF) grant to further the development of microarray analysis methods. He has a Ph.D. in Statistics and Measurement from Southern Illinois University and a strong research record in the area of statistical methodology, focused in methodological problems in statistical genetics; nonparametric statistics; simulation studies; and the use of linear models. He also has a strong background in measurement theory and the multivariate methods (e.g., factor analysis, structural equation models; regression models). Dr. Beasley teaches courses on Applied Multivariate Analysis and General Linear Models at UAB and is currently Editor of *Multiple Linear Regression Viewpoints*, a journal focused on applications of general linear models and multivariate analysis. He has published articles in applied statistics journals such as the *Journal of Educational & Behavioral Statistics, Journal of the Royal Statistical Society, Computational Statistics & Data Analysis, Multivariate Behavioral Research*, and *Communications in Statistics*. He has also published articles on methodological problems in statistical genetics in leading journals such as the *American Journal of Human Genetics; Behavior Genetics; Genetic Epidemiology; Genetics, Selection, and Evolution* and *Human Heredity*.

Jode W. Edwards received a Ph.D. in plant breeding and genetics with a minor in statistics from Iowa State University in 1999. He then spent 3 years with Monsanto Company as a statistical geneticist working in the areas of marker-assisted plant breeding and QTL mapping. Dr. Edwards joined the Section on Statistical Genetics as a Postdoctoral Fellow in 2002. His research involved application of Empirical Bayes methods to microarray analysis and development of software for microarray data analysis. Using SAS as a prototyping platform, he designed experimental versions of the HDBStat! software that is now distributed by the Section on Statistical Genetics. Additionally, Dr. Edwards helped initiate efforts to build the microarray Power Atlas, a tool to assist investigators in designing microarray experiments. In 2004, he completed his postdoctoral studies and assumed a position as a Research Geneticist with the Agricultural Research Service of the United States Department of Agriculture, in Ames, IA. His research is focused on quantitative genetics of maize, application of Bayesian methods in plant breeding, and breeding for amino acid balance in maize protein.

Contributors

David B. Allison
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

T. Mark Beasley
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Jacob P.L. Brand
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Jane Y. Chang
Department of Applied Statistics and Operational Research
Bowling Green State University
Bowling Green, Ohio

Kei-Hoi Cheung
Department of Genetics
Center for Medical Informatics
Yale University School of Medicine
New Haven, Connecticut

Tzu-Ming Chu
SAS Institute
Cary, North Carolina

Christopher S. Coffey
University of Alabama at Birmingham
Birmingham, Alabama

Stacey S. Cofield
University of Alabama at Birmingham
Birmingham, Alabama

Robert R. Delongchamp
Division of Biometry and Risk Management
National Center for Toxicological Research
Jefferson, Arizona

Shibing Deng
SAS Institute
Cary, North Carolina

Jode W. Edwards
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

David Finkelstein
Hartwell Center
St. Jude Children's Research Hospital
Memphis, Tennessee

Gary L. Gadbury
Department of Mathematics and Statistics
University of Missouri-Rolla
Rolla, Missouri

Patrick M. Gaffney
University of Minnesota
Minneapolis, Minnesota

Elizabeth Garrett-Mayer
Division of Oncology Biostatistics
Sidney Kimmel Comprehensive Cancer Center
Baltimore, Maryland

Pulak Ghosh
Department of Mathematics and Statistics
Georgia State University
Atlanta, Georgia

Bernard S. Gorman
Nassau Community College and
Hofstia University
Garden City, New York

Jason C. Hsu
Department of Statistics
Ohio State University
Columbus, Ohio

Michael Janis
Department of Chemistry,
Biochemistry, and
Molecular Biology
University of California
at Los Angeles
Los Angeles, California

Christina M. Kendziorski
Department of Biostatistics and
Medical Informatics
University of Wisconsin-Madison
Madison, Wisconsin

Jeanne Kowalski
Division of Oncology Biostatistics
Johns Hopkins University
Baltimore, Maryland

Jeffrey D. Long
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Tapan Mehta
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Kathy L. Moser
Department of Medicine
Institute of Human Genetics and
Center for Immunology
University of Minnesota
Medical School
Minneapolis, Minnesota

Michael V. Osier
Yale Center for Medical Informatics
Yale University School of Medicine
New Haven, Connecticut

Grier P. Page
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Rudolph S. Parrish
Department of Bioinformatics and
Biostatistics School of Public
Health and Information Sciences
University of Louisville
Louisville, Kentucky

Jacques Retief
Iconix Pharmaceuticals
Mountain View, California

Douglas M. Ruden
Department of Environmental Health
Sciences
University of Alabama at Birmingham
Birmingham, Alabama

Chiara Sabatti
Department of Human Genetics
University of California at Los Angeles
Los Angeles, California

Kathryn Steiger
Division of Biostatistics
University of California at Berkeley
Berkeley, California

Murat Tanik
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Alan Williams
Affymetrix
Santa Clara, California

Russell D. Wolfinger
SAS Institute
Cary, North Carolina

Qinfang Xiang
Department of Mathematics and Statistics
University of Missouri-Rolla
Rolla, Missouri

Stanislav O. Zakharkin
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Kui Zhang
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, Alabama

Zhen Zhang
Department of Pathology
School of Medicine
Johns Hopkins University
Baltimore, Maryland

Contents

1 Microarray Platforms

Patrick M. Gaffney and Kathy L. Moser

CONTENTS

1.1 INTRODUCTION

As with the development of any novel and potentially powerful technology, the prospect of revealing new information that may dramatically change our understanding of biological processes can generate much excitement. Such is true for the emerging genomic approaches that make possible high-density assays using microarray platforms. Indeed, it is difficult, if not impossible, to imagine any area of biology that could not be affected by the wide range of potential applications of microarray technology. Numerous examples, such as those from the field of oncology, provide striking evidence of the power of microarrays to bring about extraordinary advances in molecularly defining important disease phenotypes that were otherwise unrecognized using conventional approaches such as histology.

In this chapter, we present a general overview of microarray platforms currently in use with particular emphasis on high-density DNA arrays. We touch briefly on approaches to data analysis leaving most of the details for the ensuing chapters. For those just entering the microarray arena or interested in more details, a series of particularly useful reviews have recently been published that take stock of the latest developments and discuss the most pressing challenges of this technology [1].

1.2 MICROARRAY TECHNOLOGY

Microarray technology provides an unprecedented and uniquely comprehensive probe into the coordinated workings of entire biological pathways and genomic-level

TABLE 1.1
Potential Objectives of Studies Utilizing Microarray Technology

1. Distinguish patients from normal controls
2. Identify subsets of patients
3. Characterize host responses
4. Examine cellular pathways
5. Compare alternative experimental conditions
6. Examine drug response
7. Follow temporal changes in gene expression
8. Identify candidate genes for genetic studies

processes. In general terms, microarrays refer to a variety of platforms in which high density assays are performed in parallel on a solid support. Thousands to tens of thousands of datapoints may be generated in each experiment. The growth of scientific literature since the mid-1990s may provide some indication for the potential impact of this technology in biomedical sciences. A majority of applications have been in oncology, although many examples from other fields are rapidly emerging and include examination of host response to pathogens, examination of drug responses, identification of temporal changes in gene expression, and comparisons of various experimental conditions.

Three major types of microarrays exist — tissue, protein, and DNA. Tissue microarrays immobilize small amounts of tissue from biopsies of multiple subjects on glass slides for immunohistochemical processing, while protein arrays immobilize peptides or intact proteins for detection by antibodies or other means (see Section 1.3). For the last several years, much excitement and attention has focused on DNA microarrays and most of this book will concentrate on DNA microarray analysis. Regardless of the specific platform used, these approaches offer new opportunities to address biologic questions in a way never possible before. Table 1.1 provides just a few examples of the potential ways in which microarray technology can be utilized.

1.3 AUTOANTIGEN AND CYTOKINE MICROARRAYS

Applications of protein microarrays include assessment of enzyme–substrate, protein–protein, and DNA–protein interactions. Although efforts to develop these proteomic tools predate the first descriptions of DNA microarrays [2], progress has been relatively slower — in part due to challenges posed by natural inherent differences in proteins compared with DNA. As examples, proteins consist of highly diverse conformational structures that result from 20 amino acids vs. the 4 nucleic acid building blocks that generate a relatively uniform structure in DNA. Proteins may exist as large complexes, can be hydrophilic or hydrophobic, acidic or basic, and contain posttranslational modifications such as acetylation, glycosylation, or phosphorylation. Functional and conformational properties of proteins must often remain intact when immobilized onto a microarray in order to retain the desired binding properties for detection of target ligands.

The development of protein microarrays to detect immunologic targets such as cytokines or autoantibodies has enormous potential for research and diagnostic applications in autoimmune diseases. Several groups, including Joos and colleagues in Germany [3], and Robinson and colleagues at Stanford University [4], have made important strides in developing autoantigen microarrays for multiplex characterization of autoimmune serum. Joos and colleagues spotted 18 common autoantigens onto silane-treated glass slides and nitrocellulose at serial dilutions. Bound antibodies from minimal amounts of 25 characterized autoimmune serum samples and ten normal blood donors were titered by using variable amounts of autoantigen. The autoimmune serum samples were obtained from patients with autoimmune thyroiditis (Hashimoto's thyroiditis and Graves' disease), systemic lupus erythematosus (SLE), Sjogren's syndrome (SS), mixed connective tissue disease (MCTD), scleroderma, polymyositis, systemic vasculitis, and antiphospholipid syndrome. These assays proved to be highly specific and similar in sensitivity when compared to a standard ELISA format. Further developments will include optimizing the nature of the autoantigen material to minimize possible loss of antigenicity and expanding the representation of autoantigens on the array.

Similarly, Robinson and colleagues have developed a 1152-feature array containing 196 distinct biomolecules representing major autoantigens targeted by antibodies produced by rheumatic autoimmune disease patients [4]. The autoantigens included hundreds of proteins, peptides, DNA, enzymatic complexes, and ribonucleoprotein complexes. Examples of autoantigens spotted include Ro52, Ro60, La, jo-1, Sm-B/B′, U1-70 kD, U1 snRNP-C, topoisomerase 1, pyruvate dehydrogenase (PDH), and histone H2A. The arrays were characterized using multiple sera from eight human autoimmune diseases and included SLE, SS, MCTD, polymyositis, primary biliary cirrhosis, rheumatoid arthritis (RA), and both limited and diffuse forms of scleroderma. This work demonstrates the feasibility of using large-scale, fluorescence-based autoantigen microarrays to detect human autoantibodies with simple protocols and widely available equipment in a low-cost and low-sample volume format. Some of the potential applications for this technology include (1) rapid screening for autoantibody specificities to facilitate diagnosis and treatment, (2) characterization of the specificity, diversity, and epitope spreading of autoantibody responses, (3) determination of isotype subclass of specific autoantibodies, (4) guiding development and selection of antigen-specific therapies, and (5) use as a discovery tool to identify novel autoantigens or epitopes.

Microarrays that simultaneously detect multiple cytokines have been developed by Huang and colleagues at Emory University [5]. Their method utilizes capture antibodies spotted onto membranes, incubation with biological samples such as patient serum, and detection by biotin-conjugated antibodies and enzymatic-coupled enhanced chemiluminescence. Twenty-eight cytokines were detected using this method, including interleukins-1$\alpha, 2, 3, 5, 6, 7, 8, 10, 13,$ and 15; tumor necrosis factors α, and β; interferon-γ, and others. In addition to detecting multiple cytokines simultaneously, these assays were shown to be more sensitive than conventional ELISAs, with broader detection ranges. The ability to readily scale up this approach to include much larger numbers of cytokines and other proteins will undoubtedly fuel further development of this powerful tool for studying complex and dynamic cellular processes such as immune reactions, apoptosis, cell proliferation, and differentiation.

1.4 DNA AND OLIGONUCLEOTIDE MICROARRAYS

DNA microarrays were first introduced in the mid-1990s [6] and have been the most widely utilized application of microarray technology. There are two commonly available DNA microarray systems. First are the cDNA microarrays fabricated by robotic spotting of PCR products, derived primarily from the 3′ end of genes and expressed sequence tags (ESTs), onto glass slides — this is the method popularized by, among others, Dr. Patrick Brown at Stanford and Dr. Louis Staudt at the NIH [7,8]. The second method uses *in situ* synthesized oligonucleotide arrays that are fabricated using photolithographic chemistry on silicon chips — this is the method used in the proprietary Affymetrix$^{\text{TM}}$ system [9] and recently by NimbleGen$^{\text{TM}}$. A third method involves spotting previously synthesized longer (40 to 70mer) oligonucleotides on either glass (Amersham$^{\text{TM}}$ and Agilent$^{\text{TM}}$) or nylon and plastic (clonetech$^{\text{TM}}$ and SuperArray$^{\text{TM}}$). The data generated using these systems are highly concordant, as demonstrated in parallel studies of the yeast cell cycle [10,11]. In the spotted cDNA and long oligo microarray systems, two probes with different fluorescent tags are hybridized to the same array, one serving as the experimental condition and the other as a control. The ratio of hybridization between the two probes is calculated, allowing a quantization of the hybridization signal for each spot on the array. In this system, the probe is 1st strand cDNA generated by oligo-dT primed reverse transcription from an RNA sample (for additional details see http://cmgm.stanford.edu/pbrown/). In the Affymetrix$^{\text{TM}}$ system, only a single labeled probe is used and each gene on the chip is represented by 8 to 10 wild-type 25-mer oligonucletides and the same number of single base mutant 25-mer oligonucleotides synthesized next to one another on the array. Signal intensity and the ratio of specific to nonspecific hybridization allows the generation of quantitative data regarding gene expression in the sample (for more details see http://www.affymetrix.com/technology/tech_probe.html).

1.5 TILING ARRAYS

Recently several groups have developed arrays with long stretches of chromosomes or whole-genomic sequences probed onto arrays. Potential uses for such whole-genome arrays include empirical annotation of the transcriptome [12], identification of novel transcripts [13,14], analysis of alternative and cryptic splicing, characterization of the methylation state of the genome, polymorphism discovery and genotyping, comparative genome hybridization, and genome resequencing [15]. These arrays have great future potential for studying new aspects of the genome and providing greater insights into the function of living organisms.

1.6 DATA ANALYSIS

Microarray analysis is often considered a discovery-based rather than hypothesis-driven approach [16,17], largely due to the potential for discovering altered expression of novel genes for which little or no prior information was available to suggest a role in the disease or experimental condition examined. However, high quality experiments are driven by addressing a scientific question (even if it is simply — "are there genes that are differentially expressed between a group of patients and controls?"), consistency in execution of experimental protocols, use of sample sizes with as many replicates as is feasible, and a plan for statistical analysis and interpretation of the data. Including statistical expertise during the early phase of experimental design (i.e., prior to any data collection) is critical, particularly in the setting of microarray analysis where each experiment can carry significant cost.

1.7 FUTURE DIRECTIONS

The majority of human diseases undoubtedly involves the complex interplay of many genes. Although the number and type of genes are not yet known, global assessment of gene expression is a very powerful approach for gaining insight into these processes. Identification of these genes will certainly contribute to advancing our understanding of the molecular basis for human diseases and identifying novel therapeutic targets. Within a relatively short period of time, the information learned from the application of microarray technology to address complicated biological questions has not only met, but often exceeded expectations. Despite their success, microarray studies are not without their challenges. Continued refinement of these techniques, including development of improved statistical methods for extracting information from large datasets and software tools for data processing, management, and storage as described in the following chapters of this book, will likely increase the applicability and general use of these technologies. Additionally, establishing common standards for the publishing and sharing of microarray generated data will be important. The applicability of this technology in translational medicine is only beginning to be appreciated and it is likely that microarray technologies will have a substantial impact on our understanding of human disease now and into the future.

REFERENCES

1. J.M. Trent and A.D. Baxevanis. Chipping away at genomic medicine. *Nat. Genet.* (Suppl): 462, 2002.
2. G. MacBeath. Protein microarrays and proteomics. *Nat. Genet.* 32 (Suppl): 526–532, 2002.
3. T.O. Joos, M. Schrenk, P. Hopfl, K. Kroger, U. Chowdhury, D. Stoll, D. Schorner, M. Durr, K. Herick, S. Rupp, K. Sohn, and H. Hammerle. A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics. *Electrophoresis* 21: 2641–2650, 2000.
4. W.H. Robinson, C. DiGennaro, W. Hueber, B.B. Haab, M. Kamachi, E.J. Dean, S. Fournel, D. Fong, M.C. Genovese, H.E. de Vegvar, K. Skriner, D.L. Hirschberg, R.I. Morris, S. Muller, G.J. Pruijn, W.J. van Venrooij, J.S. Smolen, P.O. Brown, L. Steinman, and P.J. Utz. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat. Med.* 8: 295–301, 2002.
5. R.P. Huang. Simultaneous detection of multiple proteins with an array-based enzyme-linked immunosorbent assay (ELISA) and enhanced chemiluminescence (ECL). *Clin. Chem. Lab. Med.* 39: 209–214, 2001.
6. M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–470, 1995.
7. A. Alizadeh, M. Eisen, D. Botstein, P.O. Brown, and L.M. Staudt. Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol.* 18: 373–379, 1998.
8. J.L. DeRisi, V.R. Iyer, and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680–686, 1997.
9. A.C. Pease, D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S.P. Fodor. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl Acad. Sci., USA* 91: 5022–5026, 1994.
10. R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 2: 65–73, 1998.
11. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9: 3273–3297, 1998.
12. E.E. Schadt, S.W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K.W. Hart, A. Russell, G. Li, G. Cavet, J. Castle, P. McDonagh, Z. Kan, R. Chen, A. Kasarskis, M. Margarint, R.M. Caceres, J.M. Johnson, C.D. Armour, P.W. Garrett-Engele, N.F. Tsinoremas, and D.D. Shoemaker. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* 5: R73, 2004.
13. V. Stolc, M.P. Samanta, W. Tongprasit, H. Sethi, S. Liang, D.C. Nelson, A. Hegeman, C. Nelson, D. Rancour, S. Bednarek, E.L. Ulrich, Q. Zhao, R.L. Wrobel, C.S. Newman, B.G. Fox, G.N. Phillips Jr., J.L. Markley, and M.R. Sussman. *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci., USA* 102: 4453–4458, 2005.
14. P. Bertone, V. Stolc, T.E. Royce, J.S. Rozowsky, A.E. Urban, X. Zhu, J.L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global

identification of human transcribed sequences with genome tiling arrays. *Science* 306: 2242–2246, 2004.

15. T.C. Mockler and J.R. Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85: 1–15, 2005.
16. L.M. Staudt and P.O. Brown. Genomic views of the immune system. *Annu. Rev. Immunol.* 18: 829–859, 2000.
17. S.M. Albelda and D. Sheppard. Functional genomics and expression profiling: be there or be square. *Am. J. Respir. Cell Mol. Biol.* 23: 265–269, 2000.

2 Normalization of Microarray Data

Rudolph S. Parrish and Robert R. Delongchamp

CONTENTS

2.1 OBJECTIVES OF NORMALIZATION

2.1.1 What Is Normalization?

Normalization of microarray data is any procedure meant to reduce or account for systematic variation among or within arrays. This variation is a component common to all the genes that are measured on an array or, more generally, to a subset of genes on the array. Normalization methods are often applied prior to the application of statistical analysis methods, which are usually designed to detect differential expression. However, normalization includes procedures that adjust for known effects as part of the statistical analysis as well as those that replace the actual data with modified values prior to the statistical analysis. In either case, normalization represents an effort to obtain more powerful tests by reducing variation in the data or otherwise accounting for it mathematically.

2.1.2 Sources of Variation

The raw data from gene microarrays involve variation due to several sources [1,2]. The intent of a typical experiment is to determine whether treatment groups of experimental units (e.g., subjects, patients, mice, etc.) exhibit differential gene expression patterns. Such comparisons are based on an assessment of the variation among the experimental units within groups, which is the experimental error variance. In addition to variation in mRNA levels from unit to unit, there is variation arising from the measurement process. Normalization methods attempt to remove or reduce the influence of these additional sources of variation.

Some writers distinguish between “biological” and “technical” variation. Variation that is inherent to the characteristics of the experimental units is considered as biological variation. Variation that derives from the characteristics of the arrays themselves (due to manufacturing issues), the processing of the samples applied to arrays (e.g., sample preparation, mRNA extraction, labeling), hybridization of sample material onto the arrays, and measurement of intensities (e.g., optical properties, label intensity, scanner settings) all are considered as technical variation; see also [3]. If all technical variation could be eliminated, there would be some sense of purity in the data that should reflect group effects and involve only natural unit-to-unit variation. An ideal normalization method would remove all effects of technical variation.

Although some procedures may result in transformed data that are approximately normally distributed (i.e., Gaussian distribution), achieving normality is not the primary objective of normalization.

2.1.3 Background Correction

Many proposed methods incorporate the use of a background correction procedure in which measured intensities are adjusted according to some level of background noise. Although these methods also result in a modification to the data, in this chapter background correction algorithms are regarded as attempts to reduce bias, whereas normalization methods are regarded as attempts to reduce variance due to technical sources. Background corrections mainly affect the low expressions. Thus, background correction will not be considered in this treatment of normalization.

2.1.4 Platforms

Normalization methods generally apply to any platform, although some methods obviously are designed for one- or two-color systems or are specific to a platform. These platforms include high-density oligonucleotide arrays [4], cDNA spotted arrays using two labels per spot, and spotted arrays using one label per spot [5].

2.2 STATISTICAL BASIS OF NORMALIZATION

2.2.1 Microarray Data

Microarray data from an individual array basically form just a high-dimension multivariate observation of gene expressions. The array corresponds to an experimental unit or a sampling unit within the experimental unit, and genes correspond to the variables measured on the unit. In two-color systems, the red and green dyes often correspond to paired specimens from the same or different experimental units. As part of multivariate observations, it is natural to assume that correlations exist among the variables. Obviously, genes may be correlated through biological relationships. Variables may be correlated also by virtue of being associated with the same spots on two-color arrays.

Typically, there is only a single array for each experimental unit. That is, there usually is no replication of multiple arrays per experimental unit. In classical experimental designs, such replication forms the basis for obtaining purer estimates of the experimental error variance, and this principle can be applied to microarray experiments [6,7]. With microarrays, a different technique is employed that is based on assumed relationships involving hundreds or thousands of genes on each array. Basically, normalization methods are developed under the assumption that the average gene does not change significantly among the experimental units even under the various experimental conditions.

2.2.2 Transformations

Nearly all investigators employ a logarithmic (usually base 2) transformation on expression values prior to analysis or normalization. For 16-bit images, this means that $\log_2$-transformed expression values will be real values between 0 and 16. The purpose of transforming the data is mostly to address potential multiplicative error structures that give rise to instability of variances, but also to achieve normality so that subsequent statistical inferences will be valid. However, considering all choices, the logarithmic transformation may not be the one that most nearly produces a normal distribution, and the most appropriate choice of transformation is likely to be different for various genes when seeking normality [8,9]. Nonetheless, it is usually assumed that a logarithmic transformation stabilizes the variance, at least approximately [10].

2.2.3 Analysis of Variance Models

It is useful to consider statistical models for microarray data in order to characterize variability mathematically and to assess normalization methods. Several models, as described below, have been proposed which provide a framework for understanding the variances that are present in microarray expression data. Among these are analysis of variance type models with additive errors as given by several authors [3,11–16]. Models involving multiplicative errors have also been introduced [17,18]. Most utilize a logarithmic transformation of the expression data. Normalization methods that involve modeling probe intensity levels have also been proposed [19].

Various normalization methods make use of presumed relationships with other genes (or their overall characteristics) to modify the data so that the among-arrays variance is reduced. In a linear model context, the data are not modified directly but rather other effects in the model are adjusted for in order to reduce estimates of standard errors of treatment differences.

Most papers consider normalization as an adjustment that precedes the analysis for treatment effects. Examples include the mean (median) subtraction or locally-weighted regression (loess) adjustments. In these cases, the data are normalized and then the normalized values are analyzed for treatment effects. However, normalization can be directly incorporated into the analysis, as with the analysis of variance models. The analysis of variance approach is attractive because it explicitly accounts for sources of variation that impact inferences about treatments including the "array effect," which invariably is a major source of variation.

2.2.4 Variance Components

A simple and typical experimental design for single-channel arrays involves two treatment groups ($k = 2$), multiple subjects per treatment (n subjects), and a single array per subject ($r = 1$). There are two components of random variation: one associated with variation among subjects treated alike (subjects within treatment groups) and the other associated with arrays within subjects. The first of these is the experimental error variance, denoted by σ_s^2. The second is an array-specific variance, denoted by σ_a^2. The analysis of variance involves three sources of variation in the indicated expected mean squares. Because the number of degrees of freedom for arrays is zero when $r = 1$, the corresponding variance component is not estimable. In this model, subjects and arrays are considered as random effects.

Source	**DF**	**Expected mean square**
Treatment (Trt)	$k - 1$	$Q_{\text{Trt}} + r\sigma_s^2 + \sigma_a^2$
Subjects (Trt)	$k(n - 1)$	$r\sigma_s^2 + \sigma_a^2$
Arrays (Subjects Trt)	$kn(r - 1)$	σ_a^2

Normalization may be thought of as an attempt to reduce the magnitude of the "among-arrays" variance component σ_a^2. The "among-subjects" component σ_s^2 is the experimental error component and should not be modified by normalization. Q_{Trt} is a quadratic form based on treatment means.

2.2.5 Significance Testing

In view of the ultimate objectives in microarray analysis, significance testing is conducted in one form or another in order to discover genes that exhibit differential expression. As such, normalization methods should seek to improve power of such tests and to reduce false discovery rates. Thus, in a real sense, the impact on significance testing is one of the most important characteristics of a normalization procedure. Methods that replace the data with adjusted values, by definition, attempt to reduce variability and are likely to produce a higher frequency of significant results. Consideration of the F test, which is formed by the ratio of the mean square among treatments, $\text{MS}_{(\text{Trt})}$, to the mean square among subjects within treatments, $\text{MS}_{\text{Subjects(Trt)}}$, gives rise to the following ratios of variances estimated by these mean squares in the presence or absence of variance associated with arrays

$$\frac{E(\text{MS}_{\text{Trt}})}{E(\text{MS}_{\text{Subjects(Trt)}})} = \frac{Q_{\text{Trt}} + r\sigma_s^2 + \sigma_a^2}{r\sigma_s^2 + \sigma_a^2} < \frac{Q_{\text{Trt}} + r\sigma_s^2}{r\sigma_s^2}$$

Thus, a normalization procedure that is effective in reducing σ_a^2 will increase the F statistics (or, equivalently, t statistics for the case of two treatments), and therefore reduce p values, assuming σ_s^2 remains constant. A problem arises if the normalization method reduces variability associated not only with the array effects but also that associated with the experimental error (i.e., among-subjects variance). This issue has been discussed for two normalization methods applied to prostate cancer data [20].

2.2.6 Bias and Variance Reduction

Normalization procedures attempt to reduce variance among arrays, which improves the resolution of treatment differences. However, normalization also may bias the estimated treatment effects, which impairs the resolution of treatment differences. The bias and variance can be examined in a simple case, which allows a heuristic evaluation of the properties to be expected for less tractable cases. Let y_{ga} denote the logarithm of the observed intensity for gene "g" on array "a." A simple normalization method is to subtract the global mean for each array. This formula can be written in matrix form as $(\mathbf{I} - \mathbf{J}/m)\mathbf{y}_a$ where $\mathbf{I}$ denotes an $(m \times m)$ identity matrix, $\mathbf{J}$ denotes an $(m \times m)$ matrix of 1s, and $\mathbf{y}_a = (y_{1a}, y_{2a}, \ldots, y_{ma})'$. The gth element of $(\mathbf{I} - \mathbf{J}/m)\mathbf{y}_a$ is simply $y_{ga} - \bar{y}_{\cdot a}$. If μ_a and Σ denote the mean and covariance of y_{ga}, then the mean and covariance of the normalized data are, respectively, $(\mathbf{I} - \mathbf{J}/m)\mu_a$ and $(\mathbf{I} - \mathbf{J}/m)\Sigma(\mathbf{I} - \mathbf{J}/m)'$.

The underlying logic for normalization is an assumption that the m genes measured on an array share a component of variation, the array effect. The aim of normalization is to eliminate this effect. To see this, denote the variance of the array effect by σ_a^2, then the variance Σ can be partitioned into two parts such that $\Sigma = \mathbf{D} + \mathbf{J}\sigma_a^2$, where $\mathbf{D}$ is a diagonal matrix with elements equal to the gene-specific variances (over subjects). Then, the variance of the normalized values is $(\mathbf{I} - \mathbf{J}/m)\mathbf{D}(\mathbf{I} - \mathbf{J}/m)'$, which does not depend on σ_a^2. This matrix has diagonal elements equal to $\sigma_g^2 - (2/m)\sigma_g^2 + (1/m^2)\sum_{g=1}^{m}\sigma_g^2$, representing variances of the normalized variables. The covariances are given by $-(1/m)\sigma_i^2 - (1/m)\sigma_j^2 + (1/m^2)\sum_{g=1}^{m}\sigma_g^2$. In effect, sources of variation shared by all intensities on the array are eliminated by this normalization. Because the array variance can be quite large relative to the variance among subjects for specific genes, the data may have substantially lower variance after normalization. For large m, the variance of the normalized data will be approximately $\mathbf{D}$.

Global normalization of this form also introduces bias. Let Δ denote the logarithm of the true fold changes between two arrays, that is, the expectation of $\mathbf{y}_a - \mathbf{y}_b$. Then the expected difference after subtracting the respective means is $(\mathbf{I} - \mathbf{J}/m)\Delta \neq \Delta$. In general, estimated fold changes, which are based on normalized data, are biased in the opposite direction of the average logarithm of the fold change. If most of the interrogated genes are unaffected by treatment and m is large, this bias is negligible and the accompanying reduction in variance far outweighs any detriment from bias. However, the potential to seriously corrupt inferences about treatment effects through normalization should be a concern whenever large numbers of genes exhibit differential expression and/or whenever the data are highly smoothed as this conceptually corresponds to normalizing within subsets of the interrogated genes (i.e., effectively small m).

There are two generic directions in which the global means normalization can be modified. These modifications encompass a large percentage of the normalization procedures that have been proposed. One direction is to apply the mean normalization within subsets of the interrogated genes. For example, genes can be placed into subsets based on their physical location or the magnitude of their intensity. Procedures that adjust for each print pin or those that regress on the magnitude can be viewed this way. The other direction of modification is to replace the mean with alternative estimators. The bias can be mitigated somewhat by the use of ‘outlier’ resistant methods (e.g., using the median rather than the mean). An outlier in the context of normalization is any intensity that is affected by factors in addition to the array effect. The model outlined above only accounts for treatment effects and array effects, so “outliers” in the context of this model are the genes affected by treatment. Finessing in both directions of modification leads to more complex procedures such as loess regression.

Two other procedures also can be interpreted in the context of means normalization. If one replaces the rows of $\mathbf{J}$ with an indicator of genes for which there are no treatment effects, then the variance is reduced and there is no bias. However, this requires a priori knowledge of a subset of interrogated genes which are not affected by treatment. This is the basis for normalizations based upon housekeeping genes, although the assertion that they are unaffected by treatment is problematic. Another approach is to use the normalizing value (global mean) as a covariate in an analysis of the observed log-intensities. In essence, $\mathbf{J}$ is replaced by $\mathbf{BJ}$ where $\mathbf{B}$ is a diagonal matrix of gene-specific coefficients estimated by the analysis of covariance.

Conceptually, this accounts better for different correlations between the observed intensities and the normalizing value.

2.2.7 Variance within Arrays

Normalization by subtraction of the global means does not adjust for possible different array-specific variability among genes that might exist. In this situation, individual gene expressions will have variances across subjects (σ_g^2) that may be very large, even with respect to σ_a^2. By considering features of the array-specific distributions of gene expressions or probe intensities (e.g., the interquartile range or particular quantiles), additional adjustments may be in order. The method based on the median and interquartile range is an example of a combined approach.

2.2.8 Analysis of Covariance Models

It is not straightforward to test for treatment effects in analysis of variance (ANOVA) models unless one makes the assumptions that all genes have the same residual variance and that all genes are independent, which are unlikely to be satisfied in practice. From a strict statistical perspective, a better approach is to analyze the intensity data by individual genes and to extract the "normalizing" information residing in the other interrogated genes by enlisting a summary statistic as a covariate. For example, instead of subtracting the median from each log-intensity and then analyzing for treatment effects, one could analyze the log-intensities for treatment effects incorporating the median as a covariate. Conceptually, such an analysis makes a better adjustment because it estimates the attenuation associated with measurement error that is implicit in using the median (or other summary estimate) as a surrogate measure of the array effect.

To elaborate, consider a simplified hypothetical setting where the analysis involves two genes per array with several arrays receiving a treatment and several arrays serving as controls. We make the additional assumption that the second gene is unaffected by treatment (essentially, the basis for normalizing by housekeeping genes). Then $\mathbf{y}_a$ as previously defined can be written

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}_a \sim \left[\begin{pmatrix} \mu_1 + \tau \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 + \sigma_a^2 & \sigma_{12} + \sigma_a^2 \\ \sigma_{12} + \sigma_a^2 & \sigma_2^2 + \sigma_a^2 \end{pmatrix}\right]$$

where $\tau = 0$ if the array is from a control sample. In the analysis of covariance, we are interested in testing the first gene for a treatment effect using the distribution of Y_1, given an observed value of Y_2. In general, the expectation of this distribution depends upon the distributional assumptions in addition to the means and variances. For a bivariate normal distribution, it is known that $E(Y_1|y_2) = \mu_1 + \tau + \beta(y_2 - \mu_2)$ where

$$\beta = \frac{\sigma_{12} + \sigma_a^2}{\sigma_2^2 + \sigma_a^2}$$

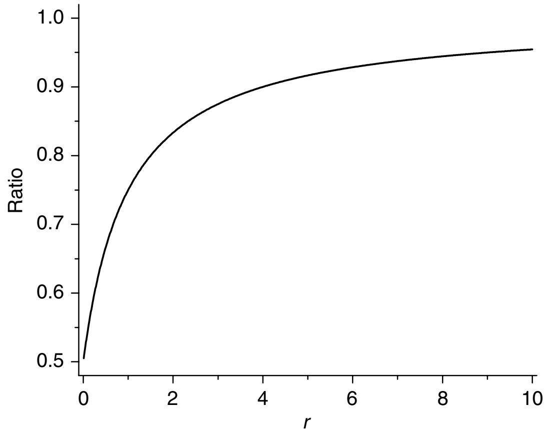


FIGURE 2.1 Ratio of the variances of treatment effect: variance using analysis of covariance divided by variance using the difference.

In particular, the variance is $\text{var}(Y_1|Y_2) = \sigma_1^2 + \sigma_a^2 - \beta(\sigma_{12} + \sigma_a^2)$ giving the variance of the estimated treatment effect, $\text{var}(\hat{\tau}) = (2/n)(\sigma_1^2 + \sigma_a^2 - \beta(\sigma_{12} + \sigma_a^2))$, assuming there are n treated and n control arrays. The variance of the treatment effect estimated from the differences, $y_1 - y_2$, is $(2/n)(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$. Arguably, the better of these methods is the one that has the smallest variance, which is straightforward to evaluate if all the component variances are specified. As an example, suppose

$$\boldsymbol{\sigma} = \sigma^2 \begin{pmatrix} 1+r & r \\ r & 1+r \end{pmatrix}$$

This represents cases where $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, $\sigma_{12} = 0$, and $\sigma_a^2 = r\sigma^2$, that is, r is the relative magnitude of the variance of the array effect. In this example, the ratio of these variances only depends on r. In Figure 2.1, the ratio is formed with $\text{var}(\hat{\tau})$ in the numerator and plotted as a function of r. Since the ratio is less than 1, the analysis of covariance produces more precise estimates of the treatment effect. At least under the assumed $\boldsymbol{\sigma}$, an analysis of covariance would be preferable to analyzing the differences. This preference would likely apply whenever normalization is based upon the mean of a few “housekeeping” genes.

In practice, when the median of all the interrogated genes is employed as a covariate, the mathematics becomes intractable. However, the median should behave similarly as the mean where the ratio is essentially 1, so there is little advantage in regard to the precision of estimated effects. Simulations using the median confirm that there is little if any increase in the precision of estimated treatment effects when a covariance analysis is compared to an analysis based upon the differences. Like the difference, the analysis of covariance also suffers from bias, which justifies use of the median rather than the mean. Hence, our preference for the analysis of covariance is largely aesthetic. An analysis of covariance seems better because it renders an assessment of the variation explained by normalization.

2.3 NORMALIZATION ALGORITHMS

A large number of normalization methods have been proposed, several of which are identified in Table 2.1; this list is not exhaustive. Bolstad et al. [19] distinguished between "complete data" methods, in which data from all arrays are used to normalize, and methods that use a baseline array to establish the normalization relation. The global normalization methods generally can be applied to all arrays together or separately to arrays within treatment groups. Nonetheless, there is no consensus on classification of these methods.

In the following sections, some methods that have been frequently reported are described in detail; however, this does not imply that these methods are necessarily more appropriate or more effective than others. Comparisons among normalization methods have not established which ones have the most desirable statistical performance characteristics.

2.3.1 Reference Genes

2.3.1.1 Housekeeping Genes

A set of housekeeping genes may be used as a group of reference genes for adjusting array values [21,22], provided it can be assumed that true expression values for these genes are unaffected by experimental conditions and do not vary across subjects or samples. Such genes are selected in advance of the experiment. Methods for selection

TABLE 2.1
Selected Normalization Methods

Reference genes	
Housekeeping and control genes	[21,44]
Global and local methods	
Global mean or median	
Linear scaling	[4]
Nonlinear scaling	[25]
Invariant set	[17]
Consistent set	[45]
Median-Interquartile range	[28]
Signal dependent q-spline	[46]
Variance stabilization	[8,10,47]
Quantile	[19,32,33]
Local regression	[29,30,36,48]
Variance regularization	[49]
Spatial normalization	[46]
Linear models	
Analysis of variance	[11]
Mixed-effects models	[12,35]
Split-plot design	[13]
Subset/global intensity	[14]

of control genes have been described [23]. Some arrays (e.g., HG-U133) are designed with probe sets that are not expected to vary across different tissue types. Typically, an adjustment factor would be computed for each array that makes the means of the reference genes all equal. Some investigators have reported difficulty in selecting suitable housekeeping genes [24,25].

2.3.2 Global and Local Methods

2.3.2.1 Linear Scaling to a Common Mean and Range

The Affymetrix algorithm [4] involves simple scaling according to the expression

$$y_{ij}^{*} = y_{ij} \times \left(y_{\text{baseline}}^{(m)} / y_{i}^{(m)} \right)$$

where $y_{\text{baseline}}^{(m)}$ is the (trimmed) mean of the expression values on the reference array, and similarly $y_{i}^{(m)}$ is that corresponding to the ith array. This may be applied at the probe intensity level. Under a linearity assumption, this is effectively fitting a line through the origin for array i values vs. baseline array values, paired according to individual genes or probes. This produces for array i the same mean and same range of variation as for the baseline array. This method does not correct for situations where the low-level expressions or intensities have a slope different from that for the larger values (i.e., if the expression distributions are very different).

2.3.2.2 Nonlinear Scaling

Instead of assuming linearity between each array and the baseline array, one can employ nonlinear relationships. Schadt et al. [25] described use of a nonlinear normalizing relation based on a subset of genes considered to be invariant relative to the ordering of expressions on an array compared to the baseline array. They proposed an algorithm based on ranks that finds an approximately invariant set of genes, which then are used with the generalized cross-validation smoothing spline algorithm (GCVSS) given by Wahba [26]. This algorithm has been implemented in dChip software [25]. This data transformation may be represented generally as

$$y_{ij}^{*} = f_i(y_{ij})$$

where f_i represents the nonlinear scaling function.

Li and Wong [17] employed a piecewise running median line instead of the GCVSS approach, and Bolstad et al. [19] utilized a loess smoothing approach [27] on probe intensities.

2.3.2.3 Overall Mean or Median

An additive adjustment factor can be computed for each array in order to make the array means or medians all equal to one another. That is, it makes the total intensity for all arrays equal or nearly so. This can be implemented by choosing a reference (i.e., baseline) array arbitrarily and adjusting the other arrays to that total intensity value, or it can be implemented by computing the mean of the means or median of the medians as the target value. This type of normalization might be valid if log expression values are affected as a direct result of differing quantities of mRNA. Mathematically, this is represented as

$$y_{ij}^{*} = y_{ij} + (y.^{(m)} - y_i^{(m)})$$

where y_{ij} is the prenormalized log-expression value for array i and gene j, $y_i^{(m)}$ is the mean or median value for array i based on values of all genes on the array or it may be the value from the baseline array, $y.^{(m)}$ is the mean of means or median of medians over all arrays, and y_{ij}^{*} is the normalized log-expression value. An arithmetic mean or a trimmed mean can be used. In a distribution sense, this method equates the central tendencies for all arrays. A global mean or median adjustment can be applied across all arrays together or within treatment groups separately.

2.3.2.4 Scaling for Heterogeneity of Variance

The global mean or median adjustment provides shift corrections but does not alter the variance of gene expression values within arrays. A simple adjustment for differing variances can be accomplished by selecting a scaling constant for each array. The scaling constants can be based on a measure of dispersion using either a baseline array or the mean or median of that measure over all arrays. Commonly, the interquartile range (IQR) is used because it is not affected by outliers as is the case with the standard deviation [28]. Incorporation of both global mean or median adjustment and an IQR-based dispersion adjustment is given mathematically by

$$y_{ij}^{*} = (y_{ij} + y.^{(m)} - y_i^{(m)}) \times (D./D_i)$$

where D_i is the measure of dispersion for the ith array and $D.$ is the mean or median or maximum of the D_i values over all arrays. The IQR-based method can be generalized for use of any two other quantiles, such as the normal range based on the 2.5th and 97.5th percentiles. When the IQR criterion is used, all arrays will have the same IQR after normalization.

2.3.2.5 Loess on Two-Channel or Paired Arrays

Dudoit et al. [29] described a loess-based method based on M vs. A (MvA) plots. This method plots

$$M_j = \log_2(y_{ij}^{(1)}/y_{ij}^{(2)}) \quad \text{vs.} \quad A_j = 0.5 \log_2(y_{ij}^{(1)} \times y_{ij}^{(2)})$$

or, equivalently,

$$M_j = \log_2(y_{ij}^{(1)}) - \log_2(y_{ij}^{(2)}) \quad \text{vs.} \quad A_j = [\log_2(y_{ij}^{(1)}) + \log_2(y_{ij}^{(2)})]/2$$

for each array i over all genes (or probes) j, and then fits a loess smoothing function. Here, $y_{ij}^{(1)}$ represents the expression (or probe) values for one channel and $y_{ij}^{(2)}$ similarly for the other channel.

Normalized values are a function of the deviations, denoted by M_j', from the fitted regression line; particularly,

$$\log_2(y_{ij}^{(1)*}) = A_j + 0.5M_j' \quad \text{and} \quad \log_2(y_{ij}^{(2)*}) = A_j - 0.5M_j'$$

where $y_{ij}^{(1)*}$ and $y_{ij}^{(2)*}$ represent the normalized values for the two channels.

2.3.2.6 Cyclic Loess on Single-Channel Arrays

The method of Dudoit et al. [29] was adapted by Bolstad et al. [19] for application to single-channel arrays by considering all pairs of arrays when constructing MvA plots. Their algorithm iteratively finds adjustments that ultimately result in normalized values.

2.3.2.7 Loess on an Orthonormal Basis

A version of the MvA method was introduced by Astrand [30] in which he first transforms the log probe intensity vector for each array using an orthonormal contrast matrix with dimensions equal to the number of probes, in order to create an alternative basis of the data. This is followed by applying a loess method to the MvA plots where a fixed reference vector in the alternative basis is paired with each of the other arrays in the alternative basis. This algorithm is implemented in the *R* software *maffy* [2,31].

2.3.2.8 Quantile Normalization

Bolstad and coworkers [19,32–34] introduced a method based on quantiles of the underlying distribution of probe intensities. That method creates identical distributions of probe intensities for all arrays by replacing the intensity values in order to attain a straight-line relationship on quantile–quantile plots for any two arrays. This is accomplished by projecting the points of an n-dimensional quantile plot onto a unit diagonal vector, according to the following steps (a) Form a data matrix $\mathbf{X}$ of dimension $p \times n$ where p is the number of probes and n is the number of arrays; (b) Sort each column of values from low to high to produce order statistics for each column; (c) Replace all values in each row of the sorted matrix by the mean of that row's values (i.e., the mean of the ith order statistics from all columns); (d) Rearrange the elements of each column back to the original ordering. The resulting matrix is the normalized data matrix from which expression values then are calculated.

2.3.3 Linear Model Based Methods

In the following statistical models, the response variable is generally taken to be the $\log_2$ of expression.

2.3.3.1 ANOVA-Based Model

Kerr et al. [11] introduced the use of ANOVA models that accounted for array, dye, and treatment effects for cDNA arrays. In this fashion, normalization was accomplished intrinsically without preliminary data manipulation. The model they proposed may be written as

$$y_{ijkg} = \mu + A_i + T_j + D_k + G_g + AG_{ig} + TG_{jg} + e_{ijkg}$$

where μ is the mean expression, A_i is the effect of the ith array, T_j is the effect of the jth treatment, D_k is the effect of the kth dye, G_g is the gth gene effect, and AG_{ig} and TG_{jg} represent interaction effects. Of interest for testing differential expression are the interaction effects, TG_{jg}, for which appropriate contrasts can be estimated for each gene. In this model, all effects were considered as fixed effects. Other terms could be incorporated into this model.

2.3.3.2 Mixed-Effects Model

Wolfinger et al. [12] utilized a mixed-effects model for cDNA data where the array effect and related interaction terms are considered as random effects. A model involving all effects simultaneously can be written as

$$y_{ijg} = \mu + A_i + T_j + AT_{ij} + G_g + AG_{ig} + TG_{jg} + e_{ijg}$$

Like the ANOVA model, this also intrinsically adjusts for the effects of arrays without modifying the data directly, although they recommend first fitting the model

$$y_{ijg} = \mu + A_i + T_j + AT_{ij} + e_{g(ij)}$$

and calculating residuals, denoted by r_{ijg}. Then the residuals are used as the dependent variables in the gene-specific models given by

$$r_{ijg} = G_g + AG_{ig} + TG_{jg} + e_{ijg}$$

or, equivalently, for each gene

$$r_{ij}^{(g)} = \mu^{(g)} + A_i^{(g)} + T_j^{(g)} + e_{ij}^{(g)}$$

The residuals are the normalized values. The effect of interest for testing differential expression is the $T_j^{(g)}$ term.

For single-channel arrays, in which different arrays are used within treatments, this approach can be represented with the overall model

$$y_{ijg} = \mu + T_j + A(T)_{i(j)} + G_g + TG_{jg} + e_{gi(j)}$$