Validation of Toxicogenomic Technologies: A Workshop Summary

Committee on Validation of Toxicogenomic Technologies: A Focus on Chemical Classification Strategies, Committee on Emerging Issues and Data on Environmental Contaminants, National Research Council

ISBN: 0-309-66825-5, 98 pages, 6 x 9, (2007)

VALIDATION OF TOXICOGENOMIC TECHNOLOGIES

A Workshop Summary

Committee on Validation of Toxicogenomic
Technologies: A Focus on Chemical
Classification Strategies

Committee on Emerging Issues and Data on
Environmental Contaminants

Board on Environmental Studies and Toxicology

Board on Life Sciences

Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This project was supported by Contract No. G-NAG 9-1451 between the National Academy of Sciences and the National Aeronautics and Space Administration. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number-13: 978-0-309-10413-5
International Standard Book Number-10: 0-309-10413-0

Additional copies of this report are available from

The National Academies Press
500 Fifth Street, NW
Box 285
Washington, DC 20055

800-624-6242
202-334-3313 (in the Washington metropolitan area)
http://www.nap.edu

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON VALIDATION OF TOXICOGENOMIC TECHNOLOGIES: A FOCUS ON CHEMICAL CLASSIFICATION STRATEGIES

Members

JOHN QUACKENBUSH *(Co-Chair)*, Harvard School of Public Health, Boston, MA
KENNETH S. RAMOS *(Co-Chair)*, University of Louisville, KY
CYNTHIA A. AFSHARI, Amgen, Inc., Thousand Oaks, Louisville, CA
LINDA E. GREER, Natural Resources Defense Council, Washington, DC
CASIMIR A. KULIKOWSKI, Rutgers University, New Brunswick, NJ
GEORGE ORPHANIDES, Syngenta Central Toxicology Laboratory, Cheshire, UK
LAWRENCE M. SUNG, University of Maryland School of Law, Baltimore, MD
RUSSELL D. WOLFINGER, SAS Institute Inc., Cary, NC

Staff

KARL E. GUSTAVSON, Project Director
MARILEE K. SHELTON-DAVENPORT, Project Director
JENNIFER E. SAUNDERS, Associate Program Officer
RUTH E. CROSSGROVE, Senior Editor
MIRSADA KARALIC-LONCAREVIC, Research Associate
RADIAH A. ROSE, Senior Editorial Assistant
LUCY V. FUSCO, Senior Project Assistant

Sponsor

NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES

COMMITTEE ON EMERGING ISSUES AND DATA ON ENVIRONMENTAL CONTAMINANTS

Members

KENNETH S. RAMOS *(Chair)*, University of Louisville, Louisville, KY
PATRICIA A. BUFFLER, University of California, Berkeley
JAMES S. BUS, Dow Chemical Company, Midland, MI
GREGORY J. CARR, The Procter & Gamble Company, Cincinnati, OH
JOSEPH J. DEGEORGE, Merck Research Laboratories, West Point, PA
DAVID J. GALAS, Battelle Memorial Institute, Columbus, OH
LINDA E. GREER, Natural Resources Defense Council, Washington, DC
ROBERT J. GRIFFIN, Marquette University, Milwaukee, WI
AMY D. KYLE, University of California, Berkeley
PETER G. LORD, Johnson & Johnson, Raritan, NJ
WILLIAM B. MATTES, Critical Path Institute, Poolesville, MD
AUBREY MILUNSKY, Boston University School of Medicine, Boston, MA
GILBERT S. OMENN, University of Michigan Medical School, Ann Arbor
GEORGE ORPHANIDES, Syngenta Central Toxicology Laboratory, Cheshire, UK
FREDERICA P. PERERA, Columbia University, New York, NY
JOHN QUACKENBUSH, Harvard School of Public Health, Boston, MA
MARK A. ROTHSTEIN, University of Louisville School of Medicine, Louisville, KY
LEONA D. SAMSON, Massachusetts Institute of Technology, Cambridge
MARTHA S. SANDY, California Environmental Protection Agency, Oakland
TODD SHERER, Emory University, Atlanta, GA
PETER S. SPENCER, Oregon Health and Science University, Portland
LAWRENCE M. SUNG, University of Maryland, Baltimore
MAHLET G. TADESSE, University of Pennsylvania School of Medicine, Philadelphia
CHERYL L. WALKER, University of Texas, Smithville

Staff

KARL E. GUSTAVSON, Project Director
MARILEE K. SHELTON-DAVENPORT, Project Director
JENNIFER E. SAUNDERS, Associate Program Officer
RUTH E. CROSSGROVE, Senior Editor
RADIAH A. ROSE, Senior Editorial Assistant
LUCY V. FUSCO, Senior Project Assistant

BOARD ON ENVIRONMENTAL STUDIES AND TOXICOLOGY

Members

JONATHAN M. SAMET *(Chair)*, Johns Hopkins University, Baltimore, MD
RAMÓN ALVAREZ, Environmental Defense, Austin, TX
JOHN M. BALBUS, Environmental Defense, Washington, DC
DALLAS BURTRAW, Resources for the Future, Washington, DC
JAMES S. BUS, Dow Chemical Company, Midland, MI
COSTEL D. DENSON, University of Delaware, Newark
E. DONALD ELLIOTT, Willkie Farr & Gallagher LLP, Washington, DC
MARY R. ENGLISH, University of Tennessee, Knoxville
J. PAUL GILMAN, Oak Ridge Center for Advanced Studies, Oak Ridge, TN
SHERRI W. GOODMAN, Center for Naval Analyses, Alexandria, VA
JUDITH A. GRAHAM, American Chemistry Council, Arlington, VA
WILLIAM P. HORN, Birch, Horton, Bittner and Cherot, Washington, DC
JAMES H. JOHNSON, JR., Howard University, Washington, DC
WILLIAM M. LEWIS, JR., University of Colorado, Boulder
JUDITH L. MEYER, University of Georgia, Athens
DENNIS D. MURPHY, University of Nevada, Reno
PATRICK Y. O'BRIEN, ChevronTexaco Energy Technology Company, Richmond, CA
DOROTHY E. PATTON (retired), Chicago, IL
DANNY D. REIBLE, University of Texas, Austin
JOSEPH V. RODRICKS, ENVIRON International Corporation, Arlington, VA
ARMISTEAD G. RUSSELL, Georgia Institute of Technology, Atlanta
ROBERT F. SAWYER, University of California, Berkeley
LISA SPEER, Natural Resources Defense Council, New York, NY
KIMBERLY M. THOMPSON, Massachusetts Institute of Technology, Cambridge
MONICA G. TURNER, University of Wisconsin, Madison
MARK J. UTELL, University of Rochester Medical Center, Rochester, NY
CHRIS G. WHIPPLE, ENVIRON International Corporation, Emeryville, CA
LAUREN ZEISE, California Environmental Protection Agency, Oakland

Senior Staff

JAMES J. REISA, Director
DAVID J. POLICANSKY, Scholar
RAYMOND A. WASSEL, Senior Program Officer for Environmental Sciences and Engineering
KULBIR BAKSHI, Senior Program Officer for Toxicology
EILEEN N. ABT, Senior Program Officer for Risk Analysis
KARL E. GUSTAVSON, Senior Program Officer
K. JOHN HOLMES, Senior Program Officer
ELLEN K. MANTUS, Senior Program Officer
SUSAN N.J. MARTEL, Senior Program Officer
STEVEN K. GIBB, Program Officer for Strategic Communications
RUTH E. CROSSGROVE, Senior Editor

BOARD ON LIFE SCIENCES

Members

KEITH YAMAMOTO *(Chair)*, University of California, San Francisco
ANN M. ARVIN, Stanford University School of Medicine, Stanford, CA
JEFFREY L. BENNETZEN, University of Georgia, Athens
RUTH BERKELMAN, Emory University, Atlanta, GA
DEBORAH BLUM, University of Wisconsin, Madison
R. ALTA CHARO, University of Wisconsin, Madison
JEFFREY L. DANGL, University of North Carolina, Chapel Hill
PAUL R. EHRLICH, Stanford University, Stanford, CA
MARK D. FITZSIMMONS, John D. and Catherine T. MacArthur Foundation, Chicago, IL
JO HANDELSMAN, University of Wisconsin, Madison
ED HARLOW, Harvard Medical School, Boston, MA
KENNETH H. KELLER, University of Minnesota, Minneapolis
RANDALL MURCH, Virginia Polytechnic Institute and State University, Alexandria
GREGORY A. PETSKO, Brandeis University, Waltham, MA
MURIEL E. POSTON, Skidmore College, Saratoga Springs, NY
JAMES REICHMAN, University of California, Santa Barbara
MARC T. TESSIER-LAVIGNE, Genentech, Inc., San Francisco, CA
JAMES TIEDJE, Michigan State University, East Lansing
TERRY L. YATES, University of New Mexico, Albuquerque

Senior Staff

FRANCES E. SHARPLES, Director
KERRY A. BRENNER, Senior Program Officer
MARILEE K. SHELTON-DAVENPORT, Senior Program Officer
EVONNE P.Y. TANG, Senior Program Officer
ROBERT T. YUAN, Senior Program Officer
ADAM P. FAGEN, Program Officer
ANN H. REID, Senior Program Officer
ANNA FARRAR, Financial Associate
ANNE F. JURKOWSKI, Senior Program Assistant
TOVA G. JACOBOVITS, Senior Program Assistant

PREFACE

Toxicogenomics has been described as a discipline combining expertise in toxicology, genetics, molecular biology, and environmental health to elucidate the response of living organisms to stressful environments. It includes the study of how genomes respond to toxicant exposures and how genotype affects responses to toxicant exposures. As the technologies for monitoring these responses rapidly develop, it is critical that scientists and regulators are confident that the technologies are reliable and reproducible and that the data analyses have been validated. To discuss these issues in a public forum, the Committee on the Validation of Toxicogenomic Technologies designed a workshop to consider the current practice and advances in the validation of toxicogenomic technologies. The workshop focused on the technical aspects of validation, recognizing it as a prerequisite for considering other important issues, such as biological validation (e.g., validating the use of microarray "signatures" to describe a toxic effect).

This workshop summary has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the National Research Council's (NRC) Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published workshop summary as sound as possible and to ensure that the summary meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following people for their review of this workshop summary: Federico Goodsaid, William Mattes, Gavin Sherlock, and Mahlet Tadesse.

Although the reviewers listed above have provided many constructive comments and suggestions, they did not see the final draft of the workshop summary before its release. The review of the workshop summary was overseen by Timothy R. Zacharewski, of Michigan State University. Appointed by the NRC, he was responsible for making certain that an independent examination of the workshop summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the workshop summary rests entirely with the committee and the institution.

The committee gratefully acknowledges the following for making presentations at the workshop: Kevin K. Dobbin, National Cancer Institute; Hisham K. Hamadeh, Amgen, Inc.; Wherly P. Hoffman, Eli Lily & Company; Rafael A. Irizarry, Johns Hopkins University Bloomberg School of Public Health; Kyle L. Kolaja, Iconix Pharmaceuticals; Leonard M. Schechtman, Food and Drug Administration; Guido Steiner, F. Hoffmann-La Roche AG; and Weida Tong, FDA National Center for Toxicological Research.

The committee is grateful for the assistance of the NRC staff in preparing this workshop summary: Karl Gustavson and Marilee Shelton-Davenport, program directors; James Reisa, director of the Board on Environmental Studies and Toxicology; Fran Sharples, director of the Board on Life Sciences; Jennifer Saunders, associate program officer; Ruth Crossgrove, senior editor; Mirsada Karalic-Loncarevic, research associate; Radiah Rose, senior editorial assistant; and Lucy Fusco, program associate.

Finally, we thank the members of the committee for their dedicated efforts throughout the development of this workshop summary.

John Quackenbush
Kenneth S. Ramos
Co-Chairs, Committee on Validation of
Toxicogenomic Technologies

CONTENTS

SUMMARY OF THE WORKSHOP

ATTACHMENTS

APPENDIXES

VALIDATION OF TOXICOGENOMIC TECHNOLOGIES

A Workshop Summary

SUMMARY OF THE WORKSHOP

INTRODUCTION

A workshop on the validation of toxicogenomic technologies was held on July 7, 2005, in Washington, DC, by the National Research Council (NRC). The workshop concept was developed during deliberations of the Committee on Emerging Issues and Data on Environmental Contaminants (see Box 1 for a description of the committee and its purpose) and was planned by the ad hoc workshop planning committee (The ad hoc committee membership and biosketches are included in Appendix A.) These activities are sponsored by the National Institute of Environmental Health Sciences (NIEHS). The day-long workshop featured invited speakers from industry, academia, and government who discussed the validation practices used in gene-expression (microarray) assays[1,2] and other toxicogenomic technologies. The workshop also included roundtable discussions on the current status of these validation efforts and how they might be strengthened.

[1] The *microarray* technologies referred to in this report measure mRNA levels in biologic samples. DNA from tens of thousands of known genes (for example, genes that code for toxicologically important enzymes such as cytochrome P450) are placed on small glass slides, with each gene in a specific position. These chips are exposed to mRNA isolated from biologic samples (for example, from rats that have been exposed to a pharmaceutical compound of interest). The mRNA in the sample is treated so that when it hybridizes with the complementary DNA strand on the chip, the resulting complex can be detected. Because the chips can hold DNA from thousands of genes, gene expression (the level of each mRNA) of all these genes can be simultaneously detected.

[2] These technologies are commonly referred to as *gene-expression arrays*, *transcript/transcriptional profiling*, *DNA microarray expression analysis*, *DNA microarrays*, or *gene chips*; more broadly, the use of these technologies is referred to as *transcriptomics*.

BOX 1 Overview of the Committee on Emerging Issues and Data on Environmental Contaminants

The Committee on Emerging Issues and Data on Environmental Contaminants was convened by the National Research Council (NRC) at the request of NIEHS. The committee serves to provide a public forum for communication between government, industry, environmental groups, and the academic community about emerging issues in the environmental health sciences. At present, the committee is focused on toxicogenomics and its applications in environmental and pharmaceutical safety assessment, risk communication, and public policy. A primary function of this committee is to sponsor workshops on issues of interest in the evolving field of toxicogenomics. These workshops are developed by ad hoc NRC committees largely composed of members from the standing committee.

In addition, the standing committee benefits from input from the Federal Liaison Group. The group, chaired at the time of the meeting by Samuel Wilson, of NIEHS, consists of representatives from various federal agencies with interest in toxicogenomic technologies and applications. Members of the Federal Liaison Group are listed in Appendix C of this report.

The workshop agenda (see Appendix B) had two related sections. Part 1 of the workshop, on current validation strategies and associated issues, provided background presentations on several components essential to the technical validation of toxicogenomic experiments including experimental design, reproducibility, and statistical analysis. In addition, this session featured a presentation on regulatory considerations in the validation of toxicogenomic technologies. The presentations in Part 2 of the workshop emphasized the validation approaches used in published studies where microarray technologies were used to evaluate a chemical's mode of action.[3]

This summary is intended to provide an overview of the presentations and discussions that took place during the workshop. This summary only describes those subjects discussed at the workshop and is not intended to be a comprehensive review of the field. To provide greater depth and insight into the presentations from Part 1 of the workshop,

[3] *Mode of action* refers to the pharmacologic or toxicologic end point or event in an organism that is elicited by a compound.

original extended abstracts by the presenters are included as Attachments 1 through 4. In addition, the presenters' slides and the audio from the meeting are available on the Emerging Issues Committee's Web site.[4]

WORKSHOP SUMMARY

Introduction

Kenneth S. Ramos, of the University of Louisville and co-chair of the workshop planning committee, opened the workshop with welcoming remarks, background on the standing and workshop planning committees, and speaker introductions. Ramos also provided a brief historical perspective on the technological advances and applications of toxicogenomics. Beginning in the early 1980s, new technologies, such as those based on polymerase chain reaction (PCR),[5] began to permit evaluation of the expression of individual genes. Recent technological advances (for instance, the development of microarray technologies) have expanded those evaluations to permit the simultaneous detection of the expression of tens of thousands of genes and to support holistic evaluations of the entire genome. The application of these technologies has enabled researchers to unravel complexities of cell biology and, in conjunction with toxicologic evaluations, the technologies are used to probe and gain insight into questions of toxicologic relevance. As a result, the use of the technologies has become increasingly important for scientists in academia, as well as for the regulatory and drug development process.

John Quackenbush, of the Dana-Farber Cancer Institute and co-chair of the workshop, followed up with a discussion of the workshop concept and goals. The workshop concept was generated in response to the standing committee's and other groups' recognition that the promises of toxicogenomic technologies can only be realized if these technologies are validated. The application of toxicogenomic technologies, such as DNA microarray, to the study of drug and chemical toxicity has improved the ability to understand the biologic spectrum and totality of the toxic response and to elucidate potential modes of toxic action. Although early studies energized the field, some scientists continue to question

[4] At http://dels.nas.edu/emergingissues.

[5] PCR is a highly sensitive method that uses an enzyme system to amplify (increase) small amounts of mRNA so that it can be more easily detected.

whether results can be generalized beyond the initial test data sets and the steps necessary to validate the applications. In recognition of the importance of these issues, the NRC committee dedicated this workshop to reflecting critically on the technologies to more fully understand the issues relevant to the establishment of validated toxicogenomic applications. Because transcript profiling using DNA microarrays to detect changes in patterns of gene expression is in many ways the most advanced and widely used of all toxicogenomic approaches, the workshop focused primarily on validation of mRNA transcript profiling using DNA microarrays. Some of the issues raised may be relevant to proteomic and metabolic studies.

Validation can be broadly defined in different terms depending on context. Quackenbush delineated three components of validation: technical validation, biologic validation, and regulatory validation (see Box 2).[6] Because of the broad nature of the topic, the workshop was designed to primarily address technical aspects of validation. For example, do the technologies actually provide reproducible and reliable results? Are conclusions dependent on the particular technology, platform, or method being used?

Part 1: Current Validation Strategies and Associated Issues

The first session of the workshop was designed to provide background information on the various experimental, statistical, and bioinformatics issues that accompany the technical validation of microarray analyses. Presenters were asked to address a component of technical validation from their perspective and experience; the presentations were not intended to serve as comprehensive reviews. A short summary of the topics in each presentation and a discussion between presenters and other workshop participants is presented below. This information is intended

[6] Another aspect of validation discussed by Russell Wolfinger, of the SAS Institute and workshop planning committee member, was statistical validation, which involves verifying that data processing algorithms are performing as intended and are producing results that are reliable, reproducible, specific, and sensitive. However, he commented that consideration of statistical validation separately is debatable because statistical and bioinformatics methods could be viewed as being an integral part of the other three kinds of validation described (technical, biologic, and regulatory).

BOX 2 Validation: Technical Issues Are the First Consideration in a Much Broader Discussion

In general, the concept of validation is considered at three levels: technical, biologic, and regulatory.

Technical validation focuses on whether the technology being used provides reproducible and reliable results. The types of questions addressed are, for example, whether the technologies provide consistent and reproducible answers and whether the answers are dependent on the choice of one particular technology versus another.

Biologic validation evaluates whether the underlying biology is reflected in the answers obtained from the technologies. For example, does a microarray response indicate the assayed biologic response (for example, toxicity or carcinogenicity)?

Regulatory validation begins when technical and biologic validation are established and when the technologies are to be used as a regulatory tool. In this regard, do the new technologies generate information useful for addressing regulatory questions? For example, do the results demonstrate environmental or human health safety?

to be accessible to a general scientific audience. The reader is referred to the attachments by the presenters of this report for greater technical detail and a comprehensive discussion of each presentation.

Experimental Design of Microarray Studies

Kevin Dobbin, of the National Cancer Institute, provided an overview of experimental design issues encountered in conducting microarray assays. Dobbin began by discussing experimental objectives and explaining that there is no one best design for every case because the design must reflect the objective a researcher is trying to achieve and the practical constraints of the experiments being done. Although the high-level goal of many microarray experiments is to identify important pathways or genes associated with a particular disease or treatment, there are different ways to approach this problem. Thus, it is important to clearly define the experimental objectives and to design a study that is driven by those objectives. Experimental approaches in toxicogenomics can typically be grouped into three categories based on objective: class comparison, class prediction, or class discovery (see Box 3 and the description in Attachment 1).

BOX 3 Typical Experimental Objectives in mRNA Microarray Analyses

Class Comparison

Goal: Identify genes differentially expressed among predefined classes of samples.

Example: Measure gene products before and after toxicant exposure to identify mechanisms of action (Hossain et al. 2000).

Example: Compare liver biopsies from individuals with chronic arsenic exposure to those of healthy individuals (Lu et al. 2001).

Class Prediction

Goal: Develop a multigene predictor of class membership.

Example: Identify gene sets predictive of toxic outcome (Thomas et al. 2001).

Class Discovery

Goal: Identify sets of genes (or samples) that share similar patterns of expression and that can be grouped together. Class discovery can also refer to the identification of new classes or subtypes of disease rather than the identification of clusters of genes with similar patterns.

Example: Cluster temporal gene-expression patterns to gain insight into genetic regulation in response to toxic insult (Huang et al. 2001).

Dobbin's presentation outlined several experimental design issues faced by researchers conducting microarray analyses. He discussed the level of biologic and technical replication[7] necessary for making statistically supported comparisons between groups. He also discussed issues related to the study design that arise when using dual-label microarrays,[8]

[7] Biologic replicates are mRNA samples from separate individual subjects that were experimentally treated in an identical manner (for example, five mRNA isolates from each identically exposed animal). Technical replicates would, for example, be tests of different sample aliquots drawn from the same biologic sample.

[8] Microarray technologies use two different approaches to detecting RNAs that have hybridized to the DNA probes on the array. Single-label technologies use a single fluorescent dye to detect hybridization of a single RNA sample to a single array, and comparisons are then made between arrays. Dual-label technologies compare two samples on each array by labeling each RNA with a unique fluorescent dye (often represented as red and green) before applying them to the arrayed probes.

including strategies for the selection of samples to be compared on each microarray, the use of control samples, and issues related to dye bias.[9] The costs and benefits of pooling RNA samples for analysis on microarrays were discussed in relation to the study's design and goals. As an example to help guide investigators, Dobbin presented a sample-size formula to determine the number of arrays needed for a class comparison experiment (see Equation 1). This formula calculates the statistical power of a study based on the variability estimates of the data, the number of arrays, the level of technical replication, the target fold-change in expression that would be considered acceptable, and the desired level of statistical significance to be achieved (see Attachment 1 for further details).

The ensuing workshop discussion on Dobbin's presentation focused on the interplay between using technical replicates and using biologic replicates. Dobbin emphasized the importance of biologic replication compared with technical replication for making statistically powerful comparisons between groups, because it captures not only the variability in the technology but also samples the variation of gene expression within a population.

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m} \right), \qquad (1)$$

where
n = number of arrays needed
m = technical replicates per sample
δ = effect size on base 2 log scale (e.g., 1 = 2-fold)
α = significance level (e.g., .001)
1-β = power
z = normal percentiles (*t* percentiles preferable)
t_g^2 = biological variation within class
s_g^2 = technical variation.

[9] When two dyes are used, slight differences in their efficiencies at each step in the process—labeling, hybridization, and detection—can cause systematic biases in the measurements that must be estimated from the data and then removed so that effective comparisons can be made.

Multiple-Laboratory Comparison of Microarray Platforms

Rafael Irizarry, of Johns Hopkins University, described published studies that examined issues related to reproducibility of microarray analyses and focused on between-laboratory and between-platform comparisons. The presentation examined factors driving the variability of measurements made using different micorarray platforms (or other mRNA measurement technologies), including the "lab effect,"[10] practitioner experience, and use of different statistical-assessment and data-processing techniques to determine gene-expression levels. Irizarry's presentation focused on understanding the magnitude of the lab effect, and he described a study where a number of laboratories analyzed the same RNA samples to assess the variability in results (Irizarry et al. 2005). Overall, the results suggest that labs using the Affymetrix microarray systems have better accuracy than the two-color platforms, although the most accurate signal measure was attained by a lab using a two-color platform. In this analysis, a small group of genes had relatively large-fold differences between platforms. These differences may relate to the lack of accurate transcript information on these genes. As a result, the probes used in different platforms may not be measuring the same transcript. Moreover, disparate results may be due to probes on different platforms querying different regions of the same gene that are subject to alternative splicing or that exhibit divergent transcript stabilities.

Beyond describing the results of the analysis, Irizarry provided suggestions for conducting experiments and analyses to compare various microarray platforms. The suggestions included use of relative, as opposed to absolute, measures of expression; statistical determinations of precision and accuracy; and specific plots to determine whether genes are differentially expressed between samples. These techniques are described in Attachment 2. Irizarry also commented that reverse transcriptase PCR (RTPCR) should not be considered the gold standard for measuring gene expression and that the variability in RTPCR data is very similar to microarray data if enough data points are analyzed. In this regard, the large quantity of data produced by microarrays is useful in describing the variability in the technology's response. However, this attribute is sometimes portrayed as a negative because the data can appear variable. Conversely,

[10] The *lab effect* relates to differences in results from different laboratories that may relate to, for example, analyst techniques, lab equipment, or differences in reagents.

RTPCR produces comparatively few measurements, and one is not able to readily assess the variability.

Irizarry also commented that obtaining a relatively low correspondence between lists of genes generated by different platforms is to be expected when comparing just a few genes from the thousands of genes analyzed. On this point, it was questioned how and whether researchers can migrate from the common practice of assessing 1,000s of genes and selecting only a few as biomarkers to the practice of converging on a smaller number of genes that reliably predict the outcome of interest. Also, would a high-volume, high-precision platform be a preferred alternative? Further questions addressed measurement error in microarray analyses and whether, because of the magnitude of this error, it was possible to detect small or subtle changes in mRNA expression. In response, Irizarry emphasized the importance of using multiple biologic replicates so that consistent patterns of change could be discerned.

Statistical Analysis of Toxicogenomic Microarray Data

The next presentation by Wherly Hoffman, of Eli Lilly and Company, discussed the statistical analysis of microarray data. This presentation focused on the Affymetrix platform and discussed the microarray technology and statistical hypotheses and analysis methods for use in data evaluation. Hoffman stated that, like all microarray mRNA expression assays, the Affymetrix technology uses gene probes that hybridize to mRNA (actually to labeled cDNA derived from the mRNA) in biologic samples. This hybridization produces a signal with intensity proportional to the amount of mRNA contained in the sample. There are various algorithms that may be used to determine hybridized mRNA signal intensity from background signals.

Hoffman emphasized the importance of defining the scientific questions that any given experiment is intended to address and the importance of including statistical expertise early on in the process to determine appropriate statistical hypotheses and analyses. During this presentation, three types of experimental questions were addressed along with the statistical techniques for their analysis (as mentioned by Hoffman, these techniques are also described in Deng et al. 2005). The first example presented data from an experiment designed to identify differences in gene expression in animals exposed to a compound at several different doses. Hoffman discussed the statistical techniques used to evaluate differences in expression between exposure levels while considering variation in responses from similarly dosed animals and variation in responses from replicate microarrays. In this analysis (using a one-factor [dose] nested analysis of variance [ANOVA] and t-test), it is essential to accurately define the degrees of freedom. Hoffman pointed out that the degree of freedom is determined by the number of animal subjects and not the number of chips (when the chips are technical replicates that represent application of the same biologic sample to two or more microarrays). Thus, technical replicates should not be included when determining the degrees of freedom. If this is not factored into the calculation, the *P* value is inappropriately biased because exposure differences appear to have greater significance. The second example included data from an experiment designed to evaluate gene expression over a time course. The statistical analysis on this type of experiment must capture the dose effect, the time effect, and the dose-time interaction. Here, a two-factor (dose and time) ANOVA is used. The third example provided by Hoffman was an experiment to determine those genes affected by different classes of compounds (alpha, beta, or gamma receptor agonists). This analysis evaluated dose-response trends of microarray signal intensities when known peroxisomal proliferation activated receptor (PPAR) agonists were tested on agonist knockout and wild-type mice to determine those probe sets (genes) that responded in a dose-response manner. Here, a linear regression model is used for examining the dose-response trends at each probe set. This model considers the type of mice (wild type or mutant), the dose of the compound, and their interaction.

Hoffman also discussed graphical tools to detect patterns, outliers, and errors in experimental data, including box plots, correlation plots, and principal component analysis (PCA). Other visualization tools, such as clustering analysis and the use of volcano plots used to show the general patterns of microarray analysis results, were also presented. These tools are further discussed in Attachment 3.

Finally, multiplicity issues were discussed. Although microarray analyses are able to provide data on the expression of thousands of genes in one experiment, there is the potential to introduce a high rate of false positives. Hoffman explained various approaches used to control the rate of false positives, including the Bonferroni approach, but commented that recent progress in addressing the multiple testing problems has been made, including work by Benjamini and Hochberg (1995). (These approaches as well as the relative advantages and disadvantages are further discussed in Attachment 3.)

The short discussion following this presentation centered primarily on the visualization tools presented by Hoffman and the type of information that they convey.

Diagnostic Classifier—Gaining Confidence Through Validation

Clinical diagnosis of disease primarily relies on conventional histological and biochemical evaluations. To use toxicogenomic data in clinical diagnostics, reliable classification methods[11] are needed to evaluate the data and provide accurate clinical diagnoses, treatment selections, and prognoses. Weida Tong, of the Food and Drug Administration (FDA), spoke about classification methods used with toxicogenomic approaches in clinical applications. These classification methods (learning methods) are driven by mathematical algorithms and models that "learn" features in a training set (known members of a class) to develop diagnostic classifiers and then classify unknown samples based on those features. Tong's presentation focused on the issues and challenges associated with sample classification methods using supervised[12] learning methods.

The development of a diagnostic classifier can be divided into three steps: training, where gene expression or other toxicogenomic profiles are correlated with clinical outcomes to develop a classifier; validation, where profiles are validated using cross-validation[13] or external valida-

[11] Classification methods are algorithms used to assign test cases to one of a number of designated classes (StatSoft, Inc. 2006). Most classification schemes referred to in this workshop report refer to classifying a chemical compound based on mode of toxicologic action. Another common scheme is the classification of a biologic sample (for example, classifying a tumor into subtypes based on invasiveness potential).

[12] The term *supervised* learning is usually applied to cases in which a particular classification is already observed and recorded in a training data set, and one wants to build a model to predict the class of a new test sample. For example, one may have a data set from compounds with a known mode of toxicologic action. The purpose of the classification analysis would be to build a model to predict which compounds (from tests of unknown compounds) would be in the same class as the test data set.

[13] Cross-validation is a model evaluation method that indicates how well the learning method will perform when asked to make new predictions for data not already seen. The basic premise is not to use the entire data set when training a

tion[14] approaches; application, where the classifier is used to classify an unknown subject for a clinical diagnosis or for biomarker identification (see Figure 1 and Attachment 4). In this presentation, the "decision forest" method, developed by Tong et al. (2004), was discussed with an emphasis on prediction confidence and chance correlation.[15] The decision forest approach is a consensus modeling method; that is, it uses several classifiers instead of a single classifier (hence, the decision forest instead of a decision tree) (see Box 4). This technique may be used with microarray, proteomics, and single-nucleotide polymorphism data sets. An example of this technique was presented that used mass spectra from protein analyses of serum from individuals to distinguish patients with prostate cancer from healthy individuals. Here, mass spectra peaks were used as independent variables for classifiers. Initially, only a few peaks were identified as classifiers and run on the entire pool of healthy individuals and cancer patients; this analysis is considered a decision tree and has an associated error (misclassification) rate. Combining decision trees (additional runs with distinct classifiers) into a decision forest improves the predictive accuracy.

Tong emphasized that validating a classifier has three components: the first is determining whether the classifier accurately predicts unknown samples; the second is determining the prediction confidence for classifying different samples or individuals; and the third is establishing that correlations between a diagnostic classifier and disease are not just because of chance (chance correlation). Tong's presentation focused on the techniques to evaluate predictive confidence and chance correlation and emphasized the usefulness of a 10-fold cross-validation technique in providing an unbiased statistical assessment of prediction confidence and chance correlation (see Attachment 4).

Discussion following Tong's presentation focused on the distinction between external validation methods and details surrounding the cross-validation methods (described in Attachment 4 and Tong et al.

learning method, so some data are removed before training begins. After training is completed, the removed data can be used to test the performance of the learned model on "new" data (Schneider and Moore 1997).

[14] External validation is the process where the accuracy of a model's prediction is tested on samples independent of those used in the training set.

[15] Because of the large number of predictor variables (proteins, mRNA transcripts, etc.) and the relatively small number of samples, it is possible that the patterns identified by a classification model could be due to chance.

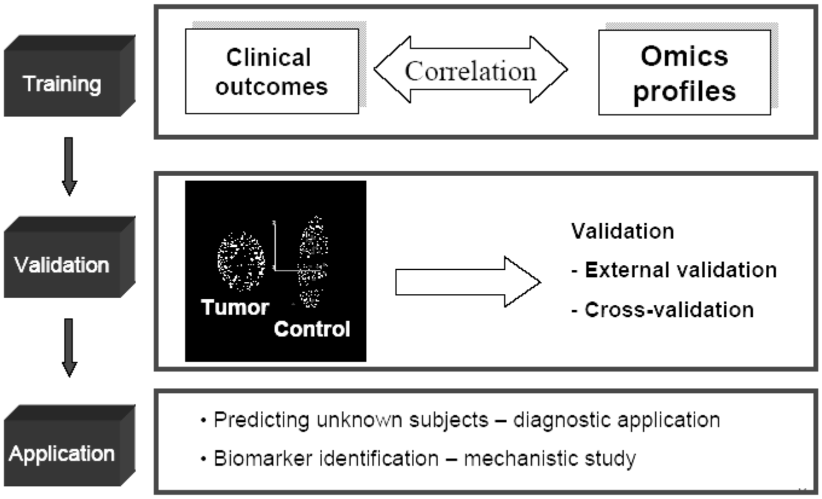


FIGURE 1 Three steps in the development of a diagnostic classifier. Source: Tong 2004.

BOX 4 Decision Forest Analysis for Use with Toxicogenomics Data

Decision forest (DF) is a consensus modeling technique that combines multiple decision tree models in a manner that results in more accurate predictions than those derived from an individual tree. Since combining several identical trees produces no gain, the rationale behind decision forests is to use individual trees that are different (that is, heterogeneous) in representing the association between independent variables (gene expression in DNA microarray, m/z peaks in SELDI-TOF data, and structural descriptors in SAR modeling) and the dependent variable (class categories) and yet are comparable in their prediction accuracy. The heterogeneity requirement assures that each tree uniquely contributes to the combined prediction. The quality comparability requirement assures that each tree makes a similar contribution to the combined prediction. Since a certain degree of noise is always present in biologic data, optimizing a tree inherently risks overfitting the noise. Decision forest tries to minimize overfitting by maximizing the difference among individual trees to cancel some random noise in individual trees. The maximum difference between the trees is obtained by constructing each individual tree using a distinct set of dependent variables.

Source: Modified from Tong 2006.

2004). In addition, questions were raised about the extent to which established classifiers could be extrapolated beyond the original training set. Tong indicated that the results of the cross-validation technique could describe the predictive accuracy of an established classifier within the confines of the original data set but not new, independent data sets.

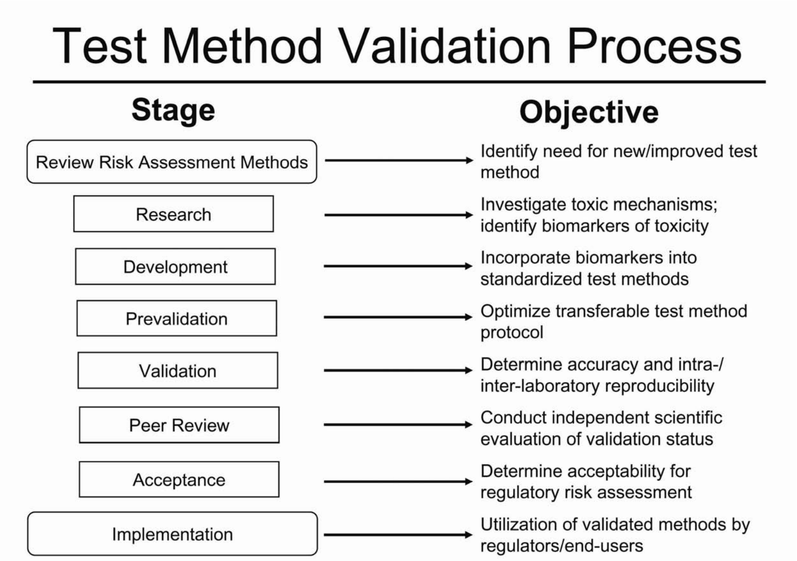
Toxicogenomics: ICCVAM Fundamentals for Validation and Regulatory Acceptance

The final presentation of the morning session was by Leonard Schectman, the chair of the Interagency Coordinating Committee on Validation of Alternative Methods (ICCVAM). This presentation described the validation and regulatory acceptance criteria and guidelines that are currently in place and have been compiled and adopted by ICCVAM and its sister agency the European Center for Validation of Alternative Methods.

At present, the submission of toxicogenomic data to regulatory agencies is being encouraged (for example, FDA 2005). However, the regulatory agencies generally consider it premature to base regulatory decisions solely on toxicogenomics data, given that the technologies are rapidly evolving and in need of further standardization, validation, and understanding of the biologic relevance. In addition, regulatory acceptability and implementation will in part depend on whether these methods have utility for a given regulatory agency and for the products that that agency regulates.

Schectmann described ICCVAM's 2003 updated guidelines for nomination and submission of methods (ICCVAM 2003). These guidelines detail ICCVAM validation and regulatory acceptance criteria. Figure 2 outlines the generalized scheme of the validation process, as presented by Schechtman. Components of this process include standardization of protocols, variability assessments, and peer review of the test method. The presentation concluded with the overall comment that validation in the regulatory arena is, for the most part, a prerequisite for regulatory acceptance of a new method.

In response to the presentation, it was questioned whether regulatory agencies were required to go through the ICCVAM process before they could use or accept information from a new test. Schectmann responded that it was not required—the process is made available to help guide a validation effort, and because multiple agencies are part of the



Test Method Validation Process

Stage	Objective
Review Risk Assessment Methods	Identify need for new/improved test method
Research	Investigate toxic mechanisms; identify biomarkers of toxicity
Development	Incorporate biomarkers into standardized test methods
Prevalidation	Optimize transferable test method protocol
Validation	Determine accuracy and intra-/ inter-laboratory reproducibility
Peer Review	Conduct independent scientific evaluation of validation status
Acceptance	Determine acceptability for regulatory risk assessment
Implementation	Utilization of validated methods by regulators/end-users

FIGURE 2 ICCVAM test method validation process. Source: ICCVAM 2003.

ICCVAM process, an ICCVAM-accepted test is likely to be accepted by those agencies. However, he cautioned that acceptance of any given method goes far beyond validation, and the ICCVAM process is one that facilitates the validation of a method but does not provide or guarantee regulatory acceptance of that method.

It was suggested by a participant that one aspect of the test-validation process (distribution of chemicals for testing), as outlined in the presentation, would not work well in the field of toxicogenomics but that the distribution of biologic samples (for mRNA quantification) would be a better alternative. Schectmann clarified that many new technologies did not exist when the ICCVAM process was initiated and that other validation approaches could be used. He emphasized that there is nothing about the ICCVAM process that is inflexible relative to the new or different technologies.

The fundamental differences between the processes for validating new technologies and those used to validate conventional, currently used toxicological methods were discussed next. It was noted that there is an apparent disconnect in that a very elaborate validation process is established for new methods, yet thousands of chemicals are currently being evaluated with methods (such as quantitative-structure-activity relationships) that would likely not pass through the current ICCVAM validation process. Overall, it was questioned whether this process sets up a system where "the perfect was the enemy of the good," because new technologies can offer information, for instance, in a chemical's weight-of-evidence evaluation? Schechtman responded that he did not believe that it was necessary to wait for a final stamp of approval. Indeed, the U.S. Environmental Protection Agency (EPA) and FDA are accepting data and mechanistic information from tests that have not undergone, and probably will never undergo, the ICCVAM validation process. Even the classical toxicological tests themselves have never been validated in this manner.

Part 2: Case Studies: Classification Studies and the Validation Approaches

The second session of the workshop featured case studies where mRNA expression microarray assays were used to classify compounds according to their toxicological mode of action. Authors of the original papers presented salient details of their studies, emphasizing validation techniques and concepts. The presentations and discussion are described below and the author's PowerPoint slides are available on the committee's Web site. As mentioned before, this report is intended to present the information at a level accessible to a general scientific audience. Technical details on the presentations are presented at a cursory level. Readers are referred to the original publications, cited in each section, for greater technical detail and a more comprehensive treatment of specific protocols.

Proof-of-Principle Study on Compound Classification Using Gene Expression

Hisham Hamadeh, of Amgen, outlined a two-part proof-of-principle study on compound classification that used microarray technologies (Hamedah et al. 2002a,b). This study was initiated in 1999 when many of these technologies were in their infancy and current validation techniques had not yet been devised. However, the experimental design and concepts used for validation and classification in those early studies remain illustrative for discussion. The purpose of the study was to determine whether gene-expression profiles resulting from exposure to various compounds could be used to discriminate between different toxicological classes of compounds. This study evaluated the gene-expression profiles resulting from exposure to two compound classes, peroxisome proliferators (including three test compounds: clofibrate, Wyeth 14,643, and gemfibrozil) and enzyme inducers (modeled by the cytochrome P450 inducer, phenobarbital). Hamadeh described the experimental design of the study, highlighting data analyses used to designate whether gene expression was significantly induced.

Gene induction results were presented using hierarchical clustering,[16] principal components analysis, and pairwise correlation. These visualization techniques demonstrated that although phenobarbital-exposed animals exhibited significant interanimal variability, they could be readily distinguished from those exposed to the peroxisome proliferators on the basis of gene expression. To expand on results obtained with this limited data set, the researchers attempted to classify blinded samples based on earlier data. A classifier using 22 genes with the greatest differential expression between the two compound classes was used to classify unknown samples into a compound class. This gene set was determined by statistical analyses of the training set (tests on the model compounds described above) using linear discriminant analysis and a genetic algorithm for pattern recognition (Hamadeh et al. 2002b). Blinded samples were classified initially by visual comparison of the levels of mRNA induction or repression in blind samples to the known compounds. Subsequently, pairwise correlation analysis of expression level of the 22 discriminant genes was also used. Correlations of $r \geq 0.8$ between blinded and known samples were used to determine whether the unknown was similar to the known class.

The analysis was able to successfully discern the identity of the blinded compounds. Phenytoin, an enzyme inducer similar to phenobarbital, was classified as phenobarbital-like; DEHP, a peroxisome proliferator, was also indicated as such; and the final compound, hexobarbital, has a similar structure to phenobarbital but is not an enzyme inducer, was not classified as being either phenobarbital-like or a peroxisome proliferator. Overall, the conclusions of this study are that it was possible to

[16] Hierarchical clustering groups similar objects into a sequence of nested partitions, where the similarity metric is predefined. In DNA microarray applications, the technique is used to identify genes with similar expression patterns.

separate compounds based on the gene-expression profiles and that it is feasible to gain information on the toxicologic class of blinded samples through interrogation of a gene-expression database.

Workshop participants were interested in the suite of discriminant genes that were used for the evaluation of chemical class and in whether that number of genes could be narrowed down. For instance, the question was asked whether it would be satisfactory to use only the induction profiles of CYP2B and CYP4A[17] to indicate the class of the unknowns. Hamadeh reported that this type of evaluation had been conducted and the number of discriminant genes could indeed be narrowed down. However, he noted that a larger number of discriminant genes allow for increased resolution between compounds. Of course, microarray analysis also provides information on many genes that would not be obtained from a simple evaluation of individual gene transcripts, and this is particularly useful when analyzing unknown samples.

The amount and origin of the variability seen within a chemical class was also discussed. Hamadeh explained that there was interanimal variability but that generally the variability in the microarray responses mirrored those seen in the animal responses (for example, whether animals within a group exhibited hypertrophy, necrosis, or the presence of lesions). Overall, the level of interanimal variability did not alter the end result that expression profiles were different for the different classes.

The discussion emphasized that mRNA expression results have several layers of intertwined information that can complicate the analysis of factors eliciting gene-expression changes. Beyond the molecular targets that are specifically affected by a compound, there are expression changes associated with the pathology resulting from exposure (for example, necrosis or hypertrophy). Gene-expression changes can also be related to an event that is secondary, or downstream, from the initial toxicologic interaction. A compound may also interact with other targets not associated with its toxic or therapeutic action. In addition, all of these effects may change, depending on time after dose, which adds another layer of complexity to the analysis. As a result, the number of genes that are used to screen for certain chemical classes is generally low and intended to screen for certain toxicities.

[17] Cytochrome P450 2B and 4A (CYP2B and CYP4A) are members of the cytochrome P450 family of proteins that catalyze mono-oxygenation of endogenous and exogenous substrates.

Acute Molecular Markers of Rodent Hepatic Carcinogenesis Identified by Transcription Profiling

Kyle Kolaja, of Iconix Pharmaceuticals, presented a study that sought to identify biomarkers of hepatic carcinogenicity using microarray mRNA expression assays (Kramer et al. 2004). In particular, identifiers of nongenotoxic carcinogenicity were desired because the conventional method for determining this mode of action (a 2-year rodent carcinogenicity assay) is time consuming and expensive. The study evaluated nine well-characterized compounds, including five nongenotoxic rodent carcinogens, one genotoxic carcinogen, one carcinogen that may not act via genotoxicity, a mitogen,[18] and a noncarcinogenic toxicant. Rats from the control group and three groups that received different dose levels of each compound were sacrificed after 5 days of dosing, and liver extracts were tested in microarray assays. The purpose of the analysis was to correlate the short-term changes in gene expression with the long-term incidence of carcinogenicity (known from previous studies of these model compounds). Kolaja highlighted the data analysis used to designate whether gene expression was significantly induced or repressed. Significantly affected genes were correlated to carcinogenic index (based on cancer incidence in 2-year rodent carcinogenicity studies).

The study resulted in the identification of two optimal discriminatory genes (biomarkers): cytochrome P450 reductase (CYP-R) and transforming growth factor-β stimulated clone 22 (TSC-22). TSC-22 negatively correlated with carcinogenic potential, and CYP-R correlated with carcinogenicity. The results were validated initially by measuring the mRNA levels using another mRNA measurement technique, quantitative PCR (Q-PCR). This analysis indicated a strong correlation between the microarray data and the Q-PCR data generated from the same set of samples. From a biologic standpoint, the role of TSC-22 in carcinogenesis is consistent with its involvement in the regulation of cellular growth, development, and differentiation.

The results of this analysis were extended by a “forward validation” of these biomarkers, that is, the independent determination of these genes as carcinogenic biomarkers by other groups or studies. Kolaja described two independent studies (Iida et al. 2005; Michel et al. 2005) using both rats and mice that identified TSC-22 as a potential marker of early

[18] Mitogens induce cell division.

changes that correlate with carcinogenesis. Additional study at Iconix Pharmaceuticals on 26 nongenotoxic carcinogens and 110 noncarcinogens indicated that the TSC-22 biomarker at day 5 after dosing had an accuracy for detecting the carcinogens of about 50% and for excluding compounds as carcinogens at about 80%. Kolaja remarked that these results were fairly robust, especially recognizing that the biomarker is a single-gene biomarker being compared across a very diverse set of compounds. Overall, it is very difficult to find one gene that is a suitable biomarker in terms of predictive performance. It was noted that multiple genes create a more integrated screening biomarker and allow for stronger predictivity, performance, and accuracy. Kolaja also stated that in the future it would be more appropriate for validation strategies to emphasize the biologic and not methodologic aspects of the validation, because the testing of a biologic question captures the technical aspects. As such, additional tests on treatments and models would follow with less emphasis on platforms and methods.

During the discussion, it was questioned whether it was possible that TSC-22 was correlative rather than mechanistic—that is, if the TSC-22 was related to another general response (such as liver weight change) and not to carcinogenesis? Kolaja mentioned that he would not be surprised if liver weight changes were also seen at day 5 and that the possibility that TSC-22 was a correlative response had not been ruled out. Another question raised was whether analyzing data sets using a multiple-gene biomarker had correspondingly greater technical difficulty compared with a single-gene biomarker? Kolaja indicated that it was the same type of binary analysis (Is a sample in the class or not?), but with multiple genes, the answer relies on the compendium of genes, and the mathematical modeling. He also noted that recent mathematical algorithms and models have become increasingly better at class separation.

Study Design and Validation Strategies in a Predictive Toxicogenomics Study

Guido Steiner, of Roche Pharmaceuticals, presented a study that used microarray analyses to classify compounds by mode of toxicologic action (Steiner et al. 2004). The goals of the study were to predict hepatotoxicity of compounds from gene-expression changes in the liver with a model that can be generalized to new compounds to classify compounds according to their mechanism of toxicity and to show the viability of a supervised learning approach without using prior knowledge about relevant genes.

In this study, six to eight model hepatoxicants for each known mode of action (for example, steatosis, peroxisome proliferation, and choleostasis) were tested. Rat liver extracts were obtained at various times (typically under 24 hours) following dosing with model compounds and tested for changes in mRNA expression. Clinical chemistry, hematology, and histopathology were used to assess toxicity in each animal. The gene-expression data from tests of the model toxicants became the training set for the supervised learning methods (in this study, support vector machines [SVMs]) (see Box 5). One aspect of this study that differed from many comparisons of gene-expression levels is that the commonly used statistical measures denoting significant gene-expression changes (magnitude of change and associated *P* value) were not used. Rather, the discriminatory features from the microarray results for classification were selected using recursive feature elimination (RFE), a method that uses the output of the SVMs to extract a compact set of relevant genes that as an ensemble yield a good classification accuracy and stabilize against the background biologic and experimental variation (see Box 5 and Steiner et al. 2004). Features for a particular class were selected from gene-expression profiles from animals exposed to model compounds. A compound's class was based on results from the serum chemistry profile and liver histopathology.

Steiner's presentation focused on the study design and validation considerations that need to be addressed when conducting this type of study. First, only a small set of compounds within a class are typically available for developing classification algorithms, and it is important to consider whether these compounds are adequately representative of class toxicity. Overall, this problem is difficult to predict or avoid a priori. In the presented study, some well-characterized treatments were initially selected, and the problem of generalizability within a toxic class was dealt with during the model validation phase. So then, how is a well-characterized training set defined? This question was approached by carefully selecting the compound using phenotypic anchoring based on the clinical chemistry and histopathologic data, subsequently confirming that the clinical results correspond with those in the literature, and then using the higher-dose treatments in the training sets that had no ambiguity regarding the toxic manifestation. One implication of this approach is that the "scale" for detecting an effect is set higher (that is, gene-expression signatures in the training set are based on higher dose, "real" effects at the organ level). Also, although the meaning for this class (in terms of gene expression) is well defined, the ability to extrapolate to lower doses is not known until this question is tested.

Another issue is that the toxicity classes established by researchers may not be accurate. Some tested compounds may show a mixed toxicity. Steiner explained that the model would pick up the various aspects of the toxicity, and indeed, results presented in Steiner et al. (2004) indicated that to be the case.

BOX 5 Classification of Microarray Data Using Algorithms and Learning Methods

Various methods are used to analyze large-scale gene-expression data. Unsupervised methods widely reported in the literature include agglomerative clustering (Eisen et al. 1998), divisive clustering (Alon et al. 1999), K-means clustering (Everitt 1974), self-organizing maps (Kohonen 1995), and principal component analysis (Joliffe 1986). Support vector machines (SVMs), on the other hand, belong to the class of supervised learning algorithms. Originally introduced by Vapnik and co-workers (Boser et al. 1992; Vapnik 1998), they perform well in different areas of biologic analysis (Schölkopf and Smola 2002). Given a set of training examples, SVMs are able to recognize informative patterns in input data and make generalizations on previously unseen samples. Like other supervised methods, SVMs require prior knowledge of the classification problem, which has to be provided in the form of labeled training data. Used in a growing number of applications, SVMs are particularly well suited for the analysis of microarray expression data because of their ability to handle situations where the number of features (genes) is very large compared with the number of training patterns (microarray replicates). Several studies have shown that SVMs typically tend to outperform other classification techniques in this area (Brown et al. 2000; Furey et al. 2000; Yeang et al. 2001). In addition, the method proved effective in discovering informative features such as genes that are especially relevant for the classification and therefore might be critically important for the biologic processes under investigation. A significant reduction of the gene number used for classification is also crucial if reliable classifiers are to be obtained from microarray data. A proposed method to discriminate the most relevant gene changes from background biologic and experimental variation is gene shaving (Hastie et al. 2000). However, we chose another method, recursive feature elimination (RFE) (Guyon et al. 2002), to create sets of informative genes.

Source: Steiner et al. 2004.

Data heterogeneity is also an issue, and a primary reason protocol standardization and chip quality control is essential. Time-matched vehicle controls[19] should be used; according to Steiner, it is a necessity because changes occur (for instance, due to variations in circadian rhythms, age, or the vehicle) through an experiment's time course. To handle the known remaining heterogeneities, the SVM models were always trained using a one-versus-all approach where data points of all toxicity classes are seen at the same time. In this setting, chances are good that any confounding pattern is represented on both sides of the classification boundary. Therefore, the SVM can "learn" the differences that are related to toxicity class (which are designated) and ignore the patterns that are not related to toxicity class (experimental factors driving data heterogeneity).

Another issue that needs to be considered is data overfitting. This effect can occur when the number of features in the classification model is too great compared with the number of samples, and it can lead to spurious conclusions. In this regard, the SVM technique has demonstrated performance when the training set is small (Steiner et al. 2004). Selection of model attributes in SVMs, including the aforementioned RFE function, also limit the potential for overfitting. However, the true performance of the model has to be demonstrated using a strict validation scheme that also takes into account that a number of marker genes have to be selected from a vast excess of available (and largely uninformative) features.

Steiner also stated that a compound classification model should not confuse gene-expression changes associated with a desired pharmacological effect with those from an unwanted toxic outcome. The SVM model addresses this concern based on the assumption that pharmacological action is compound specific and the toxic mechanism is typical for a whole class; if this is true, then the SVM will downgrade features associated with a compound-specific effect and find features for classification that work for all compounds within a class.

The final issue considered by Steiner was that of sensitivity and the need for a model and classification scheme to be at least as sensitive as the conventional clinical or histological evaluations. In this study, increased sensitivity of the developed classification scheme was demonstrated with a sample that had no effect using conventional techniques,

[19] For example, dosed animals at day 1 would be compared with control animals at day 1 and so on for each time point throughout the experiment.

but there was a small shift from the control group toward the active group in a three-dimensional scatter plot for visualizing class separation. This shift is a hint that gene-expression profiling could be more sensitive than the classical end points used in this study (Steiner et al. 2004).

Discussion from workshop participants included questions about whether the described systems were capable of detecting effects at a lower dose or if they were only detecting effects at an earlier time point (that is, the effect would have been manifested at the dose but at a later time). Steiner explained that those assessments had been completed and that the model worked quite well in making correct predictions from lower doses than those that elicit classic indicators of toxicity.

During discussion, it was noted that Steiner's data set indicated differences in responses between strains of rats, which has important implications for cross-species extrapolation (for example, between rodents and humans). Notable differences seen in microarray results between two inbred strains of rats might presage the inapplicability of these techniques to humans. Steiner replied that this was an important question not addressed in the study but that the authors did not imply that the effect in humans could be predicted from the rat data. Hamedeh pointed out that the training method used in that example data set did not consider both strains in the training set, and identifiers could have likely been found if this had been done. However, the extrapolation of this classification scheme to humans would create a whole different set of issues because those analyses would be conducted with different microarray chips (human based not rat based).

Roundtable Discussion

A roundtable discussion, moderated by John Quackenbush and open to all audience members, was held following the invited presentations, and the strengths and limitations of the current validation approaches and methods to strengthen these approaches were considered. Although technical issues and validation techniques were discussed, many of the comments focused on biologic validation, including the extent to which microarray results indicated biologic pathways, the linking of gene-expression changes to biologic events, the different requirements of biologic and technical validation, the impact of individual, species and environmental variability on microarray results, and the use of microarray assays to evaluate the low-dose effects of chemicals. The primary themes of this discussion are presented here.

Technical Issues and Validation Techniques

John Balbus, of Environmental Defense, commented that most of the presentations during the day were from case studies of self-contained data sets from a particular lab or group. However, it is commonly thought that a benefit of obtaining toxicogenomic data is that they will be included in larger databases and mined for further information. In this context, he noted the level of difficultly in drawing statistically sound conclusions from self-contained data sets and asked whether analyzing a fully populated database would create even greater complexities and whether it was possible to achieve sufficient statistical power from data mining. John Quackenbush suggested that a level of data standardization would be necessary to analyze a compiled database and that the quality of experiments in the database would exert a major influence. In addition, these analyses may require that comparisons are only made between similar technologies or applications. Irizarry suggested that a large database would also serve as a resource of independent data sets for evaluating whether a phenomenon seen in one experiment has been seen in others.

Casimir Kulikowski, of Rutgers University, raised a technical issue relating to cross-validation techniques used in binary classification models in toxicogenomics. Those models usually involve very different types of categories for a positive response for a specific compound versus other possible responses. In cross-validation, most techniques assume symmetry with random sampling from each class. In Steiner's presentation, the sampling was appropriately compound specific, but a question arises as to whether it could also take into account confounding issues not known a priori, such as the possibility of a compound differentially affecting different biologic pathways. More generally, cross-validation methods may need to be applied in a more stratified manner for problems dealing with multiple classes or mixed classes, or where there are relationships between the classes. For instance, there may be a constrained space for the hypothesis of a toxic response affecting a single (regulatory or metabolic) pathway, but one may also wish to focus on other constraints that have not yet been satisfied to generate additional information for other pathways. One approach would be to use causal pathway analysis, together with its counterfactual[20] network, to limit the possible outcomes of hypothesis generation. This means that if a set of assertions can

[20] A counterfactual conditional is an "if-then" statement indicating what would be the case if its antecedent were true.

be made based on the current state of the art, the investigator can identify those counterfactual questions that might actually be scientifically interesting. (This technique has currently been proposed for biomedical image classification, but it might also apply to microarray-based classification studies.) Kulikowski commented that it was problematic to design classifiers as simple, binary classifications and then assume that the toxic response class is the same as the class representing the mixture of all other responses. It would be more desirable to tease out what that mixture class is and then figure out how it should be stratified in a systematic manner.

Microarray Assays for the Analysis of Biologic Pathways

Federico Goodsaid, of FDA, commented that based on the presentations, the analytical validation of any given platform was fairly straightforward, but the end product of these studies (a set of genes to be used as a marker) is likely to be platform dependent, and these sets of genes will not be the same across platforms. However, he noted that identifying identical sets of genes across platforms is not essential as long as the markers are supported by sufficient biologic validation. Another participant provided an example of this concept: In a study of sets of genes indicative of breast cancer tumor metastasis, different microarray platforms indicated completely distinct sets of genes as markers of breast cancer. However, when these gene sets were mapped biologically, there was complete overlap of the pathways in which those genes were involved, thus, there was good agreement in terms of the biologic pathways. John Quackenbush also commented on the results of recent studies presented in *Nature Methods*,[21] where a variety of platforms were tested using the same biologic samples. In general, these studies indicated that although variability exists between labs and microarray platforms, and different platforms identify different biomarkers of those pathways, common biologic pathways emerge.

Linking Gene-Expression Changes to Biologic Events

Bill Mattes, of Gene Logic,[22] commented that it was necessary to

[21] May 2005, Volume 2, No. 5

[22] Dr. Mattes is currently affiliated with the Critical Path Institute

understand gene-expression changes in an appropriate biologic context, particularly if gene-expression changes are causative of a pathology or if they are just coincident (or correlative) to the pathology. Also, it is important that the time course be considered because gene-expression changes will vary through time. Overall, when the goal is understanding the etiology of a biologic effect, the gene-expression changes examined should be those preceding, not coincident to, the biologic effect of interest. Joseph DeGeorge, of Merck Pharmaceuticals, also commented on the importance of understanding the biologic context of gene-expression changes in validation efforts. For instance, if a gene-expression signal is obtained from animal tests with no associated pathology, it is important to ask if the change in mRNA expression is the first step to pathology or just an adaptive response. It is also important to know if the response is real or false or an artifact of screening a multitude of genes, and it is important to ask whether an observed response would be relevant to human risk when extrapolating from animal to humans.

Recognizing Differences in Biologic and Technical Validation

The importance of distinguishing between the vastly different needs for technical and biologic validation was also discussed by Kenneth S. Ramos. Many of the difficult concepts being addressed relate solely to biologic validation, and sufficient technical validation can be readily achieved using multiple compounds tested on a single animal strain under a single set of conditions. This would constitute the technical validation of a biomarker, even though the biology is not exactly understood. Conversely, for biologic validation, the needs are quite different, and the problems are more substantial, as has been discussed.

Geoff Patton, of EPA, commented that it was his personal opinion that when scientists seek to achieve biologic validation, they will not just rely on one method such as microarray technologies to make statements on biologic mechanisms or pathways. It is not sufficient to auto-validate within the same technique; it is necessary to look for other confirmatory information. He also stated that microarray assays provide multiple opportunities for gaining biologic insight from transcriptomics. For example, insight can be developed by evaluating events upstream of gene expression (such as the transcriptional regulatory elements and transcription factors driving gene-expression results) to solidify the understanding of co-regulated genes. Insight can also be gained by analyzing downstream events, such as the relationship to pathologic markers or the break point between adaptation and frank toxicity. Patton commented that the current understanding of modes of action can be sufficient to manipulate the biologic systems and to use other tools (such as proteomics or pathology) to demonstrate the validity of mechanistic or mode-of-action biomarkers. Thus, biologic validation will not be solely achieved with microarray technologies. To achieve biologic validation, a framework to describe how to assemble information to execute external validation of the pathway is needed.

Impact of Individual, Species, and Environmental Variability on Microarray Results

Georgia Dunstan, of Howard University, emphasized the importance of genetic variation when considering toxicogenomic results and of not extrapolating beyond the reference (for example, species and conditions) of the original experiments and platforms. The contribution of genetic variation between model systems and between individuals in response to various stressors cannot be avoided.

Leigh Anderson, of Plasma Proteome Institute, commented on the lack of microarray studies that characterize individual variation. For example, studies on the variation between inbred rat strains would be useful because understanding the interindividual variation will be critical in choosing genes as stable biomarkers. However, although that research is relatively simple, it is not being done, raising the question of how cross-species extrapolation can be discussed when the variation within a species has not been determined? It indicates the time is still early for these technologies.

Cheryl Walker, of the M.D. Anderson Cancer Center, commented on the impact of environmental variability on the stability of the genomic biomarkers. For instance, what happens to these biomarkers and signatures as you start to get away from a controlled light/dark cycle, diet, and nutritional status? Anderson replied that, at least in the field of proteomics, there are a known series of situations to be avoided (for example, animals undergoing sexual maturation and the use of proteins controlled by cage dominance). He commented that these types of variables and effects should be catalogued for microarray assays as well.

Hisham Hamedah indicated that there was an ongoing effort at the International Life Sciences Institute (ILSI) to obtain data from several companies on control animals. At the member companies of ILSI, there are different strains, different feeding regimens, and different methodologies. This effort has obtained data from about 450 chips tested on liver samples. Although the analysis is not yet complete, the goal is to identify genes with very low variance, and the results could possibly be extrapolated to other tissues or even other species.

Bill Mates, of Gene Logic, responded that in his experience, expression changes respond to these environmental variables in predictable ways. In fact, the wealth of information provided in microarray assays can permit researchers to find errors in study implementation, and he mentioned an experience where perturbations in expression profiles indicated improper animal feeding or watering.

Sarah Gerould, of the U.S. Geological Survey (USGS), commented that Dunstan's initial comments on cross-species differences were particularly important in ecotoxicology, which focuses on different kinds of fish, birds, and insects. Beyond the variety of organisms and their range of habitats, the organisms are exposed to multiple contaminants, increasing the potential difficulties in ecotoxicologic applications of these technologies.

Evaluating Low-Dose Effects of Chemicals

Jim Bus, of Dow Chemical, commented that most of the presentations during the workshop focused on screening pharmaceutical compounds in an overall attempt to avoid potential adverse outcomes as they enter the therapeutic environment. However, for the chemical manufacturing industry, the questions are of a different nature (although screening is a component) because, unlike pharmaceutical exposures, humans are not intentionally dosed, and the exposures are substantially lower. Therefore, toxicogenomic assays present an opportunity for biologic validation of effects, particularly at the low end of the dose-response curve where conventional toxicologic animal tests are insufficient. Currently, low-dose effects are addressed by, for example, 10-fold uncertainty factors or a linearized no-threshold model, but these techniques are primarily policy and are not biologically driven. Bus referred to "real world" environmental exposures that are thousands-fold lower than those assessed using conventional toxicologic models. In particular, he was interested in determining how toxicogenomics can assist in bridging the uncertainty associated with default uncertainty factors and models. These issues emphasize the need for biologic validation of these technologies and the potential for their application to the regulatory arena.

Federico Goodsaid, of FDA, brought up a related issue, commenting that conventional animal-based toxicology experiments use a limited number of samples and replicates, limiting the ability to see low-dose effects. Further, one of the most daunting tasks in using animal models to try to predict human safety issues is extrapolating from a limited sampling to what would happen with humans. In this effort, toxicogenomics perhaps represents a tool to go beyond what is currently available and to increase the power of the animal models and look at very low doses over a long period of time.

Summary Statements and Discussion

Kenneth S. Ramos moderated a summary discussion where, to initiate the discussion, he asked whether participants were comfortable with technical validation of the microarray technologies and whether it is appropriate for the field to progress to focusing on biologic validation. Indeed, similar to the roundtable session, the ensuing discussion focused on issues surrounding biologic validation, and some participants brought up themes mentioned earlier, such as the need to define mRNA expression changes that do and do not constitute a negative effect and that genetic diversity will confound extrapolation between species and among humans. Several participants also commented on the current state of validation efforts. The themes that emerged in this discussion are presented here.

Validation Issues with Microarray Assays Are Not Novel

Several participants suggested that many of the validation issues brought up throughout the day were not isolated to toxicogenomic assays. Linda Greer, of Natural Resources Defense Council, noted that in conventional animal bioassays, we often do not understand the biology underlying why, for example, animals may or may not get tumors; we do not understand the individual variation within an inbred animal strain nor how to make comparisons between species. Greer stated that she was actually relieved to hear the lack of dispute regarding technical validation of microarray technologies, because a common perception among nonspecialists is that the technology does not produce consistent results. However, she noted that technical questions have been narrowed and addressed as demonstrated by, for example, the afore-mentioned series of papers in *Nature Methods*.[23] Also, the questions regarding biologic validation apply to a lot of toxicology issues. In this regard, toxicogenomic assays are in the same state as many other conventional methodologies.

Rafael Irizarry responded to concerns that microarray assays rely heavily on “black box” mathematical preprocessing, where complex electrical and optical data is converted to a gene-expression level. He noted that other technologies also rely on mathematical algorithms or other preprocessing of raw data, for instance, RTPCR and functional magnetic resonance imaging (FMRI). However, a notable difference is that these other technologies do not produce a wealth of data like the microarray technologies, possibly explaining why the issue is rarely considered.

Bill Mattes commented on the level of consistency of microarray results generated among different labs. He stated that microarrays are like other technically demanding technologies, where everybody is not able to produce reliable data. For instance, inexperienced practitioners of histopathology via microscopy, which is considered a “gold standard” for detecting pathologic responses, will not produce reliable results. In this regard, Mattes suggested discussing the development of standards to qualify good practice. Irizarry also suggested that the scientific community use a type of internal validation of practitioners where, for example, laboratories would periodically hybridize a universal standardized reference and submit results to compare against other researchers.

Validation in What Context: Technical, Biologic, or Regulatory

Leonard Schectmann, of FDA, asked which participants had used microarray technologies and whether they were confident that the technologies had been sufficiently validated and were ready for widespread use (or “prime time” as stated by a few participants). Ramos suggested that, in this context, perhaps prime time was not the best term; rather, that these technologies had undergone sufficient validation and that it was understood they could be used to generate reliable and reproducible results. Other participants also asserted that it was necessary to understand the context in which the term *validation* was being used.

Yvonne Dragan, of FDA, said that it has been shown with microarray technologies that technical reproducibility can be achieved in the

[23] May 2005, Volume 2, No. 5

laboratory and that reproducibility across different laboratories depends on each laboratory's proficiency at the technique, with replicability being possible if the labs are proficient at the analyses. Whether reproducibility is possible across platforms depends on whether the platforms are assessing the same thing. For example, if probe sets on the microarray chip are from different locations in the same gene, then different questions are being asked. Therefore, overall, the answers depend on the level at which validation is desired. Asking whether microarrays are capable of accurately measuring mRNA levels is one question. Biologic questions are different, however, and one has to ask whether this method is the right one to address the questions being asked?

Kerry Dearfield, of USDA, commented that the question of whether a technology was ready for prime time really meant it was ready to be accepted by the regulatory agencies. In this regard, Dearfield commented that the microarray technologies were not quite there yet. Technical reproducibility, while important, does not specifically address the types of questions being asked in the regulatory field, and accurately answering the biologic questions is essential. It will be necessary to directly tie expression changes to some type of adverse end point and thus be able to address questions of regulatory interest (for example, safety or efficacy). Another application for risk assessment is when toxicogenomics will be used to examine if effects can be seen at earlier times after dosing or at lower doses. Tough questions will remain. For example, if expression changes that can be associated with a pathologic effect are seen at low doses where that pathology has not been observed, how will that information be considered in the regulatory arena? Would regulations change based on expression changes? To progress, it is necessary to ensure that the technologies are technically solid and generating reproducible, believable information. Then, that information has to be linked to biologic effects that people are concerned about. This type of technical and biologic validation needs to be tied together prior to use in the regulatory arena to address public health concerns. To get to this point, the technology will need to go through some form of internationally recognized process where, for example, performance measures for the technologies are specified, so the agencies can use the generated information.

Carol Henry, of the American Chemistry Council, commented on the potential for a group of independent researchers to recommend principles and practices necessary for the technical validation of these technologies, so the field can advance to the point where the technologies can be used in public health and environmental regulatory settings. Developing these practices could aid getting the technologies into a formal process where agencies might actually be able to accept it; right now, toxicogenomic submissions are being considered on a case-by-case basis.

Richard Canady, of the Office of Science and Technology,[24] commented that a reason to go through a validation process or certification of practitioners is to improve practice overall and reduce the odds that questionable data sets would become an accepted part of the weight of evidence for regulatory decisions (particularly for data that support a decision almost entirely). He also commented that it is important to recognize the value of data to support arguments about, for instance, dose-response extrapolation below the observable range, where it may not be possible to obtain biologically validation. In these applications, data would be used in weight-of-evidence arguments to help researchers understand the biology. Although it is good to think about validation of assays and maybe even certification of practitioners, it is bad to close the door and categorically exclude information.

Goodsaid stated that efforts were under way at FDA to develop an efficient and standard process to receive genomic information and to minimize the confusion regarding potential regulatory applications of the technologies.

Wrap-Up Discussion

To finish the workshop, John Quackenbush assembled several summary statements of themes he heard emerge from the workshop discussions and projected these for the audience (see Box 6). The statements encapsulated the technical and biologic validation considerations addressed in the speaker's presentations and the discussion that followed. Discussion on the summary statements was brief, and the workshop was adjourned.

[24] Dr. Canady is currently employed by the Food and Drug Administration.

BOX 6 Summary Statements from John Quackenbush, Harvard University

1. Technical reproducibility has been established within specific platforms.
2. Users of the technology now believe the technology is "valid" in the sense that we can extract "biology" (and possibly extrapolate to mechanistic understanding) from the data, provided that well-controlled and well-designed assays are completed.
3. Validation of the predictive algorithms and methods require sufficiently large data sets.
4. Validation in a broader context will require a focus on specific applications, and there is a need to look at dose response, time-course data, and epidemiological effects.
5. As we move to regulatory use, the best strategy is to carry out validation in an application-specific fashion.
6. Many of the classical questions derived from toxicology—cross-species extrapolation, population and environmental effects, dose effects, and so forth—apply to microarrays as well.
7. "Omics" may allow us to shed light on these questions in a more quantitative fashion than classical toxicology.
8. The only way to address validation questions is to begin to collect data and to analyze it in the proper context.

REFERENCES

Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96(12):6745-6750.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B Met. 57(1):289-300.

Boser, B.E., I.M. Guyon, and V.N. Vapnik. 1992. A training algorithm for optimal margin classifiers. Pp. 144-152 in Proceedings of the Fifth Annual International Conference on Computational Learning Theory, 27-29 July 1992, Pittsburgh, PA. Pittsburgh, PA: ACM Press.

Brown, M.P., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA 97(1):262-267.

Deng, S., T.-M. Chu, Y.K. Truong, and R. Wolfinger. 2005. Statistical methods for gene expression analysis. Computational Methods for High-throughput Genetic Analysis: Expression Profiling. Chapter 5 in Bioinformatics, Vol. 7, Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Wiley [online]. Available: http://www.wiley.com/legacy/wileychi/ggpb/aims-scope.html [accessed April 6, 2006].

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95(25):14863-14868.

Everitt, B. 1974. Cluster Analysis. London: Heinemann.

FDA (Food and Drug Administration). 2005. Guidance for Industry: Pharmacogenomic Data Submissions. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. March 2005 [online]. Available: http://www.fda.gov/cder/guidance/6400fnl.pdf [accessed April 10, 2006].

Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906-914.

Guyon, I., J. Weston, S. Barnhill, and V.N. Vapnik. 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46(1-3):389-422.

Hamadeh, H.K., P.R. Bushel, S. Jayadev, K. Martin, O. DiSorbo, S. Sieber, L. Bennett, R. Tennant, R. Stoll, J.C. Barrett, K. Blanchard, R.S. Paules, and C.A. Afshari. 2002a. Gene expression analysis reveals chemical-specific profiles. Toxicol. Sci. 67(2):219-231.

Hamadeh, H.K., P.R. Bushel, S. Jayadev, O. DiSorbo, L. Bennett, L. Li, R. Tennant, R. Stoll, J.C. Barrett, R.S. Paules, K. Blanchard, and C.A. Afshari. 2002b. Prediction of compound signature using high density gene expression profiling. Toxicol. Sci. 67(2):232-240.

Hastie, T., R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown. 2000. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. 1(2):RESEARCH0003.

Hossain, M.A., C.M.L. Bouton, J. Pevsner, and J. Laterra. 2000. Induction of vascular endothelial growth factor in human astrocytes by lead. Involment of a protein kinase C/activator protein 1 complex-dependent and hypoxia-inducible factor 1-independent signaling pathway. J. Biol. Chem. 275(36):27874-27882.

Huang, Q., R.T. Dunn, II, S. Jayadev, O. DiSorbo, F.D. Pack, S.B. Farr, R.E. Stoll, and K.T. Blanchard. 2001. Assessment of cisplatin-induced nephrotoxicity by microarray technology. Toxicol Sci. 63(2):196-207.

ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods). 2003. Guidelines for Nomination and Submission of New, Revised, and Alternative Test Methods. NIH Publication No. 03-

4508. U.S. Department of Health and Human Services, U.S. Public Health Service Research, National Institute of Health, National Institute of Environmental Health Sciences, Triangle Research Park, NC. September 2003 [online]. Available: http://iccvam.niehs.nih.gov/docs/guidelines/subguide.htm [accessed April 10, 2006].

Iida, M., C.H. Anna, W.M. Holliday, J.B. Collins, M.L. Cunningham, R.C. Sills, and T.R. Devereux. 2005. Unique patterns of gene expression changes in liver after treatment of mice for 2 weeks with different known carcinogens and non-carcinogens. Carcinogenesis. 26(3):689-699.

Irizarry, R.A., D. Warren, F. Spencer, I.F. Kim, S. Biswal, B.C. Frank, E. Gabrielson, J.G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S.C. Hilmer, E. Hoffman, A.E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S.Q. Ye, and W. Yu. 2005. Multiple-laboratory comparison of microarray platforms. Nat. Methods 2(5):329-330.

Joliffe, I.T. 1986. Principal Component Analysis. New York: Springer.

Kohonen, T. 1995. Self-Organizing Maps. Berlin: Springer.

Kramer, J.A., S.W. Curtiss, K.L. Kolaja, C.L. Alden, E.A. Blomme, W.C. Curtiss, J.C. Davila, C.J. Jackson, and R.T. Bunch. 2004. Acute molecular markers of rodent hepatic carcinogenesis identified by transcription profiling. Chem. Res. Toxicol. 17(4):463-470.

Lu, T., J. Liu, E.L. LeCluyse, Y.S. Zhou, M.L. Cheng, and M.P. Waalkes. 2001. Application of cDNA microarray to the study of arsenic-induced liver diseases in the population of Guizhou, China. Toxicol. Sci. 59(1):185-192.

Michel, C., R.A. Roberts, C. Desdouets, K.R. Isaacs, and E. Boitier. 2005. Characterization of an acute molecular marker of nongenotoxic rodent hepatocarcinogenesis by gene expression profiling in a long term clofibric acid study. Chem. Res. Toxicol. 18(4):611-618.

Nature Methods. 2005. May 2005, Volume 2, No. 5.

Schneider, J., and A.W. Moore. 1997. Cross validation. In A Locally Weighted Learning Tutorial Using Vizier 1.0. February 1, 1997 [online]. Available: http://www.cs.cmu.edu/~schneide/tut5/node42.html [accessed April 6, 2006].

Schölkopf, B., and A. Smola. 2002. Learning with Kernels. Cambridge, MA: MIT Press.

StatSoft, Inc. 2006. Electronic Statistics Textbook, Statistics Glossary. StatSoft, Inc., Tulsa, OK [online]. Available: http://www.statsoft.com/textbook/glosfra.html [accessed April 6, 2006].

Steiner, G., L. Suter, F. Boess, R. Gasser, M.C. de Vera, S. Albertini, and S. Ruepp. 2004. Discriminating different classes of toxicants by transcript profiling. Environ. Health Perspect. 112(12):1236-1248.

Thomas, R.S., D.R. Rank, S.G. Penn, G.M. Zastrow, K.R. Hayes, K. Pande, E. Glover, T. Silander, M.W. Craven, J.K. Reddy, S.B. Jovanovich, and C.A. Bradfield. 2001. Identification of toxicologically predictive gene sets using cDNA microarrays. Mol. Pharmacol. 60(6):1189-1194.

Tong, W. 2006. Decision Forest Classification Method [online]. Available: http://www.toxicoinformatics.com/DecisionForest.htm [accessed Aug. 10, 2006].

Tong, W., Q. Xie, H. Hong, L. Shi, H. Fang, R. Perkins, and E.F. Petricoin. 2004. Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: Assessing chance correlation and prediction confidence. Environ. Health Perspect. 112(16):1622-1627.

Vapnik, V.N. 1998. Statistical Learning Theory. New York: Wiley.

Yeang, C.H., S. Ramaswamy, P. Tamayo, S. Mukherjee, R.M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. 2001. Molecular classification of multiple tumor types. Bioinformatics 17(Suppl. 1):S316-S322.

Attachments

ATTACHMENT 1

EXPERIMENTAL OBJECTIVES OF DNA MICROARRAY STUDIES

Kevin K. Dobbin, Ph.D.
Biometric Research Branch
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Good gene expression microarray studies have clear objectives. These objectives will typically not be to confirm hypotheses about individual genes or pathways, because this could often be done more effectively with lower throughput assays. Instead, the hypotheses will be more general and include hundreds or thousands of genes. Having clear objectives is important for study design because no one design is best for every set of objectives, and so the choice of study design should be guided by the objectives.

Three common types of objectives in microarray studies are class comparison, class prediction and class discovery. In class comparison studies, the goal is to identify genes differentially expressed among predefined classes of samples. For example, Hossain et al. (2000) measured gene expression before and after toxic exposure to identify mechanisms of action of the toxicant, and Lu et al. (2001) compared liver biopsies from individuals in China with chronic arsenic exposure to those from healthy individuals to identify how the toxicant altered gene expression. In class prediction studies, one also has predefined classes but the goal is to develop a method for predicting class membership from gene expression data. An example of class prediction appears in Thomas et al. (2001), where a multi-gene predictor of toxic outcome was developed. In class discovery studies, one does not have predefined classes, but instead the classes are constructed during the course of the data analysis, typically by cluster analysis methods. One can cluster either the samples or the genes. Samples are typically clustered to identify structure in an otherwise homogeneous collection of samples, or to identify a novel taxonomy for the samples. An example of clustering genes in a study appears in Huang et al. (2001), where gene expression was measured at key time points after a toxic insult in order to gain insight into genetic regulation.

LEVELS OF REPLICATION

Because of the complexity of the microarray assay, replication can be performed at many different levels. But for experimental design purposes, it is important to distinguish between two general types of replicates, technical replicates and biologic replicates. Technical replicates occur when the same sample is measured multiple times. For example, one may pipette out multiple RNA samples from the same test tube and run each on a different array. Biologic replicates occur when different samples are measured on different arrays. For example, when studying populations of humans or animals, biologic replicates occur when each array is associated with a different person or animal. In cell culture experiments, the analogue to biologic replication occurs when one re-grows the cells under the same condition for each array (independent replication).

Figure 1-1 shows the effect of replication level choice on class comparison estimates. Here the number of arrays from each of two classes is fixed at 12. On the x-axis is the number of technical replicates performed on each sample; for example, a 1 corresponds to having one array performed for each of 12 samples; and a 2 corresponds to having two arrays performed for each of 6 samples. The open circles represent the accuracy of the individual sample estimates and the closed squares the accuracy of the class estimates. The figure shows that as the number of technical replicates per sample increases, so that the accuracy of each individual sample measurements gets better, the accuracy of the class estimates simultaneously gets worse. But the more accurate the class estimates are, the better the quality of the gene list. Hence the best gene list will result from the design that assigns one sample to each array with no technical replication.

Independent biologic replicates are required for valid statistical inference, because one needs some estimate of the variation in expres-

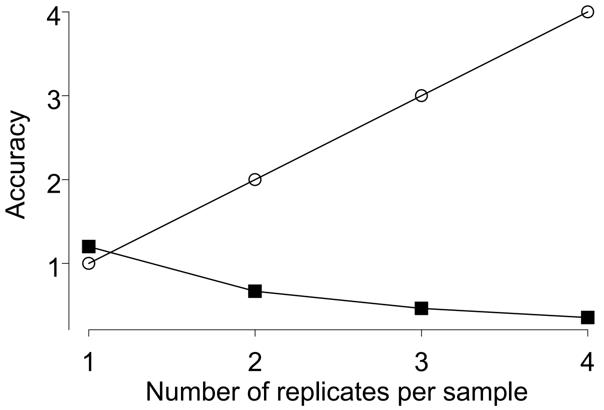


FIGURE 1-1 Accuracy (inverse of the variance of the mean estimate) of sample estimates (open circles) and class estimates (closed squared) as the number of technical replicates per sample increases from 1 to 4. Fixed number of 12 arrays per each of two class assumed. Source: Dobbin et al. 2003.

sion in the population. If, for example, one only has one sample from each of two classes, then one cannot assess the statistical significance of observed fold changes without knowing whether these are outside the range of the normal biologic variations observed in the populations. In general, the higher the level of replication the better, so that biologic replicates are preferable to technical replicates. While technical replicates can be informative in some cases, for instance for quality control, in general systematic technical replication on all samples results in poor study design. One obvious exception is when there are a limited number of samples available.

Spotted arrays typically use two dyes. This permits two expression measurements to be made at each spot. This is important because spot-to-spot variation tends to be large in spotted arrays, and may drown out the gene expression differences one is trying to detect. Having two measurements at each spot makes it possible, by designing and analyzing the study properly, to eliminate the spot variation from the gene expression comparisons.

Experimental designs that have been used for microarray studies include reference designs, balanced block designs, all pairs designs, loop designs, and variations on loop designs. For the experimental objectives under discussion here, usually a reference design or a balanced block design will be the best to use. Figure 1-2 shows an example of a reference design and of a balanced block design. The reference design uses a common reference sample which is applied to every array and is always tagged with the same dye (in the figure, this is Cy3, but it could also be Cy5). Each sample under study is tagged with the other dye and paired with one of the reference samples on an array. With this design, the expression level of each sample relative to the reference sample is measured for each gene. This permits one to also assess the expression level of any sample relative to any other sample by connecting them together using these relations to the reference. The balanced block design does not use a reference sample. With just two classes, the balanced block design pairs together one sample from each class on each array. For any class, one tags half the samples from that class with each dye.

Table 1-1 shows the relative efficiency of the balanced block design compared to the reference design for a class comparison experiment. A relative efficiency of 2.4 means that 2.4 times as many arrays are required for a reference design to equal the accuracy of a balanced block design. This is the relative efficiency for two classes, and it depends on the ratio of technical to biologic variation (Dobbin and Simon 2002). As the number of classes increases, the relative efficiency decreases, but the balanced block design will always be more efficient than the reference design.

The balanced block design does have some drawbacks. Because there is no common reference to connect all the samples, samples on different arrays cannot be compared effectively. This means, for example, that cluster analysis of the samples will perform very poorly with this type of design (it will essentially not be possible). It also means that if there is more than one way to classify the samples (for example, in cancer one can classify by tumor grade or by tumor stage), then the balanced

TABLE 1-1 Relative Efficiency of the Balanced Block (BB) Design Compared with the Reference (R) Design

Number of Classes Being Compared	Relative Efficiency (BB/R)
2	2.4
3	1.8

Reference Design

Array	Cy3	Cy5
1	R	A1
2	R	A2
3	R	A3
4	R	B1
5	R	B2
6	R	B3

Balanced Block Design

Array	Cy3	Cy5
1	A1	B1
2	B2	A2
3	A3	B3
4	B4	A4
5	A5	B5
6	B6	A6

FIGURE 1-2 Examples of a reference design and a balanced bock design. There are two classes, A and B. For example, "A1" indicates biologic sample 1 from class A. "R" indicates the reference sample, which is subsampled multiple times. Cy3 and Cy5 are dyes (green and red, respectively) used to label the samples. Source: Adapted from Dobbin et al. 2003.

block design may make some of these alternative comparisons impossible or inefficient. Another potential drawback is that data from different block design experiments cannot be effectively combined together, whereas if the same reference sample is used the data from different experiments can be combined for analysis.

For class discovery experiments in which one is classifying the samples, the reference design appears to be the best choice, although it is possible that a better design will be found. Block designs are not appropriate.

DYE BIAS IN DUAL LABEL ARRAYS

Dye bias is the tendency of some genes to fluoresce more brightly in one dye than the other. If such bias is not removed in the preprocessing and normalization of the data, then it has the potential to introduce bias into comparisons of interest. This type of dye bias can be called gene-specific dye bias and it has been shown to exist in numerous studies under a wide range of normalization and data analysis techniques (Dobbin et al. 2005).

Gene-specific dye bias cancels out of class comparisons in reference and balanced block designs. It will also not affect cluster analysis in reference design experiments if a Euclidean distance metric is used. Dye bias can affect comparisons in other types of designs, or comparisons with the reference sample in a reference design.

POOLING RNA SAMPLES

RNA samples are sometimes pooled prior to labeling and hybridization in order to either avoid RNA amplification in cases when RNA samples are insufficient for the microarray assay, or reduce the cost of the experiment by using fewer arrays. When pooling, RNA pools should be independent, so that no two pools have a sample in common.

Table 1-2 shows an example of the tradeoff between the number of arrays required and the number of samples required for a class comparison experiment when one pools RNA samples. Each row of the table represents an experiment which uses a different pooling level, and the number of arrays and sample sizes are calculated so that each experiment has the same type I and type II error rates.

As can be seen from the table, pooling may make sense when samples are cheap relative to microarrays. If the samples are valuable, then pooling will probably not make sense. One does lose information on individual samples when one pools, so, for example, human samples are rarely pooled. But it may be a viable alternative to RNA amplification when individual RNA samples are too small.

SAMPLE SIZE DETERMINATION

For class comparison studies, sample size formulas are available. In these studies, the statistical significance of the differential expression for each gene is tested using a test such as a t-test. Sample size methods for these univariate tests can be used to estimate the sample size requirements for a study. Examples of formulas for single and dual-label arrays, and various designs, appear in Dobbin and Simon (2005).

Sample size guidelines and methodologies for class prediction in microarray studies have been suggested (Mukherjeee et al. 2003; Dobbin and Simon 2007). Determining sample size for cluster analysis is more problematic.

TABLE 1-2 Example of the Tradeoff Between Number of Arrays Required and Number of Samples Required for Various Pooling Levels

Number of Samples Pooled on Each Array	Number of Arrays Required	Number of Samples Required
1	25	25
2	17	34
3	14	42
4	13	52

Source: Adapted from Dobbin and Simon 2005.

REFERENCES

Dobbin, K.K., and R.M. Simon. 2002. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18(11):1438-1445.

Dobbin, K.K., and R.M. Simon. 2005. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6(1):27-38.

Dobbin, K.K., and R.M. Simon. 2007. Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics 8(1):101-117.

Dobbin, K.K., J.H. Shih, and R.M. Simon. 2003. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. J. Natl. Cancer Inst. 95(18):1362-1369.

Dobbin, K.K., E.S. Kawasaki, D.W. Petersen, and R.M. Simon. 2005. Characterizing dye bias in microarray experiments. Bioinformatics 21(1):2430-2437.

Hossain, M.A., C.M.L. Bouton, J. Pevsner, and J. Laterra. 2000. Induction of vascular endothelial growth factor in human astrocytes by lead. Involvement of a protein kinase C/activator protein 1complex-dependent and hypoxia-inducible factor 1-independent signaling pathway. J. Biol. Chem. 275(36):27874-27882.

Huang, Q., R.T. Dunn, II, S. Jayadev, O. DiSorbo, F.D. Pack, S.B. Farr, R.E. Stoll, and K.T. Blanchard. 2001. Assessment of cisplatin-induced nephrotoxicity by microarray technology. Toxicol Sci. 63(2):196-207.

Lu, T., J. Liu, E.L. LeCluyse, Y.S. Zhou, M.L. Cheng, and M.P. Waalkes. 2001. Application of cDNA microarray to the study of arsenic-induced liver diseases in the population of Guizhou, China. Toxicol. Sci. 59(1):185-192.

Mukherjee, S., P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub, and J.P. Mesirov. 2003. Estimating dataset size requirements for classifying DNA microarray data. J. Comput. Biol. 10(2):119-142.

Thomas, R.S., D.R. Rank, S.G. Penn, G.M. Zastrow, K.R. Hayes, K. Pande, E. Glover, T. Silander, M.W. Craven, J.K. Reddy, S.B. Jovanovich, and C.A. Bradfield. 2001. Identification of toxicologically predictive gene sets using cDNA microarrays. Mol. Pharmacol. 60(6):1189-1194.

ATTACHMENT 2

COMPARISON OF MICROARRAY DATA FROM MULTIPLE LABS AND PLATFORMS

Rafael A. Irizarry
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Baltimore, Maryland

INTRODUCTION

Microarray technology is a powerful tool able to measure RNA expression for thousands of genes at once. This technology promises to be a useful tool to better understand the role of genetics in biologic responses to environmental contaminants. Various examples were presented in the 10th Meeting of the Committee on Emerging Issues and Data on Environmental Contaminants. In this extended abstract I give a short summary of a study organized to compare the three leading microarray platforms. I also give recommendations on experimental issues and statistical analyses. Many of the recommendations are applicable to general problems in technology assessment. More details are available in a Nature Methods paper (Irizarry et al. 2005) which includes various figures and tables referred to in this extended abstract.

MOTIVATION

As a statistician working in the Johns Hopkins Medical Institutions I give advice to various scientists working with microarrays. Various commercial vendors as well as custom made facilities provide many alternative platforms for researchers interested in gene-expression data. A frequently asked questions among those just getting started with this relatively new technology is “which platform performs best?” Various studies have been published comparing competing platforms with mixed results: some find agreement (Kane et al. 2000; Hughes et al. 2001; Yuen et al. 2002; Barczak et al. 2003; Carter et al. 2003; Wang et al. 2003), others do not (Kuo et al. 2002; Kothapalli et al. 2002; Li et al. 2002; Tan et al. 2003). Why the disagreement? Two possibilities were that 1) different statistical assessments were used and 2) the lab effect was not explored. That different statistical assessments can lead to different conclusions is clear and we will give an example later. To see how the lab effect can result in different conclusions, consider that in all previous studies, platform variation was confounded with technician/lab variation. In the cases where the same lab created all the microarray data it was clear that they had more experience with one of the platforms being compared. The lab effect has been shown to be particularly strong. For example, in a 1972 paper, W.J. Youden (Youden 1972) pointed out how different Physics labs published speeds of light estimates with confidence intervals that made the differences between labs statistically significant (see Figure 2-1). Notice, that if we do not take the lab effect into account this would imply that the speed of light is different in the different labs! It is no surprise that similar effects are present in biology labs and that it does not only apply to microarray measurements.

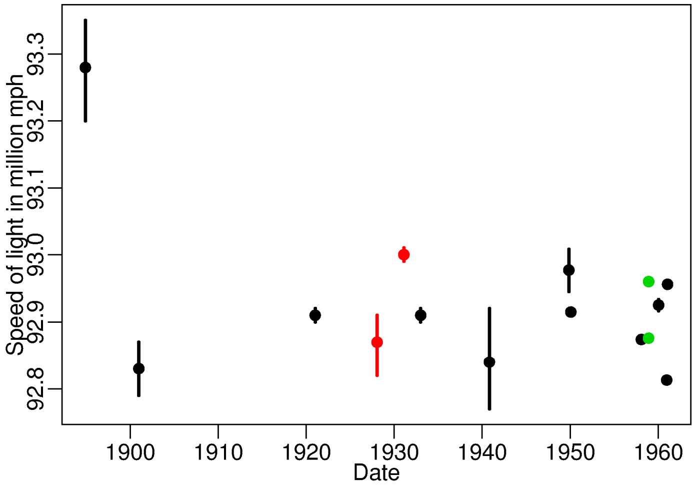


FIGURE 2-1 Speed of light estimates with confidence intervals (1900-1960). Source: Youden 1972. Reprinted with permission; copyright 1972, Journal Information for Technometrics.

Furthermore, in my opinion, previous studies had at least one of these problems: (1) precision/accuracy were not properly assessed (that is, the sensitivity/specificity trade-off was not considered and in general, assessments were based on validation of few genes). (2) There was not an a-priori expectation of truth. In general, RTPCR was considered the gold standard and measurements from that technology considered the "truth." (3) The effect of preprocessing was not explored. (4) As mentioned, the lab effect not explored.

OUR STUDY

Together with various labs from District of Columbia/Baltimore area, that volunteered their time and materials, I conducted a study for comparing microarray technologies. To overcome the problems of previous studies, we followed methodology that can be summarized by the following steps: (1) We included platforms for which results from at least two labs were available. (2) To avoid a transportation effect, we considered only labs in DC/Baltimore area. Of those we asked, five Affymetrix labs, three two-color cDNA labs, and two two-color oligo labs agreed. (3) We send each lab technical replicates of two RNA samples. (4) In the samples sent to each lab, we included technical replicates of each of the two samples. This permitted us to assess precision. (5) We designed the two RNA samples to induce a-priori knowledge of differential expression of four genes. This permitted us to assess accuracy. (6) Finally, to provide more power to the assessment of accuracy we measured fold-changes for 16 strategically chosen genes. Details are available from the *Nature Methods* publication (Irizarry et al. 2005).

STATISTICAL RECOMMENDATIONS—WHAT MEASUREMENT?

In our study we evaluated what we consider to be the basic measurement obtained from microarrays: relative expression in the form of log ratios. Thus, for each lab we had two replicate measures of relative expression: $M_1 = \log(B_1/A_1)$ and $M_2 = \log(B_2/A_2)$, with A_1, A_2, B_1, B_2 representing the two pairs of technical replicates provided to each lab.

The first important recommendation is that when comparing and/or combining measurements from different platforms one should look at relative as opposed to absolute measures of expression. This is because in microarray technology there exists a strong probe or spot effect. The *Nature Methods* publication (Irizarry et al. 2005) describes in some detail how this can lead to over-optimistic measures of precision when comparing within platform and over-pessimistic measures of agreement when comparing across platforms.

USING RELATIVE EXPRESSION INSTEAD OF ABSOLUTE EXPRESSION

Most experiments compare expression levels between different samples, thus in general this type of measure is readily available. In classification problems, it is typical to use distance measures which implicitely use relative measures. For example, the most popular methods subtract the across sample mean log expression for each gene, which produces relative (to the mean) expression measures.

PRECISION AND ACCURACY ASSESSMENTS

We believe it is important to assess precision in the context of accuracy. A platform that produces results that are perfectly reproducible (precise) is of little practical use if it fails to detect a signal (accuracy). This concept is particularly important with microarray platforms because different pre-processing methodology can lead to differences in precision and accuracy. Typically, better precision can be reached by sacrificing accuracy and vice-versa (Wu and Irizarry 2004).

PRECISION

In our study we used the following simple statistical assessments that have very intuitive interpretation in practice. To assess precision we looked at the standard deviation (SD) of $M_1 - M_2$ across all genes. This SD represents (roughly) the typical value of log-fold change when it should be 0. Correlation, which is a commonly used measure, does not have the same simple interpretation. Furthermore, in a comparison where most genes are not differentially expressed the correlation will be correctly attenuated towards 0 which might lead to the incorrect interpretation that the measurements are not reproducible. Note that measurement error should not correlate, but it should have small variance.

ACCURACY

To assess accuracy we computed the slope of M's regressed against nominal log-fold change. This represents (roughly) the expected log-fold change when it should be 1. This measure combined with the precision assessment gives us a good overall idea of the ability of the technology to detect changes among a large number of instances of no change.

CAT PLOTS

In practice, we typically screen a small subset of genes that appear to be differentially expressed. Therefore, it is more important to assess agreement for genes that are likely to pass this screen. To account for this, we introduce a new descriptive plot: the correspondence at the top (CAT) plot. This plot is useful for comparing two procedures for detecting differentially expressed genes. To create a CAT plot we form a list of *n* candidate genes for each of the two procedures and then form curves that show the proportion of genes in common plotted against the list size *n*. These figures can be seen in the *Nature Methods* publication (Irizarry et al. 2005). As an assessment measure, we can give the proportion of genes in common in lists of specific sizes, for example 25, 50, and 100.

RESULTS

Our results demonstrated that precision is comparable across platforms. With the exception of one lab, all the labs performed similarly, and it is clear that the lab effect is stronger than the platform effect. All the labs appear to give attenuated $\log_2$-fold change estimates, which is consistent with previous observations. In general, the Affymetrix labs appear to have better accuracy than the two-color platforms, although the best signal measure was attained by a two-color oligo lab. While two labs (a two-color cDNA lab and a two-color oligo lab) clearly under performed, the differences among the other eight labs do not reach statistical significance. Among the best performing labs we found a relatively good agreement of about 40% among list sizes of 100 (that is, among these labs, there was about 40% agreement on the top 100 genes with the greatest fold change in expression). It is not obvious that for variables that are highly correlated it is hard to get agreement much higher than

40%. Figure 2-2 plots the expected agreement of two bivariate normal random variables against their correlation.

DISCUSSION

Overall, the Affymetrix platform performed best. However, it is important to keep in mind that this platform is typically more expensive than the alternatives. We also demonstrated that relatively good agreement is achieved between the Affymetrix labs and the best performing two-color labs. These results contradict some previously published papers that find disagreement across platforms (Kuo et al. 2002; Kothapalli et al. 2002; Li et al. 2002; Tan et al. 2003). The conclusions reached by these studies are likely due to inappropriate statistical assessments as well as the confounding of lab effects. The existence of the sizable lab effect was ignored in all previously published comparison studies. This permits the possibility that studies using, for example, experienced technicians may find agreement and studies using less experienced technicians may find disagreement. Figure 2-3 plots the precision assessment against experience for the Affymetrix labs.

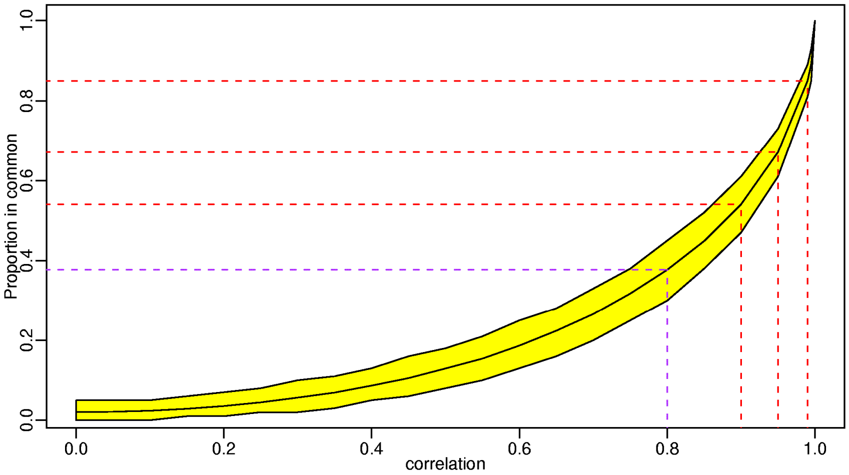


FIGURE 2-2 Expected agreement of two bivariate normal random variables against their correlation.

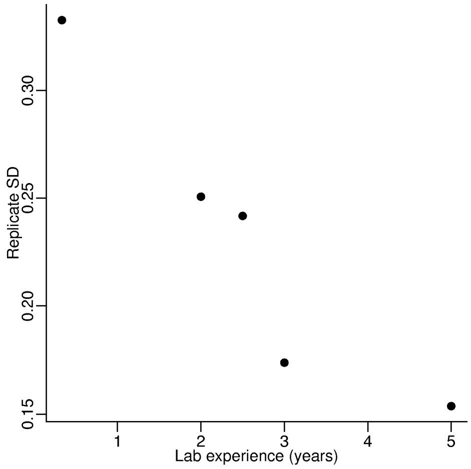


FIGURE 2-3 Precision versus years of experience for labs using Affymetrix microarrays. Source: Irizarry et al. 2005. Reprinted with permission; copyright 2005, *Nature Methods*.

Although we found relatively good across platform agreement it is quite far from being perfect. In all across-platform comparison, there was a small group of genes that had relatively large fold changes for one platform but not for the other. We conjecture that some genes are not correctly measured, not because the technologies are not performing adequately, but because transcript information and annotation can still be improved.

Our findings illustrate that improved quality assessment standards are needed. Assessments of precision based on comparisons of technical replicates appear to be standard operating procedure among, at least, academic labs. Precision and accuracy assessments are not informative unless performed simultaneously. Figure 2-4 shows observed versus nominal log-fold-changes for two labs with similar precision. Labs that perform well in terms of accuracy will show points near the diagonal line. Notice these two labs had similar precision, but Lab 3 had much better accuracy. We hope that our study serves as motivation for the creation of such standards. This will be essential for the success of microarray technology as a general measurement tool.

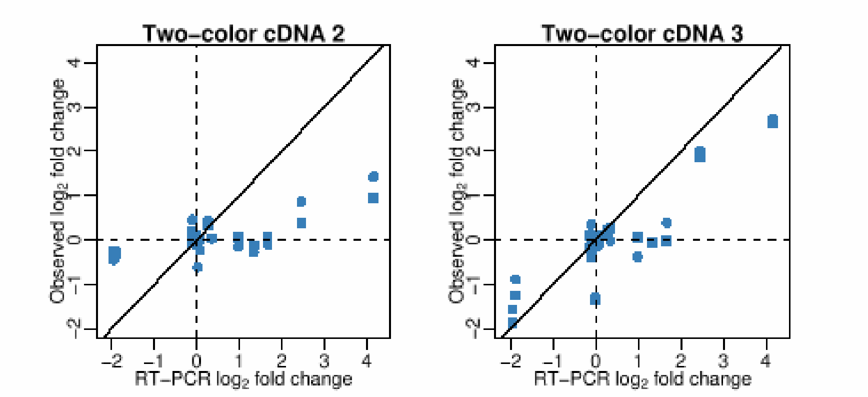


FIGURE 2-4 Observed versus nominal log-fold change for two labs (two-color cDNA 2 and two-color cDNA 3) with similar precision. Source: Irizarry et al. 2005. Reprinted with permission; copyright 2005, *Nature Methods*.

REFERENCES

Barczak, A., M.W. Rodriguez, K. Hanspers, L.L. Koth, T.C. Tai, B.M. Bolstad, T.P. Speed, and D.J. Erle. 2003. Spotted long oligonucleotide arrays for human gene expression analysis. Genome Res. 13(7):1775-1785.

Carter, M.G., T. Hamatani, A.A. Sharov, C.E. Carmack, Y. Qian, K. Aiba, N.T. Ko, D.B. Dudekula, P.M. Brzoska, S.S. Hwang, and M.S. Ko. 2003. In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling. Genome Res. 13(5):1011-1021.

Hughes, T.R., M. Mao, A.R. Jones, J. Burchard, M.J. Marton, K.W. Shannon, S.M. Lefkowitz, M. Ziman, J.M. Schelter, M.R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y.D. He, S.B. Stephaniants, G. Cavet, W.L. Walker, A. West, E. Coffey, D.D. Shoemaker, R. Stoughton, A.P. Blanchard, S.H. Friend, and P.S. Linsley. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. 19(4):342-347.

Irizarry, R.A., D. Warren, F. Spencer, I.F. Kim, S. Biswal, B.C. Frank, E. Gabrielson, J.G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S.C. Hilmer, E. Hoffman, A.E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S.Q. Ye, and W. Yu. 2005. Multiple-laboratory comparison of microarray platforms. Nat. Methods 2(5):329-330.

Kane, M.D., T.A. Jatkoe, C.R. Stumpf, J. Lu, J.D. Thomas, and S.J. Madore.

2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. 28(22):4552-4557.

Kothapalli, R., S.J. Yoder, S. Mane, and T.P. Loughran, Jr. 2002. Microarray results: How accurate are they? BMC Bioinformatics 3:22

Kuo, W.P., T.K. Jenssen, A.J. Butte, L. Ohno-Machado, and I.S. Kohane. 2002. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18(3):405-412.

Li, J., M. Pankratz, and J. Johnson. 2002. Differential gene expression patterns revealed by oligo-nucleotide versus long cDNA arrays. Toxicol. Sci. 69(2):383-390.

Tan, P.K., T.J. Downey, E.L. Spitznagel, Jr., P. Xu, D. Fu, D.S. Dimitrov, R.A. Lempicki, B.M. Raaka, and M.C. Cam. 2003. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 31(19):5676-5684.

Wang, H., R.L. Malek, A.E. Kwitek, A.S. Greene, T.V. Luu, B. Behbahani, B. Frank, J. Quackenbush, and N.H. Lee. 2003. Assessing unmodified 70-mer oligonucleotide performance on glass-slide microarrays. Genome Biol. 4(1):R5.

Wu, Z., and R.A. Irizarry. 2004. Preprocessing of oligonucleotide array data. Nat. Biotechnol. 22(6):656-658.

Youden, W. 1972. Enduring values. Technometrics 14:1-11.

Yuen, T., E. Wurmbach, R.L. Pfeffer, B.J. Ebersole, and S.C. Sealfon. 2002. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucleic Acids Res. 30(10):e48.

ATTACHMENT 3

STATISTICAL ANALYSIS OF TOXICOGENOMIC MICROARRAY DATA

Wherly Hoffman, Ph.D. and Hui-Rong Qian, Ph.D.
Statistical and Information Sciences
Eli Lilly and Company
Indianapolis, Indiana

Microarray is a relatively new technology allowing scientists to measure the expression level, or mRNA abundance of thousands of genes in a single experiment. Combined with other "omics" technologies, e.g., proteomics and metabonomics, this advance provides a completely new systems biology approach to investigate biologic problems, and makes it possible to study the huge dynamic gene interaction matrix in cells. The application of microarray in toxicity studies leads to the emergence of a new research area, toxicogenomics. Genomic studies help scientists understand the mechanism of gene function and pathway, and how a drug may alter a biologic process to treat a disease. Toxicogenomic experiments are conducted to do this with the goal of identifying the no-observed-effect level (NOEL) or no-observed-adverse-effect level (NOAEL). Since it may take a long time to develop an adverse effect while the gene expression change is relatively immediate, toxicogenomics studies have the potential for an early and accurate detection of toxic effects. Different technologies have emerged in this research area. In particular, Affymetrix's GeneChip has been widely adopted in the drug discovery process in the pharmaceutical industry. Moving from making a scientific conjecture of a compound's toxicity to analyzing gene expression profiles to concluding the association between the toxicity and gene alteration is a complex process. Statistical considerations are vital in each step of the process.

This presentation is focused on the statistical analysis of Affymetrix GeneChip. microarray data. The Affymetrix GeneChip technology, statistical hypotheses, and statistical analysis methods will be discussed in order. Visualization of the data and analysis results will be included. Conventional hypothesis testing on such a massive amount of data leads to severe multiplicity issues. Proper multiplicity adjustments for *P* values will be discussed to help distill massive amounts of information into useful information for each compound.

AFFYMETRIX GENECHIP TECHNOLOGY

On an Affymetrix GeneChip, each gene (probe set) is represented by a set (11-20 pairs) of paired short oligonucleotides of 25-base long, called perfect match (PM) and mismatch (MM) oligos. PM oligos match the gene sequence exactly so after hybridization with labeled sample RNA, they reflect the expression signal, MMs have the same sequences as PMs except that the middle base is changed to its complementary nucleotide. MMs are designed to capture the non-specific hybridized signals, or background signals. There are dozens of algorithms to extract a robust signal intensity from these 11-20 pairs of PMs and MMs for each probeset (Cope et al. 2004). The three most commonly used are MAS 5 from Affymetrix (Affymetrix 2002), the robust multi-array average (RMA) by Irizarry et al. (2003), and the model-based expression index (MBEI) by Li and Wong (2001). It is still not settled as to which method is the best. The final choice is often up to the researcher's personal preference. In this presentation, signals extracted using MAS 5 from Affymetrix are statistically analyzed.

STATISTICAL HYPOTHESES

Every experiment is designed to answer certain scientific questions. It is important that before conducting an experiment, the researchers define scientific questions and statisticians translate the scientific questions into statistical hypotheses and determine appropriate statistical analyses. It is also important to make sure that at the end of the experiment, appropriate and right amounts of data have been collected for statistical analyses and for answering the scientific questions. Considering the expensive price tag of microarray chips and the large amount of time and other resources needed to carry out these experiments, it would be unwise to have problematic designs that could not provide answers to the scientific questions. In most microarray experiments, the primary goal is to identify differentially expressed genes under different treatment conditions. We will examine important sources of variation and discuss some exploratory and inferential analyses. We will provide some examples to show how different scientific questions lead to different experimental designs and statistical hypotheses.

Microarray measures the expression abundance of essentially all the genes in a genome. With thousands of measurements and relatively few subjects (tens, rarely over a hundred), any difference in the conditions of the subjects to be tested would cause a large number of genes to show expression changes. Therefore, it is very important to understand and control different sources of variability in microarray experiments. Typically there are two sources of random variability: biologic variation and technical variation. The biologic variation exists in the tested subjects. Sometimes it is possible to reduce the biologic variation, e.g., by using more homogenous individuals. Technical variation lies in the sample preparation and the microarray technology itself, including tissue collecting, RNA isolation, labeling, chip hybridization, etc. As the microarray technology matures, the chip-to-chip variation decreases. Still, large variation is observed during the sample preparation. For example, different labeling kits could lead to a big difference in the expression signals. Even with the same labeling kit, samples processed on different days may yield very different signals.

A related issue is the pooling of RNA samples from animals. During the early discovery phase when resources are more limited, pooling of the samples from animals in the same group may be necessary. An example of this is for establishing a surrogate assay to screen peroxisomal proliferation activated receptors (PPAR). While pooling samples allows researchers to use fewer chips, it loses the ability to measure individual expression and provide the estimation of biologic variation for proper statistical tests. It may be advantageous if samples are very cheap and easy to obtain (like in cell cultures), and a partial pooling is conducted in which samples from the same treatment are pooled to form multiple independent pools; thus the biologic variation can still be properly estimated. In general, for follow-up evaluation of observed toxicity, RNA samples from animals are not pooled. For example, upon observing heart weight changes or skeletal muscle necrosis in rats, microarray technology is applied to help find biomarkers and one genechip is used for each animal.

Another important issue is to avoid potential bias by randomization and/or proper blocking. For example, the processing batch effect may be significant. Two samples can appear very different if processed on two different days, especially if it involves a change of reagents. It has been consistently shown that samples, even without any treatment, are different if the samples are collected at different times. This difference can be well detected by microarray technology.

EXPLORATORY DATA ANALYSIS BY VISUALIZATION

After a microarray experiment is done and data are generated, it is important to explore the data using some basic analytical and graphical tools to detect obvious patterns, potential outliers, errors during the experiments, etc. The following analysis and visualization approaches serve these purposes well:

(1) Side-by-side box plot of log signals of each chip. It is helpful to check the chip signal distribution and to see if normalization has been applied.

(2) Pairwise chip signal correlation. A very low correlation coefficient (CC) of one chip with other chips is an indication of a quality problem of this chip. High average CC within a treatment group and low average CC between treatment groups indicate a large treatment effect in an experiment. However, the variances of different genes are different. In the Affymetrix platform, a large signal intensity corresponds to a large variance on a raw scale, and thus has a high influence on the CC calculation. On a log scale, a large signal intensity corresponds to a small variance. The cube root transformation could be applied to stabilize the variance.

(3) MVA plot of a pair of chips or average signals of treatment groups. This is a scatter plot of average log differences of a pair of chips vs. the average mean of their log signals. This would show clearly the dependence of the variance on signal intensity. It is also helpful to see if the data scatter around 0, a good approach to check if extra normalization is needed.

(4) Principal components analysis (PCA). PCA helps reduce the data dimension and reveal the general pattern in data. A clear departure of a chip from other chips with the same treatment on a PCA plot indicates the existence of outliers. PCA plots may also reveal expression patterns that are due to unrealized blocking effects, arising from the sample preparation or array processing, the bias to eliminate during analysis.

Statistical Analysis and Visualization of Results

Although in the beginning of microarray technology, people assumed that a single model for expression could be applied to all genes on an array, it is understood now that the expression of each gene is different and should be modeled on a gene-by-gene basis. The analysis of microarray data depends on the questions we want to answer and the experimental design. To identify differentially expressed genes, in a two-group comparison, a simple t-test is usually applied. Considering the high variability when expression signal is small, it has been proposed to add an inflation factor to stabilize the variance (Tusher et al. 2001), or use a regularized t-test to minimize the dependence of variance on signal intensity (Baldi and Long, 2001). When the experiment has more than two treatment groups, an ANOVA t-test may be applied to better estimate the variability and lead to more powerful tests. In a more complicated situation, e.g., with technical replicates, or with repeated measurements, a mixed effects model may be used (Chu et al. 2002).

Sometimes in an experiment, we have multiple time points or multiple doses, and the interest is to investigate the temporal expression pattern or dose response. We may fit the data to a linear regression model to measure the general expression trend. We will present some examples to show different analysis approaches for different data. In general, when the test is applied to each gene, the statistical approach is not different from what has been used in classic biologic research. However, after the initial statistical test, since it is "fishing" significant gene expression changes out of thousands of genes, adjustments should be made to the resulting *P* values to control the false positive rate.

In a typical microarray experiment, usually there are hundreds or even thousands of genes being identified as significant. Some visualization tools to show the general patterns of analysis results are listed here:

(1) MVA plot with highlighted significant genes. As described above, this plot is helpful to show the signal intensities of significant genes, and to check if the data are properly normalized.

(2) Volcano plot. This is the plot of *P* values or adjusted *P* values with fold changes, showing if the significant genes are equally distributed over up- or down-regulation.

(3) Clustering of genes. There are different ways of clustering over different distance matrices. The general purpose is to group genes with similar expression patterns together. Physiological

data if available can be included in the clustering to identify genes closely correlated with the phenotypes.

(4) Heat map (see Figure 3-1). When there are a large number of genes that are significant, after clustering, we may rescale the gene expression data within each gene and use a heat map to show the expression patterns.

(5) For individual genes of interest, the expression signals of each sample may be plotted, overlaid with any physiological data of interest if available. This helps researchers decide for a particular gene if outliers (or non-responders) exist, if modeling approaches are appropriate, etc.

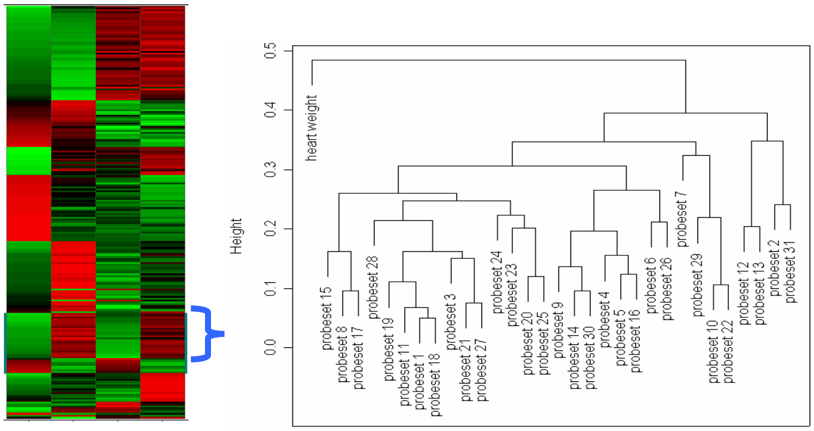


FIGURE 3-1 Heat map and gene cluster with physiological data. The cluster shows only the highlighted portion of the heat map.

MULTIPLICITY ADJUSTMENT

The power of microarray to monitor the expression of thousands of genes in one experiment enables the researchers to investigate the gene function systematically at the genome level. However, testing thousands of genes in a single experiment introduces a high rate of false positives when no efforts are taken to control the false positive rate. For example, the new human chip U133plus2 contains 54,000 probesets. If there is no treatment effect for a comparison so that all the *P* values are from a uniform distribution (0, 1), then there would be about 2,700 significant probesets ($P < 0.05$).

Bonferroni is one approach to control the false positive rate by controlling the family-wise error rate (FWER). Due to its conservativeness, when applied in microarray data analysis, it leads to too few rejections, i.e., the test power is too low to detect real useful information about gene expression changes. Some of the most recent progress in addressing multiple testing problems is the introduction of false discovery rate (FDR) by Benjamini and Hochberg (1995). FDR is defined as the expected rate of erroneous rejection of hypotheses among total rejected hypotheses. Let M be the total hypotheses being tested, R be the number of rejected null hypotheses, and V the number of falsely rejected true null hypotheses, then

$$\mathrm{FDR} = \begin{cases} E(\frac{V}{R}) \text{ if } R > 0 \\ 0 \text{ if } R = 0 \end{cases}$$

With this setting, the conventional per comparison error rate (PCER) and FWER are

$$\mathrm{PCER} = E(\frac{V}{M}) \text{ and } \mathrm{FWER} = P(V \geq 1), \text{ respectively.}$$

Instead of controlling the FWER, procedures controlling FDR control the error rate of false rejections in the rejected hypotheses. Under complete null hypotheses (i.e., all the null hypotheses are true), FDR and FWER are the same. When some of the null hypotheses are not true, FDR is smaller than FWER, thus FDR-controlling procedures are more powerful than FWER-controlling procedures (but note they control two error rates). In microarray experiments, the researchers try to identify genes showing differential expressions across treatment groups. They are more interested in how many genes identified in a microarray experiment will fail to be confirmed in the follow-up studies, i.e., the false-positive rate among the "discoveries." This provides a perfect setting to apply FDR-controlling methods.

When addressing the multiplicity issue and deciding what gene expression changes are significant using FDR or other approaches, the researchers assume that all gene expressions are independent and equally important. These assumptions are typically not true. A list of genes of interest may be available before the microarray experiment is conducted. What FDR cutoff values to use to pick follow-up genes and how to apply the FDR adjustment will be discussed.

REFERENCES

Affymetrix, Inc. 2002. Statistical Algorithms Description Document [online]. Available: http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf [accessed April 10, 2006].

Baldi, P., and A.D. Long. 2001. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics 17(6):509-519.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B Met. 57(1):289-300.

Chu, T.M., B. Weir, and R.D. Wolfinger. 2002. A systematic statistical linear modeling approach to oligonucleotide array experiments. Math. Biosci. 176(1):35-51.

Cope, L.M., R.A. Irizarry, H.A. Jaffee, Z. Wu, and T.S. Speed. 2004. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 20(2):323-331.

Irizarry, R.A., B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249-264.

Li, C., and W.H. Wong. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA 98(1):31-36.

Tusher, V.G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA 98(18):5116–5121.

ATTACHMENT 4

DIAGNOSTIC CLASSIFIER—GAINING CONFIDENCE THROUGH VALIDATION

Weida Tong, Ph.D.
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arizona

INTRODUCTION

In clinical settings, accurate diagnosis and prognosis relies mainly on histopathology, cytomorphology, or immunophenotyping. Unfortunately, some diseases are hard to classify by current clinical techniques. To overcome the inherent limitations of traditional methods of diagnosis/prognosis, much attention has been recently placed on use of molecular profiles (e.g., gene expression patterns) derived from omics experiments for clinical application, such as DNA microarray and surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS). It has been expected that recent technological advances in the fields of genomics, proteomics and other omics will offer a unique opportunity for not only improving diagnostic classification, treatment selection and prognostic assessment but also for understanding the molecular basis of health and disease. Classification methods, because of their power to unravel patterns in biologically complex data, have become one of the most important bioinformatics approaches investigated for use with omics data in clinical application. A number of classification methods have been applied to microarray gene expression data as well as other omics data. This presentation discusses the issues and challenges associated with application of classification using supervised learning methods on omics data applied to clinic. Specifically, a novel tree-based consensus method, decision forest (DF), will be discussed, which has been successfully used to develop diagnostic classifiers based on gene expression patterns, SELDI-TOF MS data, and SNPs (single nucleotide polymorphisms) profiles in a case-control study. While the general procedure to validate a classifier will be discussed, the emphasis is manly placed on assessing prediction confidence and chance correlation, two critical aspects that, unfortunately, have not been extensively discussed in the field.

ISSUES, CHALLENGES, AND RECOMMENDATIONS

Developing classifiers from omics data is difficult because (1) there are many more predictor variables than the sample size (the number of subjects); (2) the sample size is often small with a skewed patient/healthy individual distribution; and (3) the signal/noise ratio in both clinical outcomes (dependent variables) and omics profiles (independent variables) are low.

Most molecular classification approaches reported in the literature have focused on developing and validating a single classifier. Although many successes have been demonstrated, the single classifier approach is inherently susceptible to the data quality and size; as the sample size and/or the signal/noise ratio of a data set decrease, the quality of a single classifier declines rapidly. Another aspect that is unique, or at least very significant, to molecular classification is that redundant information is normally present in an omics profile. The nature of the data reflects biologic phenomena where multiple molecular expression patterns are often equally important as biomarkers in diagnosis/prognosis. Unfortunately, a single classifier tends to optimize a single pattern for classification.

Consensus modeling, that combines multiple classifiers to reach a consensus conclusion, is theoretically less prone to data quality and size and more robust to handle an unbalanced data set. Most importantly, consensus modeling makes full use of the redundant information presented in omics data to explore all possible biomarkers. Thus, consensus modeling offers a unique opportunity in molecular classification.

The critical and implicit assumption in consensus modeling is that multiple classifiers will effectively identify and encode more aspects of the variable relationships than will a single classifier. The corollaries are that combining several identical classifiers produces no gain, and benefits of combining can only be realized if individual classifiers give *different results*. In other words, benefits of combining are only expected if separate classifiers encode differing aspects of disease-omics pattern associations. More recently, we also found that the information gained from combining classifiers is valuable in assessing prediction confidence, which is usually difficult to obtain from a single classifier.

DECISION FOREST—A ROBUST CONSENSUS METHOD FOR DIAGNOSTIC CLASSIFICATION

Most consensus modeling relies on resampling approaches that use only a portion of the subjects for constructing the individual classifiers. Since we normally have a relatively small sample size, this approach will weaken individual classifiers' predictive accuracy, which follows the reduction of the improvement in a combining system gained by the resampling approach.

A preferable consensus approach is to develop multiple classifiers using different sets of omics patterns. This approach takes full advantage of the available sample size as well as the redundant information presented in the data. Accordingly, we have developed the robust DF method (see Figure 4-1).

DF emphasizes the combining of *heterogeneous* yet *comparable* trees in order to better capture the association of omics profiles and disease outcomes. The heterogeneity requirement assures that each tree uniquely contributes to the combined prediction; whereas the quality comparability requirement assures that each tree equally contributes to the combined prediction. Since a certain degree of noise is always present in biologic data, optimizing a tree inherently risks over fitting the noise. DF attempts to minimize over fitting by maximizing the difference among individual trees.

There are three benefits associated with DF compared with other similar consensus modeling methods: (1) since the difference in individual trees is maximized, the best ensemble is usually realized by combining only a few trees (i.e., four or five), which consequentially reduces computational expense; (2) since DF is entirely reproducible, the disease-patterns associations are constant in their interpretability for biologic relevance; and (3) since all subjects are included in individual tree development, the information in the original data set is fully appreciated in the combining process.

For example, we develop a DF classifier on a proteomic data set to distinguish the prostate patients from healthy individuals. The data set consists of 326 samples, of which 167 samples are from the prostate cancer patients and the noncancer group contains 159 samples including both benign prostatic hyperplasia patients and healthy individuals. The

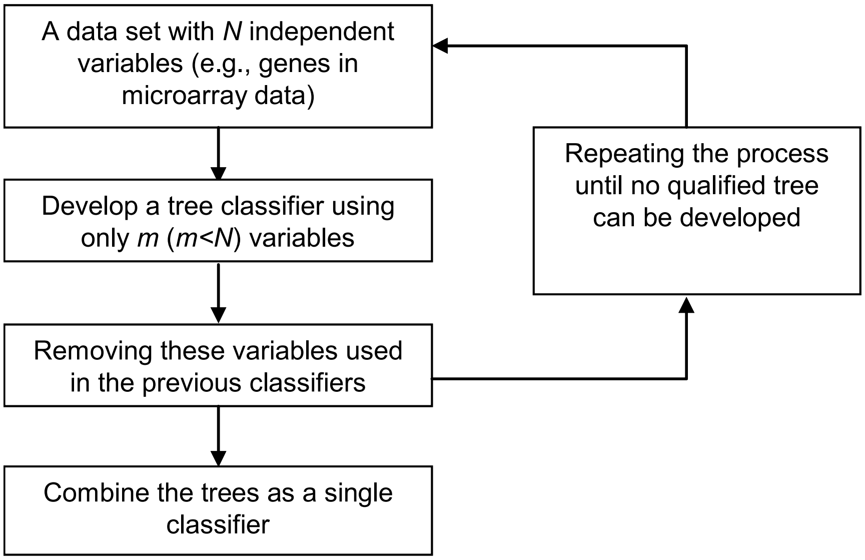

A data set with *N* independent variables (e.g., genes in microarray data)
Develop a tree classifier using only *m* (*m*<*N*) variables
Removing these variables used in the previous classifiers
Combine the trees as a single classifier
Repeating the process until no qualified tree can be developed

FIGURE 4-1 Overview of decision forest (DF). The individual tree classifiers are developed sequentially, where each tree uses a distinct set of variables. Classification (i.e., prediction) of an unknown subject is based on the mean results of all trees. Source: Tong et al. 2006. Reprinted with permission; copyright 2006, *Toxicology Mechanisms and Methods*.

classifier contains four trees, each of them having the comparable misclassifications ranging from 12 to 14 (3.7-4.3% error rate). The misclassification is significantly reduced as the number of trees to be combined increases to form a DF classifier (Figure 4-2). The four-tree DF classifier gave 100% classification accuracy.

VALIDATION 1: CROSS-VALIDATION VERSUS EXTERNAL VALIDATION

A classifier's predictive capability is commonly demonstrated using either external validation or cross-validation procedures. Although both procedures share many common features in principle, they are different in both ability and efficiency in assessing a classifier's overall prediction accuracy, applicability domain, and chance correlation during implementation and execution.

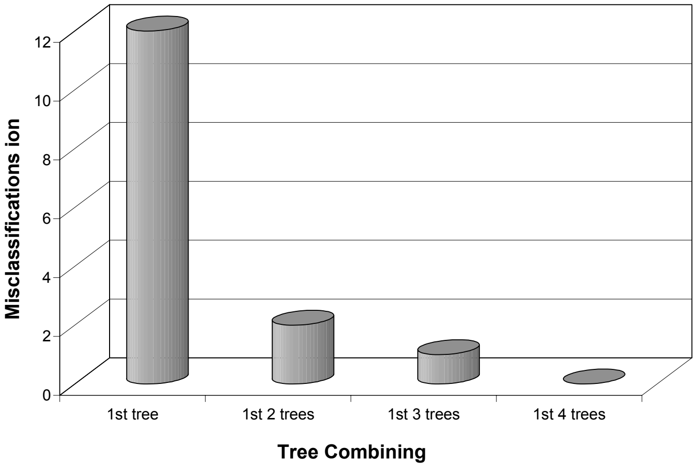


FIGURE 4-2 Plot of misclassifications versus the number of tree classifiers to be combined in DF. Source: Tong et al. 2004.

When sufficient subjects are available, a classifier should be validated by predicting subjects not used in the training set, but whose diagnostic outcomes are known (the test set). This external validation method lacks validity unless the test set is sufficiently large. Using a small number of subjects is inadequate for validation and also possibly wastes valuable data that otherwise could improve the overall quality of a classifier. Unfortunately, there is no consensus on how many subjects should be set aside to provide a valid validation.

A common practice for defining a test set in external validation is to randomly select a portion of subjects from an original population. From this perspective, cross-validation provides a similar validation for a given and fixed set of subjects. The 10-fold cross-validation procedure is commonly used to assess the predictive capability of a classifier. By comparing with external validation, cross-validation could provide a systematic measurement of a classifier's performance without the loss of subjects set aside for testing.

It is necessary to point out that the cross-validation results vary for each run due to random partitioning of the data set, and thus we recommended repeating the cross-validation process many times (complete or extensive cross-validation). The average result of the multiple cross-validation runs provides an unbiased assessment of a classifier's predictivity.

It is worthwhile to mention that classifier development and variable selection are integral in DF. Thus, DF avoids the selection bias during cross-validation that thereby provides a realistic assessment of the predictivity of a DF classifier. It is our experience that, unless the sample size is fairly large, given a constant set of subjects (a fixed data set with limited size), the cross-validation is more powerful for DF in measuring a classifier's performance than external validation with respect to assessing overall prediction accuracy, prediction confidence and chance correlation.

VALIDATION 2: ASSESSING PREDICTION CONFIDENCE

A molecular classifier is a product of a mathematic correlation between dependent and independent variables. The classifier's ability to predict an unknown sample is directly dependent on the nature of the training set. In other words, predictive accuracy for different subjects varies according to how well the training set represents the given samples. Thus, the concept of "prediction confidence" is viewed for measures of confidence in each prediction when the overall quality of a classifier is acceptable. It is critical to be able to estimate the degree of confidence for each prediction. The ability to quantify confidence greatly enhances the utility of any diagnostic classifier (to determine how best to apply the classifier).

External validation generally provides only overall quality assessment of a classifier with no indication of the confidence in individual prediction. In other words, external validation is of little value for assessing the prediction confidence which, in turn, can be readily available from cross-validation. In DF, the information derived from many runs of 10-fold cross-validation permits assessment of the prediction confidence.

Figure 4-3 gives an example illustrating how prediction accuracy and prediction confidence are related in DF. Prediction accuracy is plotted versus prediction confidence for both DF and decision tree (DT) for a problem using 2,000 runs of 10-fold cross-validation for a molecular classifier to predict prostate cancer from SELDI-TOF MS data (the same data set shown in Figure 4-2). A strong trend of increasing accuracy with increasing confidence is readily apparent for both DF and DT, as is the substantially higher accuracy for DF across almost the entire range of confidence levels. If we define high confidence predictions as those with confidence level >0.4 (Figure 4-3) and low confidence predictions as those with confidence level <0.4, we found that the high confidence prediction accuracy is ~99%, ~20% higher than the low confidence predic-

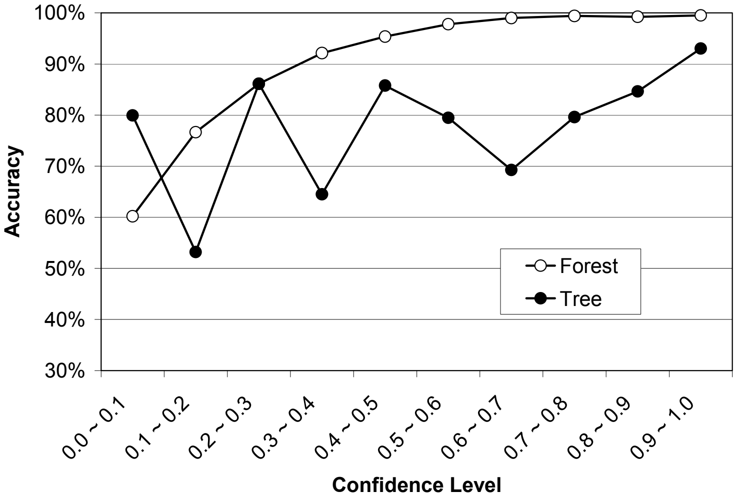


FIGURE 4-3 Prediction accuracy versus confidence level for a DF classifier of a proteomics data based on 2,000 runs of 10-fold cross-validation. The confidence level is defined as |*Pi* -0.5|/0.5, where *Pi* is the probability value for sample *i*. Source: Tong et al. 2004.

tion accuracy (~79%). The results demonstrate that the DF classifier gives a better assessment of prediction confidence than does the single tree classifier.

VALIDATION 3: DETERMINING CHANCE CORRELATION

Testing whether a classifier is, in fact, a chance correlation is highly recommended. Testing becomes increasingly imperative for smaller training sets with increasing numbers of independent variables, noise in biologic data, and unbalanced distribution of patient versus healthy individuals. All of these conditions increase the omnipresent risk of obtaining a chance correlation lacking predictive value.

Chance correlation is difficult to assess using external validation and is best obtained from cross-validation. To assess a degree of chance correlation for a DF classifier, we normally generate many pseudo data sets (e.g., 2,000 pseudo data sets) first using a randomization test, where the subject classification is randomly scrambled. Next, we apply a 10-fold cross-validation on each of pseudo data sets to generate a null distribution, i.e., the distribution of prediction accuracy from all classifiers developed on all pseudo data sets. The null distribution can then be compared with the distribution of multiple 10-fold cross-validation results derived from the real data set. The degree of chance correlation in prediction can be estimated from the overlap of the two distributions.

Figure 4-4 shows the results of a test for chance correlation of a DF classifier to predict the prostate cancer. The distribution of prediction accuracy of the real data set centers around 95% while the pseudo data sets are near 50%. The real data set has a much narrower distribution compared to the pseudo data sets, indicating that the classifiers generated from the cross-validation procedure for the real data set give consistent and high prediction accuracy. In contrast, as expected, the prediction results for the pseudo data sets varied widely, implying a large variability of signal/noise ratio across these pseudo classifiers. Importantly, there is no overlap between two distributions, indicating that a statistically and biologically relevant DF classifier can be obtained using the real data set.

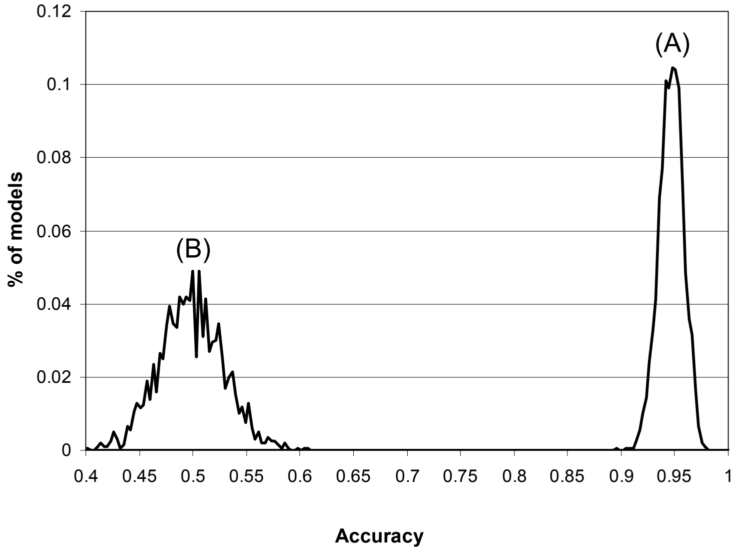


FIGURE 4-4 Prediction distribution in 2,000 runs of 10-fold cross-validation process: (A) real data set and (B) 2,000 pseudo data set generated from a randomization test. Source: Tong et al. 2004.

CONCLUSIONS

Diagnostic classification based on omics data presents challenges for most conventional supervised learning methods. Validation is a vital step towards the practical use of diagnostic classifiers. A classifier should be validated from three perspectives of assessment: (1) overall quality (prediction accuracy, sensitivity, and specificity); (2) prediction confidence; and (3) chance correlation. These can be more readily assessed in a consensus method, such as DF, than in other conventional methods using cross-validation. For DF, we have found that

- Combining multiple valid tree classifiers that use unique sets of variables into a single decision function produces a higher quality classifier than individual trees.
- The prediction confidence can be readily calculated.
- Since the feature selection and classifier development are integrated in DF, cross-validation avoids selection bias and become a more useful means than external validation in assessing a DF classifier's robustness and quality.
- Carrying out many runs of cross-validation is computationally inexpensive, which provides an unbiased assessment of a classifier's predictive capability, prediction confidence and chance correlation and facilitates identification of potential biomarkers.

REFERENCES

Tong, W., Q. Xie, H. Hong, H. Fang, L. Shi, R. Perkins, and E. Petricoin. 2004. Using Decision Forest to classy prostate samples based on SELDI-TOF MS data: Assessing prediction confidence and chance correlation. Environ. Health Perspect. 112(16):1622-1627.

Tong, W., H. Fang, Q. Xie, H. Hong, L. Shi, R. Perkins, U. Scherf, F. Goodsaid, and F. Frueh. 2006. Gaining confidence on molecular classification through consensus modeling and validation. Toxicol. Mech. Method. 16(2-3):59-68.

APPENDIX A

WORKSHOP PLANNING COMMITTEE BIOGRAPHICAL INFORMATION

John Quackenbush *(Co-Chair)* is professor of biostatistics and computational biology at the Dana-Farber Cancer Institute and professor of computational biology and bioinformatics at the Harvard School of Public Health. Until early 2005, he was an Investigator at the Institute for Genomic Research in Rockville, MD. His primary research areas are functional genomics and bioinformatics, and his work has focused on the integration of diverse data types to provide insight into biological systems. He and his group have been investigating gene expression patterns in animal models with the goal of identifying mechanisms underlying a range of human diseases. They have also used microarrays to look for diagnostic and prognostic expression fingerprints in human breast and colon cancer, and he has been active in using plant models to develop methods for integrating functional genomics and metabolomics approaches. Dr. Quackenbush has a Ph.D. in theoretical particle physics from the University of California, Los Angeles.

Kenneth S. Ramos *(Co-Chair)* is professor and chair of the Department of Biochemistry and Molecular Biology at the University of Louisville Health Sciences Center. He also serves as director of the Center for Genetics and Molecular Medicine. His research focuses on the study of molecular mechanisms of environmental disease and redox-regulated transcriptional control. He is the editor of the NIEHS journal *Environmental Health Perspectives: Toxicogenomics*. Dr. Ramos has served on numerous NRC committees including the Committee on Emerging Issues and Data on Environmental Contaminants, Committee for a Review of Evidence Regarding Link between Exposure to Agent Orange and Diabetes, Committee to Review the Health Effects in Vietnam Veterans of Exposure to Herbicides: Second Biennial Update, HHMI Predoctoral Fellowships Panel on Neurosciences and Physiology, and Committee to Review the Health Effects in Vietnam Veterans of Exposure to Herbicides: First Biennial Update. He received a Ph.D. in biochemical pharmacology and toxicology from the University of Texas at Austin.

Cynthia A. Afshari is the associate director of the Toxicology Department at Amgen, Inc. Her expertise is in the areas of molecular toxicology, functional genomics/toxicogenomics, in vitro models, and carcinogenesis. At Amgen she leads the Investigative and In Vitro Screening Toxicology groups and conducts research in new predictive toxicity assays. She is also responsible for guiding preclinical safety assessment work for several therapeutic project teams at Amgen. Previously, she was an adjunct professor of toxicology at the University of North Carolina, Chapel Hill, and was the director of basic research applications at the NIEHS Microarray Center for 4 years. At NIEHS she headed an interdisciplinary group of biologists, engineers, and computer scientists investigating applications of new genomics technologies to mechanistic toxicology. Dr. Afshari serves as the chair of the Steering Committee of the Subcommittee on Application of Genomics and Proteomics to Mechanism-based Risk Assessment organized by the Health and Environmental Science Institute (HESI) of the International Life Sciences Institute (ILSI) and was chair of the nephrotoxicity and database working groups of the same ILSI-HESI subcommittee for 2 years. She is also a member of the Pharmaceutical Research and Manufacturers of America (PhRMA) Genomics Subcommittee and is an associate editor of *Toxicologic Pathology, Toxicological Sciences*, and a reviewing editor for *Environmental Health Perspectives*. She earned her Ph.D. in toxicology from the University of North Carolina, Chapel Hill, and is a board-certified toxicologist.

Linda E. Greer is senior scientist for the Natural Resources Defense Council (NRDC) and the director of its health program. She received a Ph.D. in environmental toxicology from the University of Maryland. Dr. Greer's primary focus concerns toxic chemical pollution regulatory issues and risk assessment. She is currently focusing particularly on mercury pollution. Dr. Greer has served on numerous NRC committees, including the Committee on Industrial Competitiveness and Environmental Protection, the Committee on Ground Water Cleanup Alternatives, and the Committee on Hazardous Wastes in Highway Rights-of-Way. Dr. Greer also was a member of the NRC's Board on Life Sciences until 2004.

Casimir A. Kulikowski is Board of Governors Professor of Computer Science at Rutgers University. He earned a Ph.D. in electrical engineering from the University of Hawaii. His research interests include bioinformatics, medical informatics, and artificial intelligence. In bioinformatics, he is working on pattern recognition and clustering methods for genomics and proteomics, and consensus classification methods for comparative genomic analysis and annotation. Dr. Kulikowski is also working on models for clinical guidelines, on methods of medical decision support, biomedical imaging, and predictive data mining. He has served on the NRC Committee to Address "Information Infrastructure for Health and Health Care" and the Committee to Review the Social Security Administration's System Modernization and Strategic Plan. Dr. Kulikowski was elected to the Institute of Medicine in 1988.

George Orphanides is head of investigative toxicology and Stage 1 toxicology at Syngenta in the United Kingdom. Previously he has held roles as head of receptor biology and genomics at Syngenta and group leader for toxicogenomics at AstraZeneca. His research interests in the area of toxicogenomics include gene expression profiling, proteomics, mechanisms of gene regulation, transcription factors, nuclear receptors, chromatin dynamics, estrogenic compounds, and bioinformatics. He currently serves on several committees, including vice-chair of the ILSI-HESI Committee on the Application of Genomics in Risk Assessment; the ECVAM (European Centre for the Validation of Alternative Test Methods) Committee on the Validation of Toxicogenomics-Based Alternative Methods; and the IPCS (International Programme on Chemical Safety) Committee on Toxicogenomics and the Risk Assessment of Chemicals in the Protection of Human Health, where he is co-chair of the Subcommittee on Human Susceptibility and Exposure. He currently serves on the editorial board of the journal Biomarkers. Dr. Orphanides received his Ph.D. in biochemistry from the University of Leicester, United Kingdom.

Lawrence M. Sung holds the appointment of law school professor and director of the Intellectual Property Law Program at the University of Maryland Law School. He is also a partner with the Washington, DC intellectual property law firm of Schwartz, Sung & Webster. His area of expertise is in patent and technology transfer issues concerning the biotechnology, pharmaceutical and medical device industries. He has previously taught at the law schools of George Washington University, American University, Lewis & Clark College, and Seattle University.

Following a judicial clerkship at the U.S. Court of Appeals for the Federal Circuit, Dr. Sung was in private practice specializing in biotechnology patent litigation with several international law firms. He received a Ph.D. in microbiology from the U.S. Department of Defense, Uniformed Services University and a J.D. from American University, Washington College of Law. In addition to numerous articles, Dr. Sung is the author of two books: *Patent Law Handbook* and *Patent Infringement Remedies*.

Russell D. Wolfinger is director of scientific discovery and genomics at SAS Institute, Inc, one of the top 10 software companies in the world. He earned a Ph.D. in statistics from North Carolina State in 1989 and has been at SAS ever since. His first 10 years were devoted to developing statistical procedures in the areas of linear and nonlinear mixed models, multiple testing, and density estimation. In 2000, he started the Scientific Discovery department at SAS and since then has been leading a team in research and development of software solutions in the areas of genetics, transcriptomics, and proteomics/metabalomics. Dr. Wolfinger is an adjunct faculty member at North Carolina State University, University of North Carolina at Chapel Hill, and University of Missouri.

APPENDIX B

VALIDATION OF TOXICOGENOMIC TECHNOLOGIES: A FOCUS ON CHEMICAL CLASSIFICATION STRATEGIES

WORKSHOP AGENDA

Thursday, July 7, 2005

8:30 am	Introduction (Kenneth Ramos, University of Louisville)
8:45 am	Description of Workshop Concept and Goals (John Quackenbush, Harvard University)

PART 1: Current Validation Strategies and Associated Issues

9:00 am	Experimental Design of DNA Microarray Studies Speaker: Kevin Dobbin, National Cancer Institute
9:30 am	Speaker: Rafael Irizarry, Johns Hopkins University
10:00 am	*BREAK*
10:20 am	Statistical Analysis of Toxicogenomic Microarray Data: Hypotheses, Analysis Methods, and Multiplicity Issues Speaker: Wherly Hoffman, Eli Lilly and Company
10:50 am	Diagnostic Classifier—Gaining Confidence Through Validation Weida Tong, FDA, National Center for Toxicological Research

11:20 am Toxicogenomics: ICCVAM Fundamentals for Validation and Regulatory Acceptance
Leonard Schechtmann, The Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM); FDA, National Center for Toxicological Research

11:50 am Questions and Discussion

12:10 pm *LUNCH*

PART 2: Case Studies—Classification Studies and the Validation Approaches
[Presenters to provide an overview of the study emphasizing the validation strategies employed in this research]

1:10 pm Presenter: Hisham Hamadeh, Amgen
Hamadeh HK. Bushel PR. Jayadev S. Martin K. DiSorbo O. Sieber S. Bennett L. Tennant R. Stoll R. Barrett JC. Blanchard K. Paules RS. Afshari CA. 2002. Gene expression analysis reveals chemical-specific profiles. Toxicological Sciences. 67(2):219-231.

1:40 pm Presenter: Kyle Kolaja, Iconix Pharmaceuticals
Kramer JA. Curtiss SW. Kolaja KL. Alden CL. Blomme EA. Curtiss WC. Davila JC. Jackson CJ. Bunch RT. 2004. Acute molecular markers of rodent hepatic carcinogenesis identified by transcription profiling. Chemical Research in Toxicology 17(4):463-470.

2:10 pm *BREAK*

2:25 pm Presenter: Guido Steiner, Roche
Steiner G. Suter L. Boess F. Gasser R. de Vera MC. Albertini S. Ruepp S. 2004. Discriminating different classes of toxicants by transcript profiling. Environmental Health Perspectives 112(12):1236-1248.

2:55 pm Roundtable discussion (Moderator: John Quackenbush, Harvard University)
Discuss the strengths and limitations of the current vali-

dation approaches and methods to strengthen these approaches.

3:55 pm Summary Statements and Conclusions (Kenneth Ramos, University of Louisville)
Summary of ideas and themes from the presentations, case studies, and discussion.

4:30 pm *ADJOURN*

APPENDIX C

FEDERAL LIAISON GROUP FOR THE NRC COMMITTEE ON EMERGING ISSUES AND DATA ON ENVIRONMENTAL CONTAMINANTS

Samuel Wilson, M.D., ***Chair***
Deputy Director
National Institute of Environmental Health Sciences
Research Triangle Park, NC

William T. Allaben, Ph.D.
Associate Director for Scientific Coordination
FDA National Toxicology Program Liaison
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, AR

Nancy B. Beck, Ph.D., D.A.B.T.
Toxicologist/Risk Assessor
Office of Information and Regulatory Affairs
Office of Management and Budget
Washington, DC

David A. Belluck, Ph.D.
Senior Transportation Toxicologist
FHWA/USDOT
Washington, DC

David Brown, M.P.H.
Staff Assistant to the Director
National Institute of Environmental Health Sciences
Research Triangle Park, NC

Richard A. Canady, Ph.D.
Office of the Commissioner
Food and Drug Administration
Rockville, MD

Daniel A. Casciano, Ph.D.
Director
National Center for Toxicological Research
Food and Drug Administration
Jefferson, AR

Christopher De Rosa, Ph.D.
Director, Division of Toxicology
Agency for Toxic Substances and Disease Registry
Atlanta, GA

John Doll, Ph.D.
Director
United States Patent and Trademark Office
Arlington, VA

Amanda Edens, Ph.D.
Director, Office of Chemical Hazards-Metals
Directorate of Standards and Guidance
OSHA/DOL
Washington, DC

William Farland, Ph.D.
Acting Deputy Assistant Administrator for Science
U.S. EPA
Office of Research and Development
Washington, DC

Kevin Geiss, Ph.D.
Office of Science and Technology Policy
Executive Office of the President
Washington, DC

M. Olivia Harris, M.A., A.B.D.
Senior Environmental Health Scientist
Office of Science
National Center for Environmental Health
Agency for Toxic Substances and Disease Registry
Atlanta, GA

Robert Kavlock, Ph.D.
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, NC

Don Marlowe, Ph.D.
Agency Standards Coordinator
Office of the Commissioner/OSHC
Food and Drug Administration
Rockville, MD

Albert E. Munson, Ph.D.
Director, Health Effects Laboratory Division
National Institute for Occupational Safety and Health
Morgantown, WV

Lawrence Reiter, Ph.D.
Director
National Exposure Research Laboratory
U.S. EPA
Research Triangle Park, NC

Carl M. Schroeder, Ph.D.
U.S. Department of Agriculture
Food Safety & Inspection Service
Office of Public Health and Science
Washington, DC

Paul Schulte, Ph.D.
Director, Education and Information Division
National Institute for Occupational Safety and Health
Robert A. Taft Laboratories
Cincinnati, OH

Karen K. Steinberg, Ph.D.
Chief, Molecular Biology Branch
National Center for Environmental Health
Centers for Disease Control and Prevention
Atlanta, GA

Hal Zenick, Ph.D.
Associate Director for Health
National Health and Environmental Effects Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC