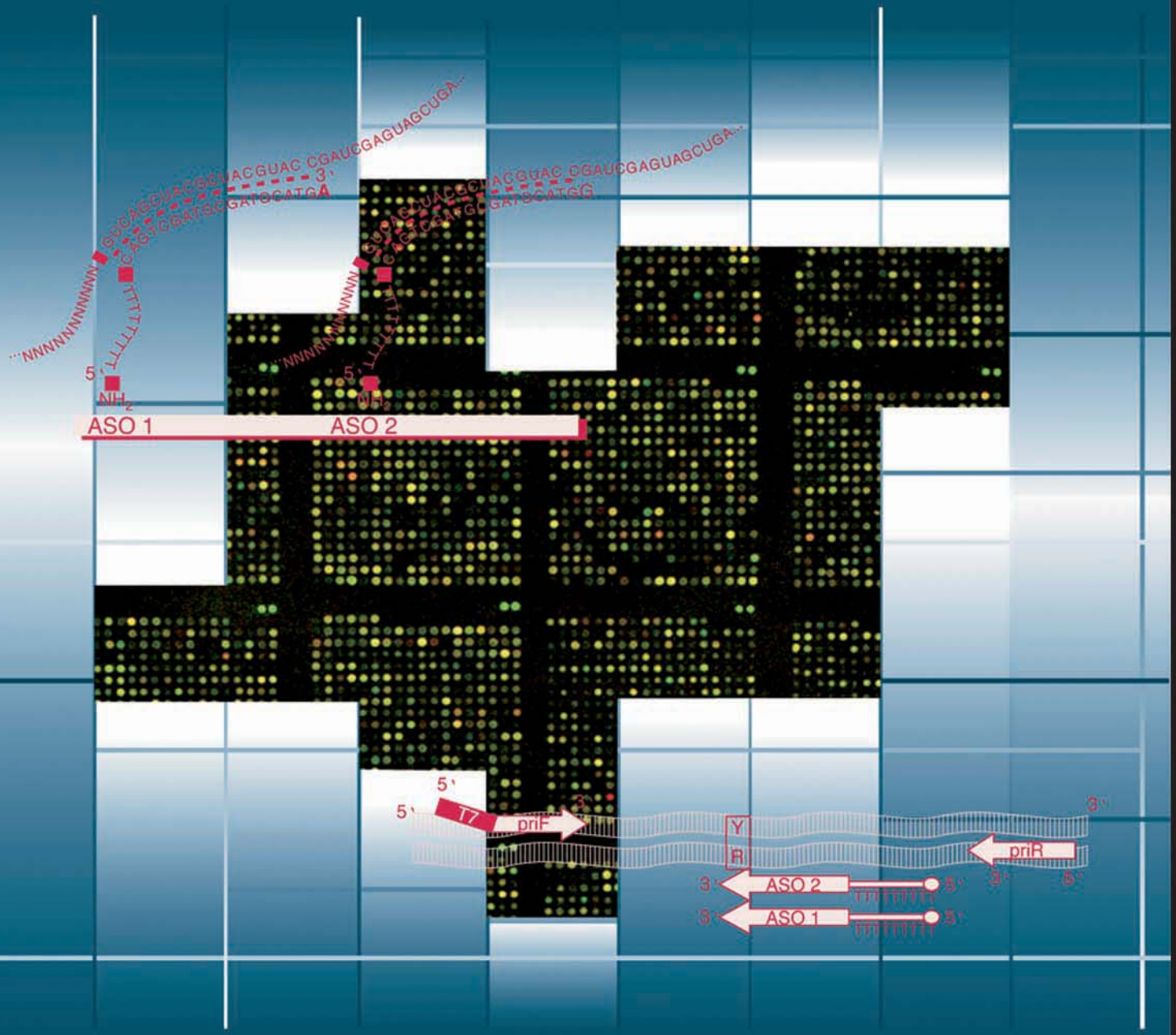
DNA Microarrays

Edited by Ulrike A Nuber



BIOS ADVANCED METHODS

DNA Microarrays

DNA Microarrays

Ulrike A Nuber (Ed.)

Department of Human Molecular Genetics,
Max Planck Institute for Molecular Genetics,
Berlin, Germany

Taylor & Francis
Taylor & Francis Group

Published by:
Taylor & Francis Group
In US: 270 Madison Avenue
New York, NY 10016
In UK: 4 Park Square, Milton Park
Abingdon, OX14 4RN

This edition published in the Taylor & Francis e-Library, 2007.

ISBN 0-203-96733-X Master e-book ISBN

ISBN: 0-415-35866-3 (Print Edition)

A catalog record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

DNA microarrays / [edited by] Ulrike Nuber.
p. ; cm. -- (BIOS advanced methods)
Includes bibliographical references and index.
ISBN 0-415-35866-3
1. DNA microarrays.
[DNLM: 1. Oligonucleotide Array Sequence Analysis--methods. 2. Gene Expression Profiling--methods. QU 450 D629 2005] I. Nuber, Ulrike. II. Title. III. Series.

QP624.5.D726.D633 2005
572.8'636--dc22

2005017596

Editor: Elizabeth Owen
Editorial Assistant: Chris Dixon
Production Editor: Georgina Lucas

T&F informa

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

Visit our web site at http://www.garlandscience.com

Contents

Colour plate section appears between pages 196 and 197

Abbreviations

ABA	abscisic acid
AFLP	amplified fragment length polymorphism
AhR	aryl hydrocarbon nuclear receptor
ALL	acute lymphoblastic leukemia
AMD	age-related macular degeneration
AML	acute myloid leukemia
APAP	acetaminophen
APEX	arrayed primer extension
ar	autosomal recessive
ASO	allele-specific oligonucleotide
ATP	adenosine triphosphate
BACs	bacterial artificial chromosomes
BDNF	brain-derived neurotropic factor
BW	body weight
CCD	charge-coupled device
ccRCC	clear cell renal cell carcinoma
chRCC	chromophobe renal cell carcinoma
CGH	comparative genomic hybridization
ChIP	chromatin immunoprecipitation
CRC	colorectal cancer
CRD	cone-rod dystrophy
CRF	corticotrophin-releasing factor
DEPC	diethylpyrocarbonate
DMSO	dimethyl sulfoxide
DOP PCR	degenerate oligonucleotide primed PCR
EB	empirical Bayes
EGF	epidermal growth factor
EST	expressed sequence tag
FDR	false discovery rate
FISH	fluorescent in situ hybridization
FWER	family-wise error rate
GCV	generalized cross-validation
GSH	glutathione
HAS	hybrid adaptive spline
HMM	hidden Markov model
IDF	inflated degrees of freedom
IP	immunoprecipitation
IVT	*in vitro* transcription
JA	jasmonic acid
LCA	Leber congenital amaurosis
LCV	large copy number variation
LOWESS	locally-weighted scatterplot smoothing
MAD	median of absolute deviation
MALDI-TOF MS	matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MCEM	Markov chain Monte Carlo EM algorithm
MDS	multidimensional scaling
MPC	mesodermal progenitor cell
MPSS	massive parallel signature sequencing
NAA	1-naphthalene acetic acid
NAPQI	N-acetyl-p-benzoquinone imine
NDO	nucleotide dispensation order
NPA	N-1-naphthylphthalamic acid
NPCs	neural progenitor cells

PACs	P1 artificial chromosomes
PAM	prediction analysis of microarrays
PBS	phosphate-buffered saline
PCA	principal components analysis
PCR	polymerase chain reaction
PMT	photomultiplier tube
POSaM	piezoelectric oligonucleotide synthesizer and microarrayer
pRCC	papillary renal cell carcinoma
RAFL	RIKEN *Arabidopsis* full-length (cDNA library)
RCC	renal cell carcinoma
RP	retinitis pigmentosa
RSS	residual sum of squares
RT	reverse transcriptase
SA	salicylic acid
SAM	significance analysis of microarrays
SAGE	serial analysis of gene expression
SBCE	single base chain extension
SBT	sequence-based typing
SDS	sodium dodecyl sulfate
SNP	single nucleotide polymorphism
SNR	signal-to-noise ratio
SOM	self-organizing maps
SPM	single-pulse model
SSC	sodium sodium citrate
SSCP	single strand conformational polymorphism
STGD	Stargardt macular dystrophy
SVM	support vector machine
TNFα	tumor necrosis factor α
TSP	top-scoring pairs
UNG	uracil-N-glycosylase

1 Introduction: DNA microarrays – ten years old, but no old hat

Falk Hertwig and Ulrike A Nuber

In October this year, we celebrate the 10th anniversary of DNA microarrays, as this technology was first mentioned in an article by Schena, M., Shalon, D., Davis, R.W., and Brown P.O. published in the 'Genome Issue' of *Science* in 1995 (1). Predecessors of this technology were dot blots, slot blots, and macroarrays with membranes used as platform. Microarrays have been fascinating the scientific community for the last decade, but still have not reached the limits of their potential and are continuing to invade new fields of biology and medicine.

In the course of deciphering the genomes of many organisms, the need for functional studies of thousands of genes arose, and one step towards this goal has been the identification of expression patterns of genes under normal and pathological conditions. Coincidentally, the 'Genome Issue' of *Science* in 1995 contained two articles that describe mRNA profiling at a large scale: the DNA microarray paper by Schena *et al*. (1), and the paper by Velculescu *et al*. (2) on SAGE (serial analysis of gene expression). However, it seems that microarrays 'won the race' in the field of gene expression profiling – which is mainly due to the high throughput (number of investigated samples per time), and rapid readout of the results.

Schena *et al*. (1) used their first microarray to monitor gene expression – which is undeniably still the most prominent application – but the possibility to use this technique for other purposes is only limited by the scientist's creativity and budget.

This book focuses on microarrays that consist of immobilized DNA molecules, but similar miniaturized hybridization formats have been developed, such as lipid microarrays (3) and protein microarrays (for review see 4). The latter can, for example, be used to monitor the interactions of immobilized proteins with proteins, nucleic acids and small molecules, and offers applications in medical diagnostics (e.g. the screening of patients' sera for specific antigenic properties) (5), and in basic research (e.g. the identification of ligands using protein receptor arrays) (6).

There is probably no field in life sciences (basic or applied) which has not been impinged upon by DNA microarrays. Some examples are their application in cancer research (Chapter 4), pharmacogenomics (Chapter 2), and stem cell research (Chapter 5). In theory, DNA microarrays of every organism can be generated and used, dependent soley on the availability

of DNA sequence information. In this book, the use of yeast, plant, mouse, rat and human arrays is described (see *Figure 1.1*).

Apart from monitoring gene expression, DNA microarrays are used to detect single nucleotide changes (Chapters 6, 7, and 8), unbalanced chromosome aberrations by array CGH (Chapters 11 and 12), or balanced chromosome aberrations by array painting (Chapter 12). The ChIP-on-chip technology was established only a few years ago as a new microarray tool enabling the search for transcription factor binding sites at a large scale (Chapters 13 and 14).

Strangely enough, there are still some scientists who are reluctant to use DNA microarrays. *'With microarrays, you see too much!' 'But there are so many differentially expressed genes…'* – If this sounds like you, don't be shy, read on. Chapters in this book will guide you through the processing of complex data and introduce you to different approaches for handling your results.

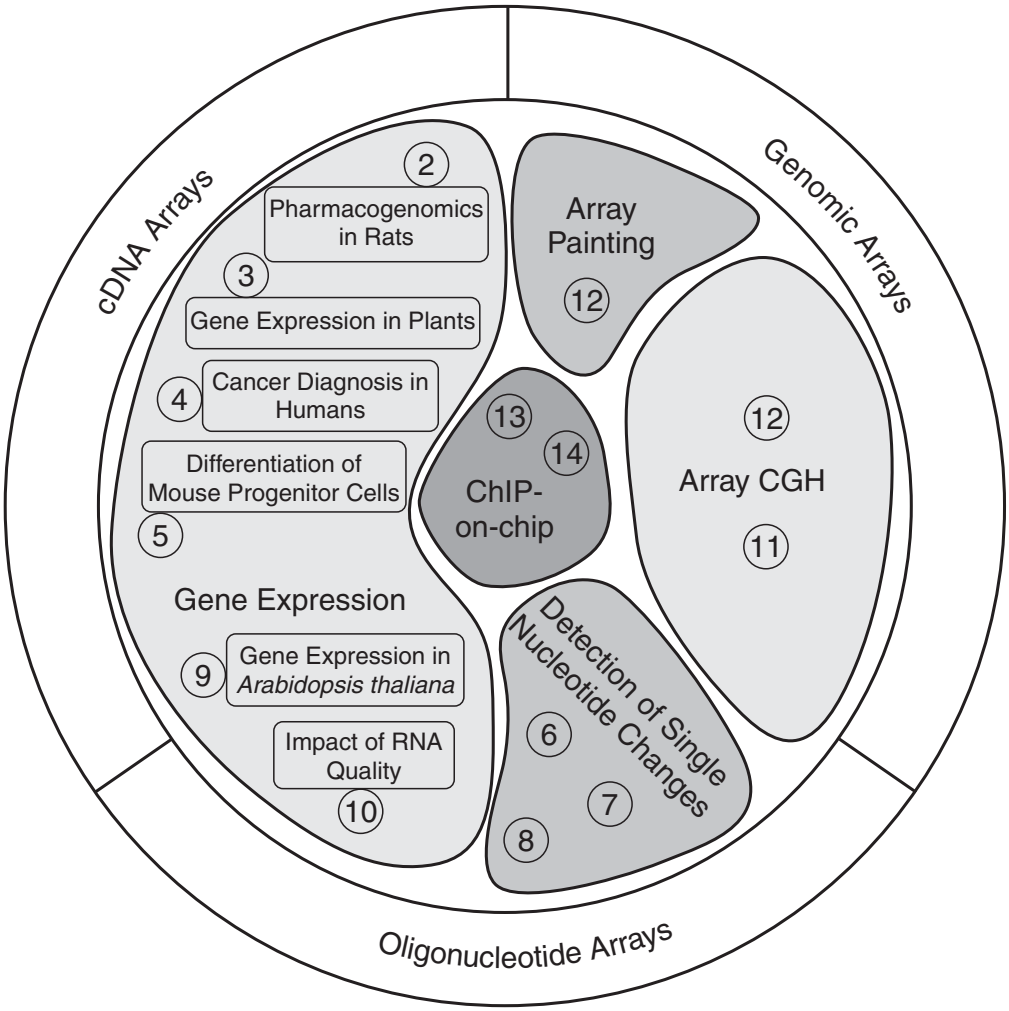


Figure 1.1.

Chapter overview showing different microarray platforms and applications described in this book. Numbers indicate book chapters.

In contrast, other scientists, seduced by the temptation to produce vast amounts of data, may face some of the pitfalls associated with DNA microarray experiments. If you belong to this category, chapters in this book will show you some strategies on how to set up a good microarray experiment, resulting in relevant data, and how to avoid producing data which cannot be processed by the best bioinformatician…

Finally, to the rest of the interested readers, who may have already gained some experience with DNA microarrays, this book should provide new aspects (for example the impact of RNA quality on Affymetrix GeneChip analyses, Chapter 10) and detailed protocols covering all steps of a microarray experiment from the production of the array to data analysis and storage.

1.1 How to perform a microarray experiment

While musing at the thought of using microarrays, many questions pop up in one's mind: What type of arrays do I want to use? How do I design my experiments? Which labeling method do I want to use? How do I analyze my microarray data?

To obtain relevant results and perform a microarray experiment that best suits your needs, you should carefully set up a plan before starting at the bench. You will find help and inspiration in the various chapters. Regarding the design of DNA microarray experiments, we would like to refer to a review by Yang and Speed (7).

DNA microarrays can be classified according to the type of probes on the array (cDNA, oligonucleotides, genomic fragments), their generation and immobilization. In many cases presynthesized molecules (PCR products, oligonucleotides, isolated DNA) are deposited on the array either by contact printing (using metal pins that carry small volumes of probe solution due to capillary action) or by non-contact printing, when probe solution is dispensed by ink-jet printing. In addition, several companies generate high-density microarrays by synthesizing oligonucleotides *in situ (*for review see 8). The synthesis is either based on specific base deprotection by light (coordinated by photomasks or digital micromirror devices) or on chemical deprotection and the use of ink-jet technology. An alternative to these commercial systems is the open-source platform POSaM (piezoelectric oligonucleotide synthesizer and microarrayer), described by Lausted *et al*. (9). These authors present the low-cost production of *in situ* synthesized oligonucleotide arrays (containing 9800 features) in their lab. An overview of commercially available oligonucleotide arrays is given in *Table 1.1*.

cDNA arrays can also be purchased (see *Table 1.2*), or produced as described in various chapters of this book (Chapters 2, 3, 4, and 5). In addition to protocols for the generation of genomic DNA arrays (see Chapters 11 and 12), you can find commercial suppliers in *Table 1.2*.

After the hybridization of labeled target molecules to DNA arrays, fluorescent signals are detected by laser scanners or CCD cameras. Chapters 15 and 16 describe image acquisition and image data conversion.

Finally, microarray data produced in large scale need to be stored and processed to obtain relevant and meaningful results. At this point the field

Table 1.1. Commercially available oligonucleotide microarrays

Principle of DNA microarray generation	Company
***In situ* synthesis**	
Photodeprotection using photomasks (~25mers)	Affymetrix (http://www.affymetrix.com/)
Photodeprotection using digital mirrors (DMD) (24mers-70mers)	NimbleGen#* (http://www.nimblegen.com/)
Chemical deprotection using ink-jet technology (60mers)	Agilent Technologies* (http://www.home.agilent.com)
Presynthesized oligonucleotides spotted onto arrays	
50mers on expoxy surface glass slides	MWG (http://www.mwg-biotech.com/)
80mers on coated glass slides	BD Biosciences (Clontech) (http://www.clontech.com/)
Long oligomers	TeleChem International, Inc. (http://www.arrayit.com/)
30mers on 3D-matrix-coated slides	Mergen Ltd. (www.mergen-ltd.com)
30mers on 3D-matrix-coated slides	GE Healthcare (Amersham Biosciences) (http://www5.amershambiosciences.com)
50mers on 3-micron beads	Illumina, Inc. (http://www.illumina.com)
70mers on proprietary slide substrate	Microarrays Inc* (http://www.microarrays.com)

Arrays for ChIP-on-chip assays (#) or array CGH (*) are also provided.

Table 1.2. Companies producing cDNA and genomic microarrays

cDNA arrays	Genomic arrays
Miltenyi Biotec GmbH (http://www.miltenyibiotec.com)	Aviva Systems Biology (http://www.avivasysbio.com)
Scienion AG (www.scienion.com)	Panomics (http://www.panomics.com)
Takara Bio, Inc. (http://bio.takara.co.jp)	Spectral Genomics, Inc (http://www.spectralgenomics.com)
Cambrex Bio Science (http://www.cambrex.com)	Vysis (Abbott Laboratories) (http://www.vysis.com)

of bioinformatics found a vast playground and the increased use of microarrays triggered the co-evolution of various bioinformatics methods (*Figure 1.2*). In general, all data need to be normalized. Not to get you lost at this early step, several normalization methods are presented in Chapter 17. The special case of normalizing array CGH data is described in Chapter 21.

After normalization, further data processing depends on the questions you intend to address. If you want to compare gene expression under different biological conditions (e.g. normal vs. tumor tissue, wild type vs. knockout/transgene cells, control vs. drug-treated cells etc.) the first and foremost question concerns differential gene expression. Chapter 18 takes you on a safe trip to a list of differentially expressed genes.

DNA microarray time course experiments are highly suitable to monitor the expression of a very large number of genes during a biological process over a defined period of time (one example is given in Chapter 5). Chapter 20 describes the statistical analysis of time-course data.

Chapter 19 deals with clustering and classification. Clustering is, for example, used to find a group of genes, which have similar expression patterns or a group of samples (e.g. tissue samples from patients), which show likewise expression of a set of genes. Classification methods can determine whether a gene expression profile of a tissue sample belongs to a certain class, and are applied to predict disease courses.

Finally, special applications require special bioinformatic analyses. Therefore, the processing of data generated by single nucleotide polymorphisms (SNP) detection and ChIP-on-chip experiments is addressed separately in Chapters 8 and 13, respectively.

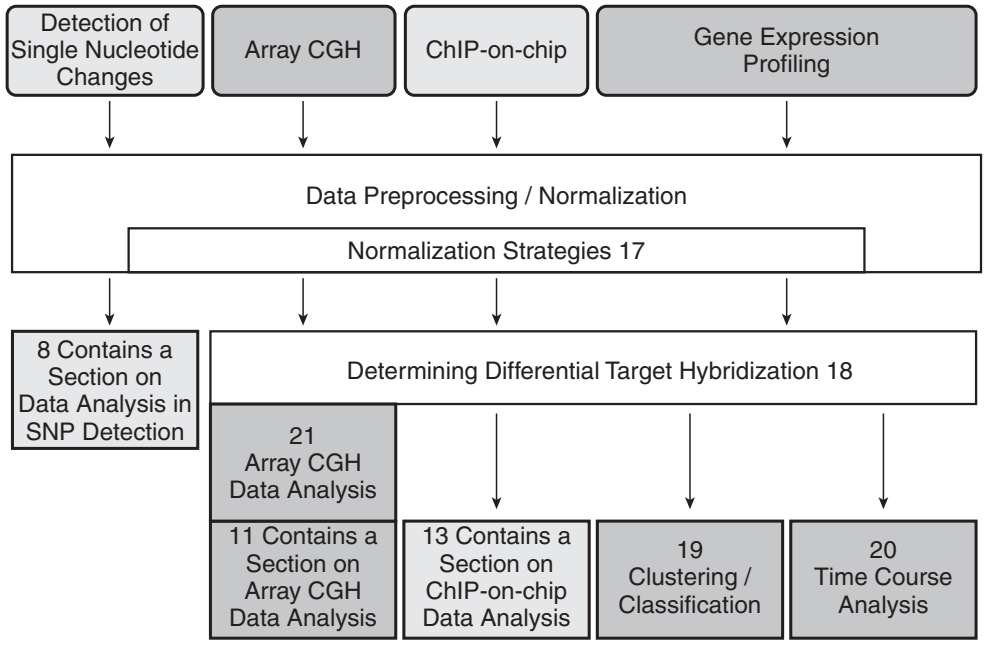


Figure 1.2.

Flowchart of bioinformatics methods that are used at different steps of microarray data analysis. Numbers indicate book chapters.

Ten years ago, DNA microarrays became fashionable because they enable high-throughput gene expression analyses. Over the years, they gained importance with their expanding use in medical diagnostics and research and even now scientists are continuing to advance this technology and its applications. The examples of ChIP-on-chip and array painting show that the combination of two techniques can lead to a new technology – so it will be exciting to see what's yet to come….

References

1. Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–470.
2. Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995) Serial analysis of gene expression. *Science* **270**: 484–487.
3. Lahiri J, Jonas SJ, Frutos AG, Kalal P and Fang Y (2001) Lipid microarrays. *Biomed Microdevices* **3**: 157–164.
4. Labaer J and Ramachandran N (2005) Protein microarrays as tools for functional proteomics. *Curr Opin Chem Biol* **9**: 14–19.
5. Lueking A, Possling A, Huber O, Beveridge A, Horn M, Eickhoff H, Schuchardt J, Lehrach H and Cahill DJ (2003) A nonredundant human protein chip for antibody screening and serum profiling. *Mol Cell Proteom* **2**: 1342–1349.
6. Fang Y, Webb B, Hong Y, Ferrie A, Lai F, Frutos AG and Lahiri J (2004) Fabrication and application of G protein-coupled receptor microarrays. *Methods Mol Biol* **264**: 233–243.
7. Yang YH and Speed T (2002) Design issues for cDNA microarray experiments. *Nat Rev Genet* **3**, 579–588.
8. Gao X, Gulari E and Zhou X (2004) In situ synthesis of oligonucleotide microarrays. *Biopolymers* **73**: 579–596.
9. Lausted C, Dahl T, Warren C, King K, Smith K, Johnson M, Saleem R, Aitchison J, Hood L and Lasky SR (2004) POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biol* **5**: R58.

cDNA microarray analysis and its role in toxicology – a case study

2

Alexandra N Heinloth, Gary A Boorman and Richard S Paules

2.1 Introduction

Toxicogenomics

Nuwasyir and colleagues (1) in 1999 defined 'toxicogenomics' as the intersection of toxicology and genomics. They proposed that the goal of this new discipline is to identify potential toxicants and to clarify their mechanism of action with the help of genomics resources. Since then, major efforts have been undertaken to establish data sets that include a diversity of compounds and environmental stressors. This will eventually allow classification of unknown or novel compounds into mechanistic groups. By doing so, researchers hope to achieve toxicant or toxicant-group-specific genomic signatures which indicate exposure and initiation of toxic events. This might not only be valid for known and already well-defined toxicants, but perhaps more importantly, for unknown toxicants or compounds under development. Achieving this goal would allow identification of potential toxicity prior to indications of overt toxicity for novel compounds and could allow for very sensitive exposure monitoring. Several groups have undertaken efforts to classify compounds based on gene expression data. One of the first classification studies in toxicogenomics was published by Waring *et al.* in 2001 (2). Here the authors retrieved gene expression data from livers of rats exposed to 15 different hepatotoxicants and showed correlations between differentially expressed genes, histopathological and clinical chemistry changes. They also demonstrated that gene expression analysis allows for the identification of mechanistically related compounds and reveals a higher degree of similarity between RNA derived from animals treated with the same compound than to those exposed to other hepatotoxicants. Hamadeh and colleagues in 2002 performed the first toxicological classification study that included blinded samples. In this study, the authors first determined gene expression patterns for three different peroxisome proliferators and one barbiturate (3). This data was utilized as a training set and identified discriminating signatures between compounds. Coded RNA samples from animals exposed to either a barbiturate or peroxisome proliferators were subjected to gene expression

analysis. This study demonstrated that it was possible to predict the class of compound to which the rats were exposed based on gene expression profiles for those blinded liver RNA samples (4).

Mechanisms of toxicity

Comparison of gene expression profiles of novel or poorly defined compounds with those from well-defined drugs or toxicants can not only assign those compounds to a known class, but also elucidate potential mechanisms of action. This is based on the assumption that monitoring global gene expression changes as a result of exposure gives indications about which physiological or pathological processes within the organ are activated or repressed. Waring and colleagues (5) demonstrated this analysis in a study in which rats were exposed to a thienopyridine inhibitor (A-277249) and liver tissue was examined for gene expression changes. Comparison of those changes with a database of profiles from 15 known hepatotoxicants elucidated greatest similarity of the test compound with two known activators of the aryl hydrocarbon nuclear receptor (AhR). They concluded that the activation of AhR mediated the hepatic toxicity observed after exposure to A-277249 (5).

Acetaminophen as a model compound

We chose acetaminophen (APAP), one of the most popular analgesics worldwide, as a model compound to study genomic responses in liver tissue. This choice was driven by several criteria we believe to be of crucial importance for compound selection. First, APAP is the focus of major health concerns in the US and Europe. Accidental overdoses and ingestions with suicidal intent make APAP the leading cause of drug-induced acute liver failure in the United States (6). Secondly, rodents metabolize APAP similar to humans and are therefore an appropriate model system. APAP is metabolized by several isoforms of cytochrome p450 to the highly reactive metabolite N-acetyl-p-benzoquinone imine (NAPQI). At low, therapeutic concentrations, this metabolite is detoxified by conjugation with glutathione (GSH). At high, toxic concentrations, the liver is depleted of GSH and NAPQI is covalently bound to proteins (7). Thirdly, significant information already exists about APAP metabolism and toxicity in the liver. Toxicogenomics as an emerging field can benefit from placing the results in context with a wealth of previously well-documented published findings – with the goal to recapitulate and expand existing knowledge.

Experimental design

In this study, we treated rats with a single dose of 0, 50, 150 or 1500 mg kg^{-1} body weight (BW) APAP and sacrificed them 6, 24 or 48 h after treatment. Livers were harvested for gene expression and histopathological analysis, and blood was collected for serum chemistry. While the two lower doses showed neither histopathological nor serum enzyme alterations, 1500 mg kg^{-1} APAP induced signs of centrilobular necrosis and significant serum enzyme elevations 24 and 48 h after treatment (8).

In order to perform gene expression analysis, total RNA was isolated from liver tissue and microarray analysis was performed as described in the Protocols. The complete data set is available at: http://dir.niehs.nih.gov/microarray/datasets/home-pub.htm. After performing cluster analysis (9) with all differentially expressed genes across all doses and time points, it became obvious that a distinct subset of genes was regulated similarly after low and high dose exposure to APAP (8). Further analysis of these gene expression responses revealed that those genes regulated in common after high- and low-dose exposure belonged to distinct metabolic pathways. Many of the genes down-regulated after treatment with 50 or 150 mg kg^{-1} APAP were involved in energy consuming biochemical pathways like gluconeogenesis, fatty acid synthesis, cholesterol synthesis, porphyrin synthesis, sterol synthesis and the urea cycle (8). Analysis of differentially expressed genes after 1500 mg kg^{-1} APAP showed, besides other changes, a strong down regulation of genes in those same energy demanding processes. Not only were similar gene changes observed after this higher dose, but more members of the same biological pathway were changed.

The converse was true with up-regulated genes involving energy production. After treatment with 150 mg kg^{-1} APAP, genes involved in energy producing biochemical pathways like glycolysis and mitochondrial ω-hydroxylation were up regulated. Exposure to 1500 mg kg^{-1} APAP resulted in a more pronounced effect on the same processes, as well as additional genes in those processes that were over-expressed in comparison to control livers. Also, genes in other energy producing pathways like the tricarboxylic acid cycle, pentose phosphate pathway, and mitochondrial β-oxidation were up-regulated after exposure to 1500 mg kg^{-1} APAP.

We concluded from these results that the liver appeared to be compensating for energy depletion after exposure to an overtly toxic dose of APAP (1500 mg kg^{-1}). Strikingly, similar responses were seen in livers following exposure to sub-toxic doses of APAP (50 and 150 mg kg^{-1}), even though there was no histopathological evidence of toxicity after those low doses. As might be predicted, these attempts of the liver to compensate for energy depletion were more pronounced after exposure to the clearly toxic dose of 1500 mg kg^{-1} APAP.

To test the hypothesis that the liver suffered from energy depletion after exposure to APAP, we performed measurements of ATP levels in liver tissue after exposure to high and low doses of APAP. As shown in *Figure 2.1*, statistically significant decreases in ATP levels were found only at 3 and 48 h after exposure to 1500 mg kg^{-1} APAP. Doses of 50 and 150 mg kg^{-1} APAP did not produce any significant decreases of ATP levels as measured in this assay.

The gene expression profile suggested energy depletion after all doses. We suspected that the ATP assay lacked the necessary sensitivity to show slight decreases, since energy depletion may have occurred only in a small subpopulation of hepatocytes immediately adjacent to the central vein where toxicity is first seen. As the production of ATP in the cell is primarily a function of mitochondria, we hypothesized that the energy depletion after APAP exposure was caused by mitochondrial damage. Therefore we performed ultrastructural analysis on liver tissue after treatment with 0, 50, 150 or 1500 mg kg^{-1} APAP. Six hours after treatment with 150 and

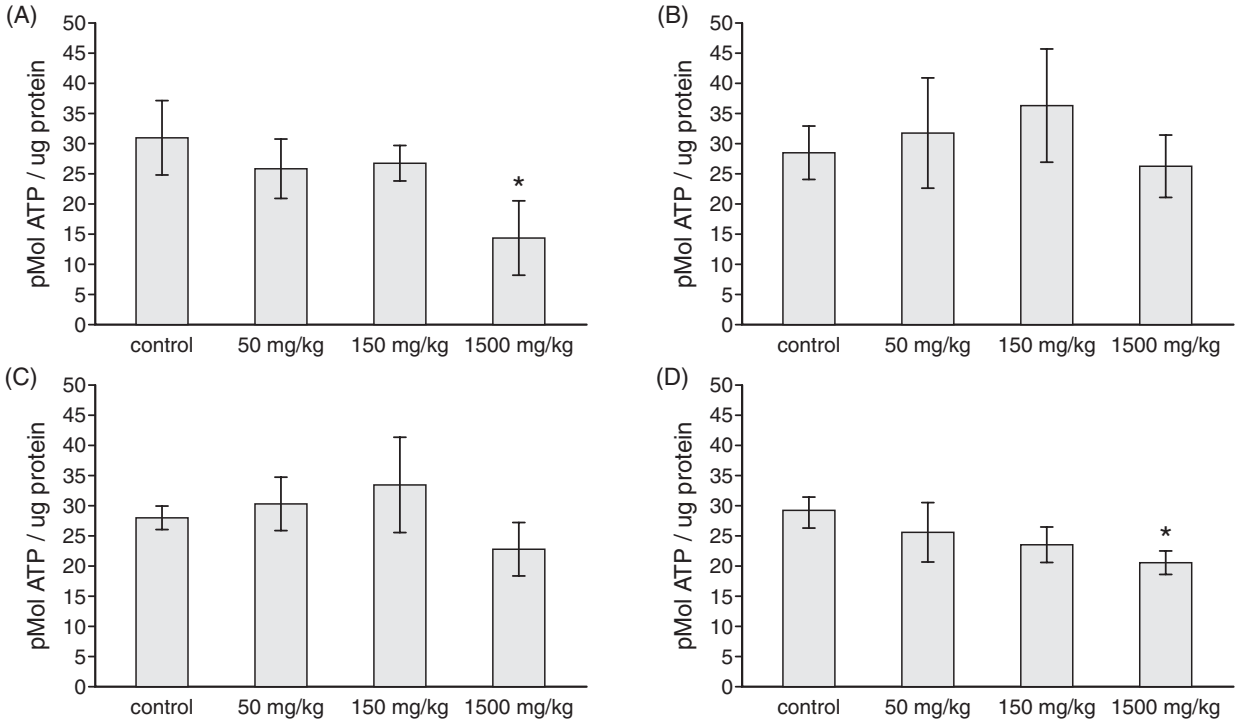


Figure 2.1.

Hepatic ATP levels after exposure to APAP: (**A**) 3 h, (**B**) 6 h, (**C**) 24 h and (**D**) 48 h after exposure to acetaminophen. Bar graphs represent pmol ATP per μg protein (mean ± S.E.). Asterisks indicates $p < 0.02$ for statistical differences between animals treated with APAP and sham-treated control animals ($n = 3$).

1500 mg kg^{-1} APAP we found mitochondria that had lost electron density, indicative of mitochondrial damage after those doses (*Figure 2.2*). At 150 mg kg^{-1} only a few hepatocytes immediately adjacent to the central vein had evidence of mitochondrial toxicity. This suggests that, at least for acetaminophen, gene expression changes may be a more sensitive indicator of potential toxicity than traditional toxicology endpoints such as histopathology and clinical chemistry.

We concluded from our study that liver gene expression profiles in response to exposure to sub-toxic doses of APAP have the ability to indicate potential toxicity of higher doses of this hepatotoxicant. We identified gene expression changes indicative of cellular ATP depletion after sub-toxic doses (50 and 150 mg kg^{-1} APAP), and found that those changes became more pronounced after exposure to a toxic dose (1500 mg kg^{-1} APAP). Therefore, microarray analysis appears to be an extremely useful and sensitive tool to predict potential adverse effects of exposures.

2.2 Gene expression profiling reveals indicators of potential adverse effects

The study presented here tested the hypothesis that gene expression analysis after exposure to sub-toxic doses would allow prediction of adverse effects that only become manifest after exposure to toxic doses. Therefore

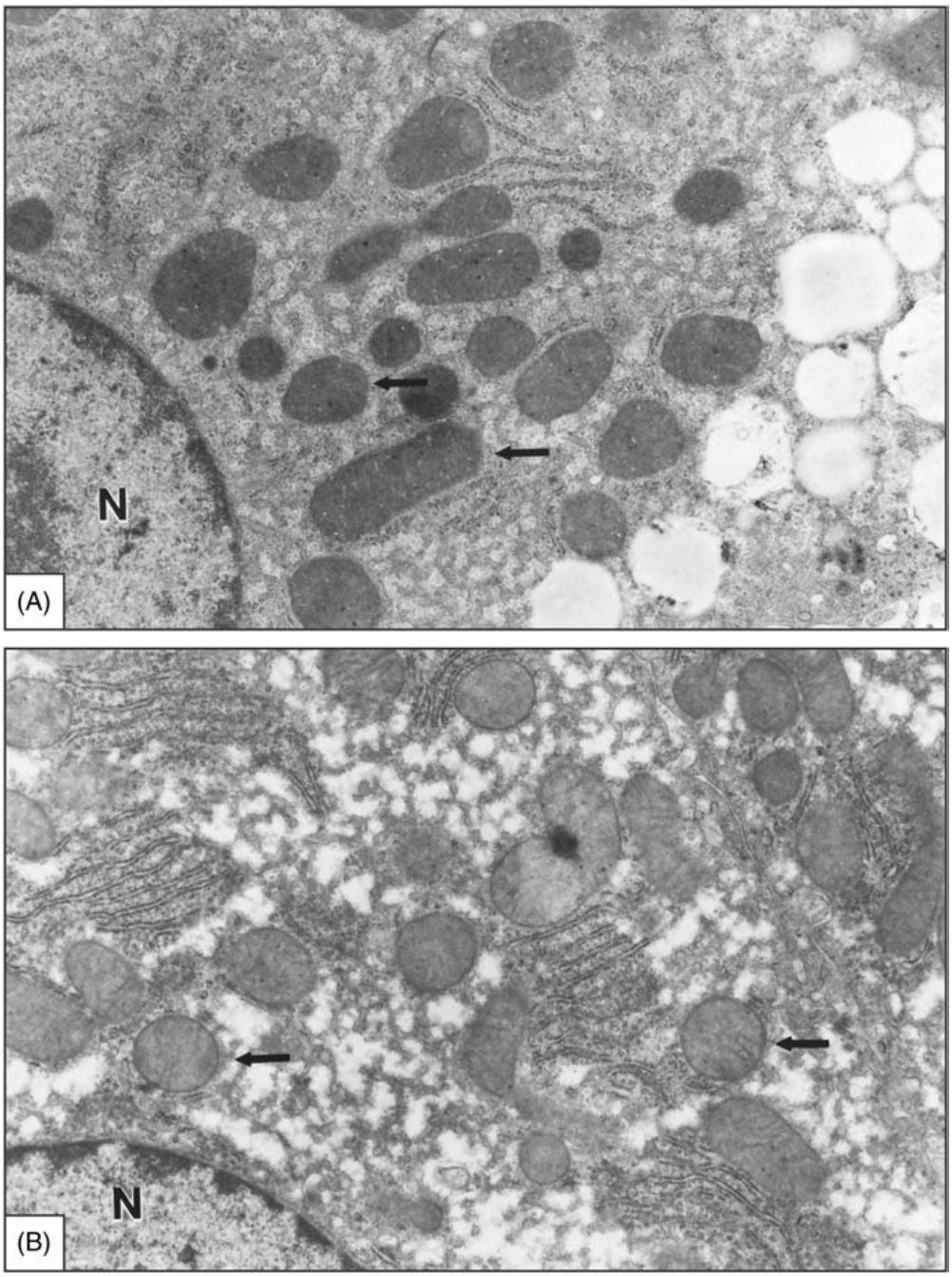
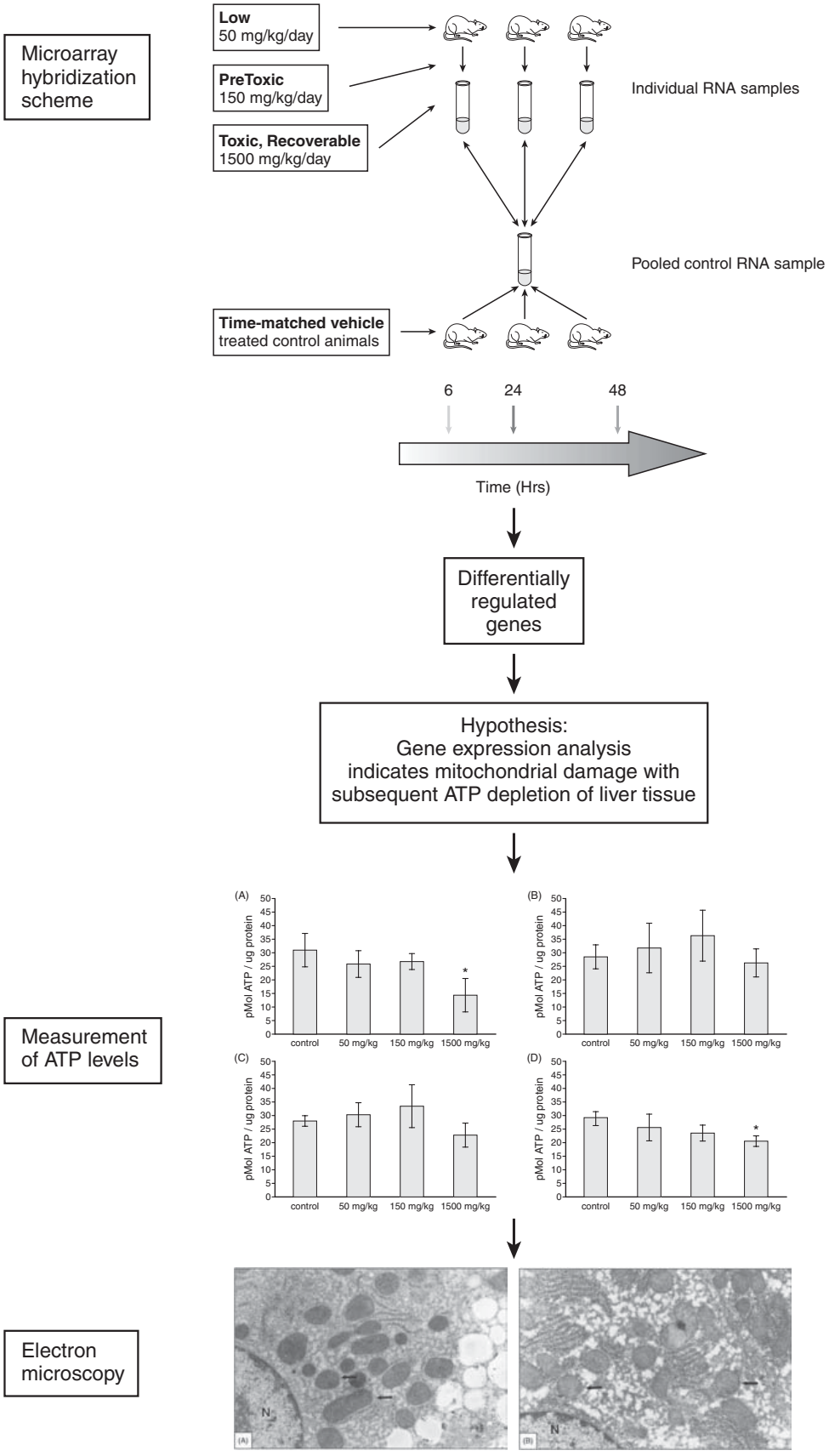


Figure 2.2.

Electron micrographs of centrilobular hepatocytes. (**A**) Hepatocyte from control animal. (**B**) Hepatocyte from animal after treatment with 150 mg kg^{-1} APAP. Arrows point to mitochondria.

we treated rats with sub-toxic (50 and 150 mg kg^{-1}) and toxic (1500 mg kg^{-1}) doses of APAP and performed microarray analysis on total RNA isolated from livers of those animals. As demonstrated, the observed gene expression changes indicated cellular energy depletion at doses of APAP that did not cause any traditional manifestations of toxicity. Ultrastructural studies


Microarray hybridization scheme
Low
50 mg/kg/day
PreToxic
150 mg/kg/day
Toxic, Recoverable
1500 mg/kg/day
Individual RNA samples
Pooled control RNA sample
Time-matched vehicle
treated control animals
6
24
48
Time (Hrs)
Differentially regulated genes
Hypothesis:
Gene expression analysis
indicates mitochondrial damage with
subsequent ATP depletion of liver tissue
Measurement of ATP levels
(A)
(B)
(C)
(D)
pMol ATP / ug protein
control
50 mg/kg
150 mg/kg
1500 mg/kg
Electron microscopy
N

revealed mitochondrial damage after exposure to 150 mg kg^{-1}, but not to 50 mg kg^{-1} APAP, although gene expression analysis suggested some modest cellular energy loss even after exposure to 50 mg kg^{-1} APAP. This indicates that gene expression changes are very sensitive markers of cellular stress, even in the absence of any apparent phenotypic changes (8).

Our results have great implications for future toxicological research. On the one hand, the prediction of toxicity based on, in the traditional sense, sub-toxic exposure levels has great potential for identification of compounds that have toxicological potential. Comparison of gene expression profiles retrieved from unknown toxicants against extensive toxicogenomics databases might reveal toxic potential of those compounds – and save the effort of doing prolonged dose-finding studies to establish phenotypic anchors. This is even more important in the case of novel and genetically engineered compounds for which limited amounts may be available.

On the other hand, with the advent of toxicogenomics, it is becoming apparent that slight alterations in the environment will manifest themselves in gene expression changes compared with sham-treated controls. The challenge is to determine which of those changes are truly meaningful and indicative of potential harm and which are pharmacological, adaptive, or confounding events – related to animal housing, animal handling, feeding and the circadian cycle.

2.3 cDNA microarrays – curse or blessing?

The biggest challenge for scientists confronted with gene expression data is the biological interpretation of large datasets. With the development of microarray technologies, biological sciences experienced an exponential leap from a paradigm where one experiment involves one or two measurements in a known area to a new paradigm where one experiment provides data on thousands of measurements often involving genes for which the scientist may have very little familiarity. This demands a new form of abstract thinking from the scientists involved – new principles are developing as to how to approach these data sets most efficiently. Scientists need to achieve both the ability for grasping the big picture provided by the gene expression changes and to focus on the truly novel insights in the data sets. At this point, traditional biological follow-up studies are necessary to test the biological hypotheses developed from the gene expression analysis. This is both a challenge and an opportunity: while it was never before possible to learn so much from a single experiment, it was also never before so difficult to extract meaningful information from the results of a single experiment.

2.4 The future of toxicogenomics – prediction of toxicity

One of the major goals of toxicogenomics is the prediction of toxicity of unknown compounds. To facilitate achieving this goal, several institutes have started efforts to establish databases that would allow the collection of gene expression data, to analyze this data and compare and contrast gene expression results. The overall goal is to learn more about known toxicants

by examining them in a full, systems biology context and to discover similarities between novel compounds and known toxicants. One of these databases is the Chemical Effects in Biological Systems (CEBS) (10), being developed by the National Center for Toxicogenomics at the National Institute of Environmental Health Sciences. The mission of CEBS is to provide a repository of data retrieved from toxicogenomic studies and to enable scientists to more easily perform cross-compound and cross-experiment analysis to gain further insight into toxicological processes.

Another challenge the field faces are the difficulties in comparing data sets created by different investigators on different platforms. Often several additional factors, like animal strains, doses, time points and normalization procedures used, are different. In the past, those differences were often not or at best only partially reported and attached to public data sets. To improve this situation, the Microarray Gene Expression Data society (http://www.mged.org) was formed and has developed standards for publication of microarray data, published as 'Minimal Information About a Microarray Experiment' (MIAME, see also Chapter 22) (11). The information collected according to these guidelines enables researchers to replicate analysis published in reports following those guidelines. As an extension of this effort, MIAME-Tox was developed (http://www.mged.org/Workgroups/tox/tox.html) which collects additional information considered necessary to interpret and replicate toxicology studies.

As described above, the future of toxicogenomics lies in moving away from the analysis of one compound at a time to being able to utilize extensive databases and performing cross-compound analysis. This promises to facilitate the discovery of novel general concepts about the pathophysiology of certain adverse phenotypes – and the discovery of potential therapeutic interventions at the inception of an adverse response to an environmental stressor.

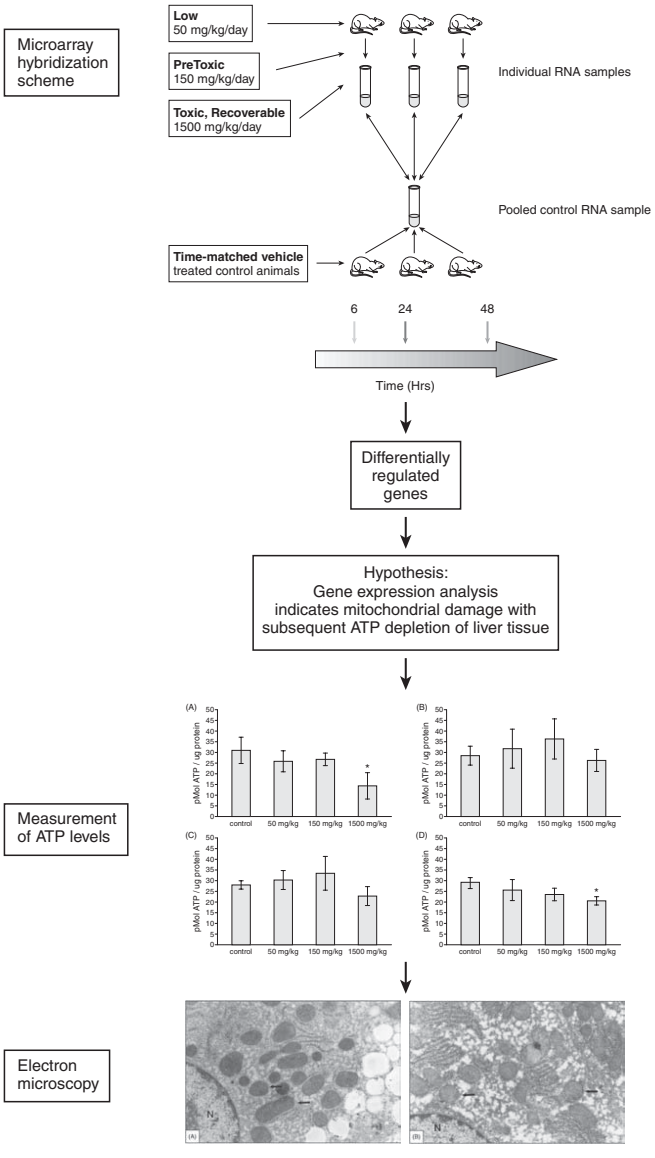
References

1. Nuwaysir EF, Bittner M, Trent J, Barrett JC and Afshari CA (1999) Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinogen* **24**: 153–159.
2. Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, Morfitt DC, Buratto B, Roberts C, Schadt E and Ulrich RG (2001) Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol Appl Pharmacol* **175**: 28–42.
3. Hamadeh HK, Bushel PR, Jayadev S, *et al.* (2002) Gene expression analysis reveals chemical-specific profiles. *Toxicol Sci* **67**: 219–231.
4. Hamadeh HK, Bushel PR, Jayadev S, *et al.* (2002) Prediction of compound signature using high density gene expression profiling. *Toxicol Sci* **67**: 232–240.
5. Waring JF, Gum R, Morfitt D, Jolly RA, Ciurlionis R, Heindel M, Gallenberg L, Buratto B and Ulrich RG (2002) Identifying toxic mechanisms using DNA microarrays: evidence that an experimental inhibitor of cell adhesion molecule expression signals through the aryl hydrocarbon nuclear receptor. *Toxicology* **181–182**: 537–550.
6. Lee WM (2003) Acute liver failure in the United States. *Semin Liver Dis* **23**: 217–226.
7. Mitchell JR, Jollow DJ, Potter WZ, Gillette JR and Brodie BB (1973) Acetaminophen-induced hepatic necrosis. IV. Protective role of glutathione. *J Pharmacol Exp Ther* **187**: 211–217.

8. Heinloth AN, Irwin RD, Boorman GA, *et al*. (2004) Gene expression profiling of rat livers reveals indicators of potential adverse effects. *Toxicol Sci* **80**: 193–202.
9. Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95:** 14863–14868.
10. Waters M, Boorman G, Bushel P, *et al*. (2003) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenom* **111**: 15–28.
11. Brazma A, Hingamp P, Quackenbush J, *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**: 365–371.
12. Council NR (1996) *Guide for the Care and Use of Laboratory Animals*. National Academy Press, Washington, DC
13. Easom RA and Zammit VA (1984) A cold-clamping technique for the rapid sampling of rat liver for studies on enzymes in separate cell fractions. Suitability for the study of enzymes regulated by reversible phosphorylation-dephosphorylation. *Biochem J* **220**: 733–738.
14. Hamadeh HK, Knight BL, Haugen AC, *et al*. (2002) Methapyrilene toxicity: anchorage of pathologic observations to gene expression alterations. *Toxicol Pathol* **30**: 470–482.
15. Lennon G, Auffray C, Polymeropoulos M and Soares MB (1996) The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**: 151–152.
16. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS and Trent JM (2002) Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* **18**: 1207–1215.
17. Bushel PR, Hamadeh H, Bennett L, Sieber S, Martin K, Nuwaysir EF, Johnson K, Reynolds K, Paules RS and Afshari CA (2001) MAPS: a microarray project system for gene expression experiment information and data validation. *Bioinformatics* **17**: 564–565.
18. Wolfinger, RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C and Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**: 625–637
19. Martin FL and McLean AE (1998) Comparison of paracetamol-induced hepatotoxicity in the rat in vivo with progression of cell injury in vitro in rat liver slices. *Drug Chem Toxicol* **21**: 477–494.

Protocols

CONTENTS



Protocol 2.1: In-life study

ANIMAL MODEL

Male F344/N rats from Taconic Laboratories, Inc., Germantown, NY were used in these studies. The animals were obtained 36 ± 3 days old, and were about 89 ± 3 days old at the start of the study. Rats were housed three per cage in 22 × 12.5 × 8 inch (l × w × h) polycarbonate cages (Lab Products, Inc., Seaford, DE) with polyester cage filters (Snow Filtration Co., Cincinnati, OH). In studies dealing with fed animals, it is especially important to keep track of the cage assignment of the single animals. We decided to perform this study with fed animals in order to be closer to a human physiologically relevant situation. One potential disadvantage of this study design is that variability is introduced due to differences in feeding between animals in the same cage, based on their social hierarchy.

The room temperature was between 71 and 75°F, with the humidity between 36% and 48%. The animals were fed *ad libitum* with irradiated NTP-2000 wafer feed (Ziegler Brothers, Gardners, PA) and had *ad libitum* access to city water (Durham, NC). The rats had a 12-h light period from 6 a.m. to 6 p.m. and a 12-h dark period from 6 p.m. to 6 a.m. The animals were dosed between 9 and 11 a.m.

STUDY DESIGN

For gene expression analysis, groups of three male rats received acetaminophen (APAP, 99% pure, Sigma, St. Louis, MO) as a suspension in 0.5% aqueous ethyl cellulose (USP/FCC grade; Fisher Scientific Company, St. Louis, MO) by gavage. Doses utilized in this study were 0 (vehicle only), 50, 150 and 1500 mg kg^{-1} BW^{-1}. The animals were sacrificed (see below) at 6, 24 or 48 h after dosing.

For ATP measurements, groups of two male rats were dosed with 0 (vehicle only), 50, 150 or 1500 mg kg^{-1} BW^{-1} APAP and sacrificed after 3, 6, 24 or 48 h.

The entire study was replicated for a biological confirmation of the results. In the case of the gene expression analysis, the data presented are primarily derived from one of the two replicates, for ATP measurements the data is presented as average of both replicates.

We performed the studies according to the guidelines in the NIH Guide for the Care and Use of Laboratory Animals (12). An approved Animal Study Protocol was on file prior to initiation of the study.

NECROPSY

For gene expression analysis, animals were euthanized with carbon dioxide from a regulated source. Blood was drawn from the posterior vena cava for clinical chemistry. The liver was weighed and a mid-sagittal section from the left lateral lobe was taken for histology, the remainder of the liver was cubed and frozen in liquid nitrogen. The time from drawing the blood to freezing the liver was less than 90 s. Tissues were stored at –80°C until processing for RNA extraction.

For ATP measurements the animals were anesthetized with pentobarbital (pentobarbital sodium injection, Abbott Laboratories, North Chicago, Il) 50 mg per animal i.p.. In pilot studies we had observed ATP loss in liver samples from animals that were euthanized with carbon dioxide. We used the cold clamp method as described by Easom and Zammit (13).

Protocol 2.2: Gene expression analysis

RNA ISOLATION

QIAGEN RNeasy Maxi Kits® (QIAGEN, Valencia, CA) were used to isolate total hepatic RNA as previously described (14). Gene expression of individual treated animals was compared against control pools. Those pools were made up from equal amounts of RNA from each of the three control animals per dose and time point.

RNA LABELING AND HYBRIDIZATION

cDNA was generated by *in vitro* transcription from 35 μg total RNA. RNA was combined with 1 μg oligo $dT_{12–18}$ primer (Amersham Pharmacis Biotech, Piscataway, NJ) and 10 U RNase-inhibitor (Invitrogen, Carlsbad, CA) and heated to 70°C for 10 min. Samples were chilled to 4°C for 2 min, then first strand buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$; Invitrogen), 11 mM dithiothreitol (Invitrogen), 2.2 nM FluoroLink Cy3-deoxy (d) UTP or Cy5-dUTP (Amersham Pharmacia Biotech), dNTP mix (0.7 mM dATP, 0.7 mM dGTP, 0.7 mM dCTP, 0.4 mM dTTP; Amersham Pharmacia Biotech), and 2 μl SUPERSCRIPT™ II Reverse Transcriptase (Invitrogen) were added. After incubation at 42°C for 1.5 h, another 2-μl aliquot of SUPERSCRIPT™ II Reverse Transcriptase was added, and samples were incubated for an additional 1.5 h at 42°C. After cDNA synthesis, RNA was degraded by addition of 30 μl of 0.1 M NaOH and incubation for 30 min at 70°C. The pH was neutralized by addition of 30 μl of 0.1 M HCl, and Cy3- and Cy5-labeled samples were pooled. Microcon-30 filters (Millipore Corp., Bedford, MA) were used to remove unincorporated label. Ten μg human COT1 DNA (Invitrogen) per 10 μg RNA and 20 μg yeast tRNA (Invitrogen) were added to the probe to limit nonspecific binding. Hybridization solution (3 × SSC, 2 × Denhardt's, and 0.8% SDS) was added to the samples and they were boiled for 2 min, and then purified with a 0.45-μm filter (Millipore Corp., Bedford, MA).

Roughly 7000 rat clone cDNAs (Research Genetics, Huntsville, AL) (http://dir.niehs.nih.gov/microarray/chips.htm) were printed on glass slides as described in Hamadeh *et al.* (14). The methods used to produce the chips are available at http://dir.niehs.nih.gov/microarray/methods.htm. The cDNA clones were sequence verified and annotated according to UniGene (15).

The labeled cDNA, representative of cellular mRNA, was applied to cDNA microarray chips, covered by a cover-slide, and incubated for 24 h in a humidified chamber at 65°C. At the end of the incubation period, slides were inverted in 0.5 × SSC, 0.01% SDS for 5 min to remove cover-slides. After that, slides were washed in 0.5 × SSC, 0.01% SDS for 5 min, and in 0.06 × SSC for 5 min. After washing, slides were dried by spinning for 3 min at 1000 *g*.

Each individual treated RNA sample was hybridized against its time- and dose-matched control pool in duplicate with reversal of the fluorescent Cy3 or Cy5 dyes. This resulted in a total of 6 microarray chips per dose and time period.

SCANNING

Fluorescent intensities were measured with an Agilent DNA microarray scanner (Palo Alto, CA). This scanner has a pixel resolution of 10 micron per pixel and we used a PMT gain of 100%.

BIOINFORMATICS

The signal intensities were quantified and normalized with IPLabs image-processing software (Scanalytics, Inc., Fairfax, VA) with the Array Suite 2.0 extension (National Human Genome Research Institute, NHGRI, Bethesda, MD) (16). Differentially expressed genes were defined in MAPS (17) at the 95% confidence level in both hybridizations per RNA pair. We also used a mixed linear model (18) to identify statistically significant differentially expressed genes.

Protocol 2.3: Clinical pathology

We performed clinical chemistry analysis on all animals in the study. This analysis included urea nitrogen, creatinine, total protein, albumin, total bile acid concentrations, and activities of alanine aminotransferase, alkaline phosphatase, creatine kinase, and sorbitol dehydrogenase. The analyses were performed with a Roche Cobas Mira chemistry analyzer (Roche Diagnostics Systems, Inc., Montclair, NJ).

Protocol 2.4: Histopathology

Tissues collected at necropsy were embedded in 10% neutral buffered formalin for 24 to 48 h. After dehydration in 70% alcohol, tissues were embedded in paraffin and H&E slides were made. Two independent pathologists evaluated the liver sections.

Protocol 2.5: Electron microscopy

Two additional rats were exposed to 0, 50, 150 or 1500 mg kg^{-1} BW^{-1} APAP and deeply anesthetized with Pentobarbital at 6 h after dosing. The vessels superior and inferior of the liver were clamped. The posterior vena cava was canulated and the portal vessels severed. The retrograde perfusion was via the posterior vena cave first with RPMI media at 37°C to clean the liver followed by cold 3% glutaraldehyde (in 0.1 M sodium cacodylate buffer pH 7.2) for 30 min. The left lateral lobe was minced into 1-mm cubes and incubated overnight in 3% glutaraldehyde solution, and post fixed in OsO_4. Centrilobular areas were identified on 0.5-μm-thick sections stained with toluidine blue. Thin sections, approximately 80 nm, were examined on a Philips EM 400 electron microscope after staining with uranyl acetate and lead citrate.

Protocol 2.6: ATP measurements

As described by Martin and McLean (19) samples were prepared. Their ATP content was measured in the supernatant with an ATP assay kit (Calbiochem, San Diego, CA) according to the instructions of the manufacturer.

Gene expression profiling in plants using cDNA microarrays

3

Motoaki Seki, Junko Ishida, Maiko Nakajima, Akiko Enju, Ayako Kamei, Youko Oono, Mari Narusaka, Masakazu Satou, Tetsuya Sakurai and Kazuo Shinozaki

3.1 Introduction

DNA microarray technology has become a powerful tool for systematic analysis of expression profiles of large numbers of genes in several plant species (3–8). It is currently being used to investigate a variety of physiological and developmental processes in plants. Expression profiles have been studied for responses to various stresses (9–11), environmental conditions, such as light (12), and day/night cycling (13), symbionts (14), pathogens (15–17), rehydration after dehydration (18), and various developmental processes (19).

Two types of microarrays are currently in use: cDNA microarrays (20) and oligonucleotide microarrays (21). The basic microarray analysis performed with both types of arrays is similar and based on the specific hybridization of a labeled target to the immobilized nucleic acids (probe) on the array. These two systems differ primarily in the nature of the DNA fixed to the solid support. A number of reviews are available on their uses and advantages (1, 2, 3, 7, 22, 23).

In this chapter, we summarize the methodology on gene expression profiling in plants using cDNA microarrays.

3.2 Gene expression profiling methods

Methods for quantifying mRNA abundance in various plant tissues and experimental conditions are: (i) RNA gel-blot (northern) analysis, (ii) differential display (24), (iii) quantitative real-time PCR, (iv) cDNA-amplified fragment length polymorphism (AFLP) analysis (25), (v) serial analysis of gene expression (SAGE) (26), (vi) massive parallel signature sequencing (MPSS) (27), (vii) cDNA macroarray analysis (28), (viii) cDNA microarray analysis, and (ix) oligonucleotide microarray analysis. These methods have several advantages and disadvantages (7, 8, 29).

RNA gel-blot (northern) analysis is an established and reliable method, which allows accurate quantification of specific transcripts, but it cannot be applied for genome-scale expression analysis.

Differential display uses low stringency PCR, a combinatorial primer set, and gel electrophoresis to amplify and visualize larger populations of cDNAs representing mRNA populations of interest. Differential display is a relatively cheap and simple means of screening for differentially expressed genes, and is particularly useful when the availability of RNA is limited. However, this technique requires a large number of reactions to achieve maximal coverage of all active transcripts and suffers from an output that is not quantitative and identified sequences are often difficult to clone and confirm (8).

Quantitative real-time PCR (QRT-PCR) has been demonstrated to generate robust, quantitative expression data for single genes and this method offers rapid and reproducible results (30). One of the major advantages of real-time PCR is its broad dynamic range with which one can precisely quantify transcript concentrations over more than at least five orders of magnitude (31). RNA gel-blot analysis and real-time PCR are often used to confirm differential expression of genes detected by DNA microarray analysis.

The principles of AFLP are applied to cDNA templates in cDNA-AFLP analysis, which has been used to identify differentially expressed genes involved in a variety of plant processes. This technique offers several advantages over traditional approaches. Of particular importance is the fact that poorly characterized genomes can be investigated in a high-throughput manner. Because the stringency of cDNA-AFLP PCR reactions is quite high (which is not the case with differential display) the fidelity of the cDNA-AFLP system allows much greater confidence in acquired data and differences in the intensities of amplified products can be informative (25). As with the other profiling methods described here, the sensitivity of cDNA-AFLP is only limited by the ability of cDNA libraries to capture low-abundance transcripts (7).

SAGE is based on the capture and sequence analysis of a short region close to the 3′ end of each cDNA in the sample (26), and it is a quantitative or digital method of gene expression analysis like EST sequencing. SAGE is time-consuming and requires an extensive foundation of sequence information. Variations in amplification efficiency between ditags may lead to distorted results (29).

MPSS, developed and commercialized by Lynx Therapeutics (Hayward, CA), is based on methods to clone individual cDNA molecules on microbeads and sequence, in parallel, short tags or signatures from these cDNAs (27). The final output of MPSS is a set of abundances for thousands of distinct 17- or 20-base signatures, most of which uniquely identify a particular transcript. The parallel sequencing method produces millions of MPSS signatures in only a few weeks, but the technology is sufficiently complex, and unlike SAGE, it cannot be performed in individual laboratories (8).

cDNA macroarray technology allows parallel and comparative analysis of the expression of thousands of genes (28) as well as DNA microarray technology. The cDNA macroarrays and the cDNA microarrays differ primarily in the type of solid support immobilized: that is, the macroarrays use a membrane-based matrix while the microarrays use a glass or plastic slide. In most cases, macroarray targets are radioactively labeled. The cDNA

macroarray analysis is also less expensive than the oligonucleotide array analysis. However, it takes a lot of work, such as printing on many membranes, to prepare a whole genome macroarray. A number of reviews have described how cDNA macroarrays are used and the advantages and disadvantages of cDNA macroarray analysis (32, 33).

3.3 DNA microarrays: cDNA and oligonucleotide microarrays

Compared with other methods, the DNA microarray technology has the following important advantages: (i) it can measure thousands of different mRNA transcripts in parallel, (ii) it provides semi-quantitative data, and (iii) it is sensitive enough to detect low-abundance transcripts that are represented on a given array. In several organisms whose complete genome sequence is available, DNA microarrays enable the monitoring of whole-genome gene expression in a single experiment.

cDNA microarrays and oligonucleotide microarrays each have several advantages and disadvantages. One advantage of cDNA microarrays is that they can be prepared directly from the isolated cDNA clones. Once a set of corresponding PCR products has been generated, microarrays can be created in multiple versions containing the entire set of cDNA sequences, resulting in large-scale arrays for identification of differentially expressed genes of interest or small-scale arrays suitable for specific research applications. The most important advantage is that the cDNA microarray is less expensive to make than a single oligonucleotide array. However, cross-hybridization between homologous sequences is problematic for cDNA microarrays.

One advantage of oligonucleotide arrays is that oligonucleotides can be synthesized either in plates or directly on solid surfaces (*in situ* synthesis), making it easier to prepare the DNA probes than for cDNA microarrays. Also, the probes in an oligonucleotide array can be designed to represent unique gene sequences such that cross-hybridization between related gene sequences is minimized to a degree dependent upon the completeness of available sequence information. One primary disadvantage of oligo-based analysis is that oligonucleotide microarrays are expensive.

There are two types of oligonucleotide microarrays: one is the direct synthesis of oligonucleotides on a solid (34). Microarrays of this sort are produced by Affymetrix and Nimble Gen. The other is the immobilization of pre-synthesized oligonucleotides, for example produced by Agilent Technologies (Palo Alto, CA) and other companies. The spotted oligo microarrays are typically comprised of a single 50- to 70-mer oligonucleotide for each gene.

3.4 cDNA clones and their application for cDNA microarray analysis

Currently there are nearly 23 million expressed sequence tags (ESTs) in the NCBI public collection as of August 2004, about 4 million of which derive from plants (http://www.ncbi.nlm.nih.gov/dbEST/). With many

large-scale EST sequencing projects in progress and new projects being initiated, the number of ESTs in the public domain will continue to increase in the coming years. cDNA clones are a useful tool for expression profiling studies, because cDNA microarrays can be prepared directly from the isolated clones.

Altogether, 19 RIKEN *Arabidopsis* full-length (RAFL) cDNA libraries from plants grown under different conditions were constructed as reported previously (4, 35, 36). We have isolated 245 946 RAFL cDNA clones and they are clustered into about 18 000 non-redundant cDNA groups, covering about 70% of predicted genes (36; Seki et al., unpublished results).

The RAFL cDNAs were used for microarray analysis of expression profiles of *Arabidopsis* genes (5, 6) under various stress conditions, such as drought, cold and high salinity (9, 10), various treatment conditions, such as abscisic acid (ABA) (37), rehydration treatment after dehydration (18), ethylene (16), jasmonic acid (JA) (16), salicylic acid (SA) (16), reactive oxygen species-(ROS-) inducing compounds such as paraquat and rose bengal (16), UV-C (16), and inoculation with a pathogen (16, 17). We have also studied the expression profiles in various mutants and transgenic plants (9, 16, 38, 39). These studies have shown that cDNA microarray analysis is useful for analyzing the expression pattern of plant genes under various stress and hormone treatments, to identify target genes of transcription factors involved in stress or hormone signal transduction pathways, and to identify potential *cis*-acting DNA elements by combining the expression data with genomic sequence data. cDNA microarrays can be used in closely related species with about 90% sequence homologies of coding regions, such as rice and barley (40), and *Arabidopsis* and *Thellungiella halophila* (41).

In this chapter, we describe the protocol of our RAFL cDNA microarray analysis.

References

1. Eisen MB and Brown PO (1999) DNA arrays for analysis of gene expression. *Methods Enzymol* **303:** 179–205.
2. Bowtell D and Sambrook J (2002) *DNA Microarrays: A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Richmond T and Somerville S (2000) Chasing the dream: plant EST microarrays. *Curr Opin Plant Biol* **3:** 108–116.
4. Seki M, Narusaka M, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y and Shinozaki K (2001) *Arabidopsis* encyclopedia using full-length cDNAs and its application. *Plant Physiol Biochem* **39:** 211–220.
5. Seki M, Kamei A, Satou M, Sakurai T, Fujita M, Oono Y, Yamaguchi-Shinozaki K and Shinozaki K (2003) Transcriptome analysis in abiotic stress conditions in Higher Plants. *Topics Curr Genet* **4**: 271–295.
6. Seki M, Satou M, Sakurai T, *et al*. (2004) Expression profiling under abiotic stress conditions using RIKEN *Arabidopsis* full-length (RAFL) cDNA microarray. *J Exp Bot* **55:** 213–223.
7. Alba R, Fei Z, Payton P, *et al*. (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *Plant J* **39:** 697–714.
8. Meyers BC, Galbraith DW, Nelson T and Agrawal V (2004) Methods for transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiol* **135:** 637–652.

9. Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y and Shinozaki K (2001) Monitoring the expression pattern of 1300 *Arabidopsis* genes under drought and cold stresses using a full-length cDNA microarray. *Plant Cell* **13:** 61–72.
10. Seki M, Narusaka M, Ishida J, *et al.* (2002) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold, and high-salinity stresses using a full-length cDNA microarray. *Plant J* **31:** 279–292.
11. Fowler S and Thomashow MF (2002) *Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *Plant Cell* **14:** 1675–1690.
12. Ma L, Jinming L, Qu L, Hager J, Chen Z, Zhao H and Deng XW (2001) Light control of *Arabidopsis* development entails coordinated regulation of genome expression and cellular pathways. *Plant Cell* **13:** 2589–2607.
13. Schaffer R, Landgraf J, Accerbi M, Simon V, Larson M and Wisman E (2001) Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. *Plant Cell* **13:**113–123.
14. Fedorova M, Mortel JVD, Matsumoto PA, Cho J, Town CD, VandenBosch KA, Gantt JS and Vance CP (2002) Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*. *Plant Physiol* **130:** 519–537.
15. Maleck K, Levine A, Eulgem T, Morgan A, Schmid J, Lawton K, Dangl J and Dietrich R (2000) The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nature Genet* **26:** 403–410.
16. Narusaka Y, Narusaka M, Seki M, *et al.* (2003) Monitoring the expression profiles and classification of *Arabidopsis* genes induced by *Alternaria brassicicola* attack using a cDNA microarray. *Plant Cell Physiol* **44:** 377–387.
17. Narusaka Y, Narusaka M, Park P, *et al.* (2004) RCH1, a locus in *Arabidopsis* that confers resistance to the hemibiotrophic fungal pathogen *Colletotrichum higginsianum*. *Mol Plant Microbe Interact* **17:** 749–762.
18. Oono Y, Seki M, Nanjo T, *et al.* (2003) Monitoring expression profiles of *Arabidopsis* gene expression during rehydration process after dehydration using ca. 7000 full-length cDNA microarray. *Plant J* **34:** 868–887.
19. Aharoni A, Keizer LC, Bouwmeester HJ, *et al.* (2000) Identification of the *SAAT* gene involved in strawberry flavor biogenesis by use of DNA microarrays. *Plant Cell* **12:** 647–662.
20. Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complimentary DNA microarray. *Science* **270:** 467–470.
21. Lipshutz RJ, Fodor SPA, Gingeras TR and Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nature Genet* **21:** 20–24.
22. Aharoni A and Vorst O (2001) DNA microarrays for functional plant genomics. *Plant Mol Biol* **48:** 99–118.
23. Donson J, Fang Y, Espiritu-Santo G, Xing W, Salazar A, Miyamoto S, Armendarez V and Volkmuth W (2002) Comprehensive gene expression analysis by transcript profiling. *Plant Mol Biol* **48:** 75–97.
24. Liang P and Pardee A (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257:** 967–971.
25. Bachem C, Van der Hoeven R, de Bruijn S, Vreugdenhil D, Zabeau M and Visser R (1996) Visualisation of differential gene expression using a novel method of RNA finger-printing based on AFLP: analysis of gene expression during potato tuber development. *Plant J* **9:** 745–753.
26. Velculescu V, Zhang L, Vogelstein B and Kinzler K (1995) Serial analysis of gene expression. *Science* **270:** 484–487.
27. Brenner S, Johnson M, Bridgham J, *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnol* **18:** 630–634.

28. Sasaki Y, Asamizu E, Shibata D, *et al.* (2001) Monitoring of methyl jasmonate-responsive genes in *Arabidopsis* by cDNA macroarray: self-activation of jasmonic acid biosynthesis and crosstalk with other phytohormone signaling pathways. *DNA Res* **8:** 153–161.
29. Green CD, Simons JF, Taillon BE and Lewin DA (2001) Open systems: panoramic views of gene expression. *J Immunol Methods* **250:** 67–79.
30. Klein D (2002) Quantification using real-time PCR technology: applications and limitations. *Trends Mol Med* **8:** 257–260.
31. Heid CA, Stevens J, Livak KJ and Williams PM (1996) Real time quantitative PCR. *Genome Res* **6:** 986–994.
32. Freeman WM, Robertson DJ and Vrana KE (2000) Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques* **29:** 1042–1055.
33. Soldatov AV, Nabirochkina EN, Georgieva SG and Eickhoff H (2001) Adjustment of transfer tools for the production of micro- and macroarrays. *Biotechniques* **31:** 848–854.
34. Lockhart DJ, Dong H, Byrne MC, *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol* **13:** 1675–1680.
35. Seki M, Carninci P, Nishiyama Y, Hayashizaki Y and Shinozaki K (1998) High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper. *Plant J* **15:** 707–720.
36. Seki M, Narusaka M, Kamiya A, *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296:** 141–145.
37. Seki M, Ishida J, Narusaka M, *et al.* (2002) Monitoring the expression pattern of ca. 7000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray. *Funct Integr Genomics* **2:** 282–291.
38. Abe H, Urao T, Ito T, Seki M, Shinozaki K and Yamaguchi-Shinozaki K (2003) *Arabidopsis* AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15:** 63–78.
39. Maruyama K, Sakuma Y, Kasuga M, *et al.* (2004) Identification of cold-inducible downstream genes of the *Arabidopsis* DREB1A/CBF3 transcriptional factor using two microarray systems. *Plant J* **38:** 982–993.
40. Negishi T, Nakanishi H, Yazaki J, *et al.* (2002) cDNA microarray analysis of gene expression during Fe-deficiency stress in barley suggests that polar transport of vesicles is implicated in phytosiderophore secretion in Fe-deficient barley roots. *Plant J* **30:** 83–94
41. Taji T, Seki M, Satou M, Sakurai T, Kobayashi M, Ishiyama K, Narusaka Y, Narusaka M, Zhu JK and Shinozaki K (2004) Comparative genomics in salt tolerance between *Arabidopsis* and *Arabidopsis*-related halophyte salt cress using *Arabidopsis* microarray. *Plant Physiol* **135:** 1697–1709

Protocols

CONTENTS

The flowchart of the RAFL cDNA microarray analysis is shown in *Figure 3.1*.

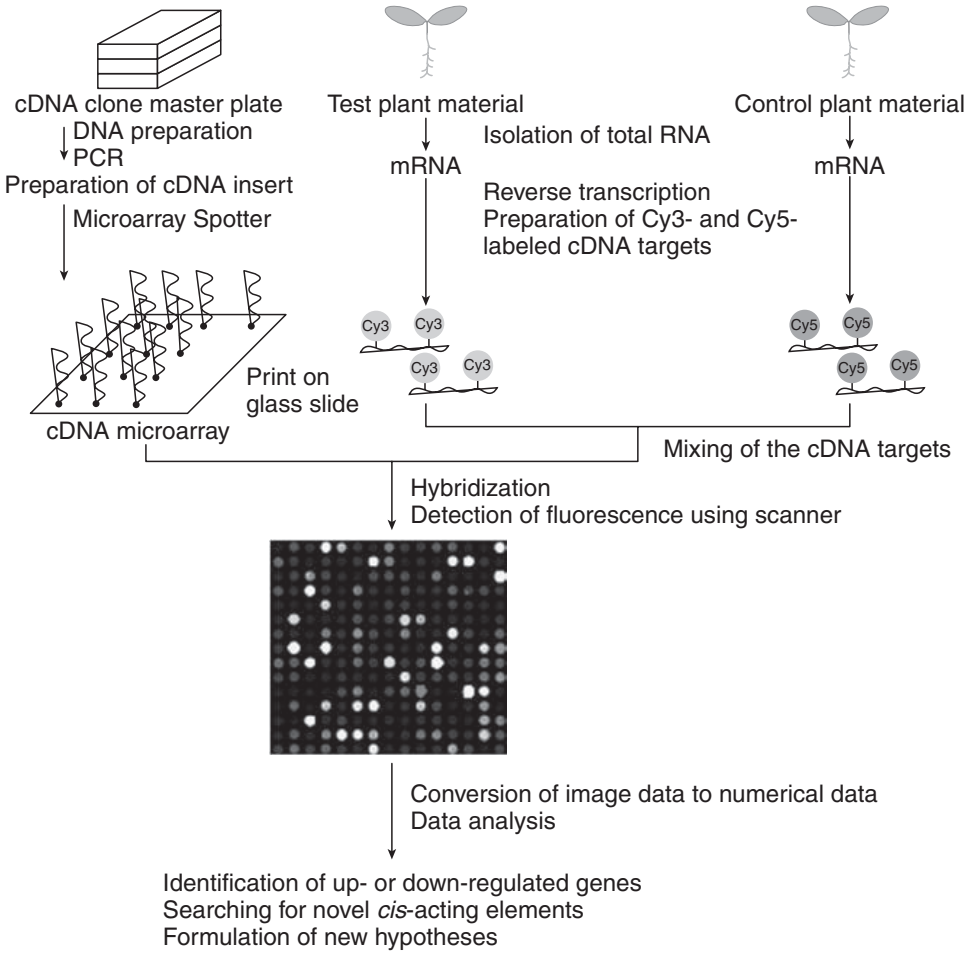


Figure 3.1.

Overall strategy for cDNA microarray analysis

Protocol 3.1: Preparation of cDNA microarrays

CDNA CLONES

We prepared a 7K RAFL cDNA microarray (10) to study the expression profiles of *Arabidopsis* genes under various conditions. The 7K RAFL cDNA microarray consists of about 7000 RAFL cDNA clones isolated from full-length cDNA libraries. It includes drought- and cold-inducible genes, *responsive to dehydration* (*rd*) and *early responsive to dehydration* (*erd*). A PCR-amplified fragment from a λ control template (TX803; Takara, Kyoto, Japan) is used as an external control, and two cDNAs derived from the mouse *nicotinic acetylcholine receptor epsilon-subunit* (nAChRE) gene and the mouse *glucocorticoid receptor homolog* gene are included as negative controls.

SEQUENCE ANALYSIS OF CDNA CLONE INSERTS

1. Extract the plasmid DNA with DNA extraction instruments of Kurabo (PI–100; KURABO, Tokyo, Japan) or Beckman (Biomek2000; Beckman Coulter, Tokyo, Japan).
2. Determine the DNA sequences using the dye terminator cycle sequencing method with a DNA sequencer (ABI PRISM 3700; Perkin-Elmer Applied Biosystems, Foster City, CA). Examine sequence homologies with the GenBank/EMBL database using the BLAST program.

AMPLIFICATION OF CDNA INSERTS

1. Amplify the inserts of cDNA clones by PCR using primers complementary to vector sequences flanking both sides of the cDNA insert, as described previously (9). Add plasmid templates (1 to 10 ng) to 50 μl of a PCR mixture containing 0.25 mM of each nucleotide, 0.2 μM of each primer, 1 × Ex Taq buffer (Takara, Kyoto, Japan), and 1.25 units of Ex Taq polymerase (Takara, Kyoto, Japan). Perform the PCR as follows: at 95°C for 3 min; 35 cycles at 95°C for 30 s, 60°C for 1 min, and 72°C for 3 min; and at 72°C for 3 min.
2. Precipitate the PCR products in isopropanol and resuspend the DNA in 5 μl of TE to a final concentration of about 2 $\mu g \mu l^{-1}$.

3. Check one aliquot of each reaction product on a 0.7% agarose gel to confirm amplification quality and quantity.
4. Add 2 μl of 2 × polymer (Fuji Photo Film Co., Kanagawa, Japan) and 4 μl of dimethyl sulfoxide (DMSO) (Kishida Chemical Co., Osaka, Japan) into 2 μl of DNA solution in 96-well plates. Transfer the mixture into 384-well plates and mix at least 10 times using an automatic dispenser (model EDS-384S; Biotech Co., Ltd., Tokyo, Japan).

PRINTING ON GLASS SLIDES

1. Array the PCR products from 384-well microtiter plates onto micro-slides (model Super Aldehyde substrates; Telechem International Inc., Sunnyvale, CA) using a microarray stamping machine (model SPBIO2000; Hitachi Software Engineering Co., Ltd., Tokyo, Japan). The tip loads 2 μl of PCR products (500 to 1000 ng μl^{-1}) from 384-well microtiter plates and deposits 0.5 nl per slide on 48 slides with spacing of 300 μm in our system (10).
2. Postprocess the slides according to the manufacturer's protocol (Telechem International Inc., Sunnyvale, CA). Dry the slides for more than 12 h in a desiccator (relative humidity <30%). This period may facilitate binding of the printed DNA and slide coating. Irradiate the slides with 65 mJ UV to obtain cross-linked DNA.
3. Rock them in 0.2% SDS for 2 min twice and then rock in distilled water for 2 min twice vigorously.
4. Transfer the slide racks into a chamber containing boiling water and leave for 2 min. Remove the slide racks to a clean glass container and leave them at room temperature for 5 min to cool. Pour the blocking solution, containing 1 g of sodium borohydride, 300 ml of phosphate-buffered saline (PBS; Invitrogen, Carlsbad, CA), and 90 ml of 100% ethanol, into the glass chamber.
5. Shake the slide racks gently for 5 min, transfer three times into a new chamber containing 0.2% SDS and shake gently for 1 min.
6. Transfer them into a chamber containing distilled water, shake gently for 1 min, and dry by centrifugation for 20 min to remove any residual solution from the slides. Store the slides in a desiccator.

Protocol 3.2: Preparation of cDNA targets

ISOLATION OF TOTAL RNA

Total RNA was prepared using TRIZOL Reagent (Invitrogen, Carlsbad, CA) in the following way.

1. Add 20–30 ml of Trizol reagent to a 50-ml Falcon tube and incubate the tube at 50°C.
2. Homogenize 3–4 g plant tissues in liquid nitrogen.
3. Transfer the plant tissues using pre-cooled spoons.
4. Vortex for 30 s.
5. Incubate at 50°C for 2 min, then at RT for 5 min and keep it on ice for 1 h.
6. Add 4–6 ml of chloroform and shake for 1min vigorously.
7. Leave for 2 min at room temperature.
8. Centrifuge at 10 000 *g* at 4°C for 20 min.
9. Transfer the supernatant to new tube, add 1/3 volume of 8 M LiCl, mix and leave at
   –80°C for 1 h. Thaw at 50°C for a while and then thaw completely (note: do not leave at 50°C longer than needed).
10. Centrifuge at 10 000 *g* at 4°C for 30 min.
11. Decant the supernatant, wash the pellet with 2 ml of 75% ethanol and dry.
12. Dissolve the pellet in 1.3 ml of DEPC-treated distilled water.
13. Thaw at 50°C for a while and then thaw completely (note: do not leave at 50°C longer than needed).
14. Centrifuge at 10 000 *g* at 4°C for 3 min. Divide it into two tubes (each 650 μl).
15. Add 65 μl of 3 M sodium acetate and 650 μl isopropanol. Mix and leave for 5 min.
16. Centrifuge at 15 000 *g* at 4°C for 10 min.
17. Decant the supernatant, wash the pellet with 75% ethanol and dry.
18. Dissolve in 400 μl of DEPC-treated distilled water and store at –80°C.

ISOLATION OF mRNA

mRNA was prepared using the protocol according to the mRNA isolation kit (MACS, Miltenyi Biotec, Bergisch Gladbach, Germany) as follows.

1. Incubate the elution buffer at 65°C, leave the lysis/binding buffer and wash buffer at room temperature.
2. Add ca. 1 mg total RNA sample to the Eppendorf tube, incubate the tubes at 65°C for 3 min and keep on ice.
3. Dilute up to 1 mg total RNA with at least 1 volume of lysis/binding buffer. The final volume should be 0.5–5 ml.
4. Add 25 μl of oligo (dT) microbeads per 100 μg of the total RNA sample, and mix by pipetting. (Note: do not allow to bubble during the mixing).
5. Place a MACS Column Type M in the magnetic field of an appropriate MACS separator.
6. Prepare column by rinsing with 250 μl of lysis/binding buffer and let buffer run through.
7. Apply total RNA sample on top of the column matrix. Let the solution pass through. Magnetically labeled mRNA is retained in the column.
8. Rinse the column with 1 × 250 μl of lysis/binding buffer.
9. Rinse the column with 4 × 250 μl of wash buffer.
10. Apply 200 μl of pre-heated elution buffer on top of the column. mRNA is eluted by gravity. Typically, the third to sixth drop will contain around 90% of the isolated mRNA.
11. Add 1/10 volume of 3 M sodium acetate and 3 volume of ethanol to the eluted mRNA sample, and leave at –80°C for 1 h.
12. Centrifuge at 15 000 *g* at 4°C for 10 min.
13. Decant the supernatant, wash the pellet with 75% ethanol and dry.
14. Dissolve in DEPC-treated distilled water and store at –80°C. The final mRNA concentration should be over 200 ng μl^{-1}.

PREPARATION OF CY-DYE-LABELED CDNA TARGETS

1. Prepare 7-μl solution containing 1 μg of denatured poly(A)$^+$ RNA with 1 ng of λ poly(A)$^+$ RNA-A (TX802; Takara, Kyoto, Japan) for external control, 50 ng μl^{-1} oligo-(dT) 12–18 mer (Invitrogen, Carlsbad, CA). Incubate the annealing reaction solution for 5 min at 70°C, and then at 42°C for 1–2 min.

2. Add 8 μl of the buffer mixture containing 4 μl of 5 × Superscript first-strand buffer (250 mM Tris-HCl, pH 8.3, 375 mM KCl, and 15 mM $MgCl_2$; Invitrogen, Carlsbad, CA), 2 μl of 0.1 M DTT and 2 μl of dNTP mixture (dATP, dCTP, dGTP, each at 5 mM, and dTTP at 2 mM), 2 μl of 1 mM Cy3-dUTP or Cy5-dUTP(Amersham Pharmacia, Piscataway, NJ), 2.5 μl of 40 units μl^{-1} RNase inhibitor (SIN–101; Toyobo, Osaka, Japan) and 1 μl of 200 units μl^{-1} Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA) to the annealing mixture.
3. Following incubation at 42°C for 35 min, add 100 units of Superscript II reverse transcriptase.
4. Incubate the reaction sample for an additional 35 min.
5. Following addition of 5 μl of 0.5 M EDTA, 10 μl of 1 N sodium hydroxide, and 20 μl of distilled water to stop the reaction and to degrade the template, incubate them for 1 h at 65°C.
6. Neutralize the solution with 25 μl of 1 M Tris-HCl (pH 7.5).
7. Combine the reaction products of two samples (one with Cy3 labeling and the other with Cy5 labeling).
8. Place the samples in a Microcon YM-30 microconcentrator (Millipore, Bedford, MA).
9. Add 250 μl of TE buffer, spin for 17 min in a benchtop microcentrifuge at a high speed to a volume of 10 μl, and discard the flow-through product. Repeat this step four times.
10. Collect the targets by inverting the filter and spinning for 5 min.
11. Add several microliters of distilled water to the Microcon.
12. Invert the filter, spin, and add distilled water so that the final volume of the collected targets is 18 μl.
13. Add 5.1 μl of 20 × SSC, 2.5 μl of 2 μg μl^{-1} yeast tRNA and 4.8 μl of 2% SDS to the probes.
14. Denature the target samples by placing them in a 100°C heat block for 2 min, leave at room temperature for 5 min, and then use for hybridization.

Protocol 3.3: Microarray hybridization and scanning

1. Spin the target sample for 1 min at 15 000 *g* to pellet any particulate matter. Pipette the target solution close to the end of the microarray slide.
2. Place a cover slip over the probe in such a way as to avoid the formation of bubbles.
3. Place the slides in a sealed hybridization cassette (Telechem International Inc., Sunnyvale, CA) and submerge in a 65°C water bath for 16–20 h.
4. After hybridization, wash the slides in 2 × SSC, 0.03% SDS for 2 min, then in 1 × SSC for 2 min, and finally in 0.05 × SSC for 2 min.
5. Centrifuge (1 min at 2500 *g*) the slides immediately to dry.
6. Scan the slides with a ScanArray 4000 (GSI Lumonics, Oxnard, CA) as described previously (9).

Protocol 3.4: Data analysis

Image analysis and signal quantification were performed with QuantArray version 2.0 (GSI Lumonics, Oxnard, CA). Background fluorescence was calculated from the fluorescence signal of the negative control genes (the mouse nicotinic acetylcholine receptor epsilon-subunit gene and the mouse glucocorticoid receptor homolog gene) in our RAFL cDNA microarray analysis. To remove the systematic variation, it is necessary to normalize the microarray data (see Chapter 17). Gene-clustering analysis (see also Chapter 19) was performed with Genespring (Silicon Genetics, San Carlos, CA).

4 Identification of gene expression patterns for a molecular diagnosis of kidney tumors

Holger Sültmann, Andreas Buneß, Markus Ruschhaupt, Wolfgang Huber, Ruprecht Kuner, Bastian Gunawan, Laszlo Füzesí and Annemarie Poustka

4.1 Introduction

Over the past 20 years, significant success in the therapy of certain cancer types has given rise to the hope that cancer will soon be curable. However, it is becoming evident that many tumor types, which were previously regarded as homogeneous disease entities, are composed of different subtypes with varying patient prognosis and survival rates. These findings may explain the varying degrees of success in cancer treatment. Since classical pathological parameters are often not sufficient to identify tumor subtypes, novel markers for tumor diagnosis and new targets for differential tumor therapies are required.

Adult renal cell carcinoma (RCC) is one of the 10 most common human malignancies in developed countries. Its global incidence has been increasing continuously over the past 30 years (1). Males are afflicted twice as often compared to females, and several genetic factors, such as the von Hippel Lindau (*VHL*) gene are known to play a role in a subset of RCC. Apart from these typical markers, other genes known to be involved in RCC include *VEGF* (2, 3), *EGFR* (4, 5), *TGFA* (6), *c-myc* proto-oncogene (7, 8) and *VIM* (9). RCC is divided into clear cell (ccRCC; 80% of all cases), papillary (pRCC, 10%), chromophobe (chRCC, 5%), and several other rare types. Although the histopathological diagnosis of kidney cancer is well established in the clinical routine, the molecular basis for the distinction of RCC types is poorly understood.

New technologies to examine tissue samples taken from cancer patients on a large scale have been developed in the genome projects. In particular, DNA microarrays have been applied to various kinds of human tumors (10–18) in order to find new cancer subclasses and to decipher their molecular basis. However, there is an important issue associated with the discovery of gene expression patterns to be used for diagnostic purposes. The number of available tumor samples usually is much smaller (10–300)

than the number of available probes (10 000–50 000). Therefore, special attention should be devoted to avoiding over-interpretation of microarray data. Analysis of differential gene expression has to account for multiple testing, and classification methods must address the problem of overfitting (see also Chapter 19). We propose the usage of various classification methods for microarray data analysis in order to reduce the risk of over-interpretation. We constructed RCC-specific cDNA microarrays encompassing 4207 cDNA clones and hybridized these with labeled cDNA derived from tumor samples of the three major RCC types. By using the microarray data of 35 RCC samples, we identified a set of 18 genes that are potentially useful for diagnosis and therapy of kidney tumors.

4.2 Experimental design

Thirty-five RCC samples (13 ccRCC, 13 pRCC, and 9 chRCC) were labeled individually with Cy3. A Cy5-labeled sample pool of 28 tumor RNA samples (15 ccRCC, 8 pRCC, and 5 chRCC) was used as a common reference for all hybridizations. Both Cy3- and Cy5-labeled samples were hybridized competitively against the cDNA probes on the microarrays. The pixel intensities for each spot were quantified, and data were normalized as described (19). The generalized log ratios were used for tumor classification.

4.3 Molecular classification of kidney tumors

For analysis, we chose a broad range of classification approaches to identify and address the potential dependencies between classification approaches and their prediction abilities (see also Chapter 19). To this end, we selected the frequently used prediction analysis for microarrays (PAM) (21), support vector machines (22) and random forest (23) algorithms. The rationale of a classification is as follows: A set of microarray data (the training set) is divided into two or more classes (here, the three RCC types). The goal is to build a classifier (a method that is able to predict the classes in an independent data set). Ideally, the performance of the classifier is evaluated by testing its prediction ability on an independent test set for which the classes are also known. In practice however, due to the small number of available samples, cross-validation is performed, where the samples are randomly divided into equally sized subsets. In each step, one subset is left aside, the classifier is built on the remaining samples, and the classes of the left-out samples are predicted and compared with the actual classes. In our data set, the prediction ability was assessed by 10-fold cross-validation, that is in each step 90% of the samples are used as a training set and 10% as a test set. The whole procedure was repeated 20 times.

Microarrays can detect small fold changes in pairwise comparisons. However, small fold changes are often not useful for diagnostic routines. Hence, we applied a filter in order to select for genes whose expression levels changed considerably between two tumor types. To avoid overfitting, our gene filtering was applied in each cross-validation step on the training set. Therefore, we calculated the mean expression value of every gene/EST and every tumor type and selected all genes whose group means' fold change was >2 in any pairwise group comparison. Depending on the samples in

the training set, approximately 440 of the 4207 genes on the microarrays fulfilled this criterion.

The microarray data of 35 kidney tumor samples were used for the molecular classification of the tumor subgroups. All classification methods (21–23) showed a high prediction ability and gave similar results: Each method correctly classified at least 32 out of 35 samples, 31 in at least 95% and one in at least 80% of the repetitions (not shown). One ccRCC sample (no. 14) was consistently misclassified by all methods, two others were misclassified by random forest and SVM in at least 60% of the repetitions (not shown). In PAM (*Plate 1*), only sample 14 was "incorrectly" classified while all other samples corresponded to the histopathological diagnosis. Thus, the misclassification rate was less than 3%. The tumor samples that did not match the histopathological classification were reanalyzed by pathology. The initial diagnosis was confirmed, suggesting that there was no error in the clinical diagnosis of these tumors.

4.4 Building a classifier for kidney tumor diagnosis

Knowing that the prediction ability of the various classifiers was high, we applied the same methods to all samples in order to build a final classifier that could be used for diagnostic purposes. In the following, we will concentrate on the classifier of the PAM method.

The classifier of the PAM method was based on 18 genes. We visualized the generalized log ratios of these genes by different colors (*Plate 2*). Among the 18 genes were *GAPD, CXCL14, ADFP, SPP1*, and *CD9*. Several of these are known key players in cancer and differentiation: CXCL14 is a cytokine that is frequently down-regulated in tumors (24) but up-regulated in inflammatory cells of the tumor microenvironment (25), ADFP is a protein involved in cell differentiation and has been found to be highly overexpressed in ccRCC (26). The osteopontin gene (*SPP1*) is a target for *TP53* (27) and a lead marker for colon cancer progression (28). Two members of the galactose-binding lectin family (*LGALS3* and *LGALS9*) are known to be involved in tumorigenic processes. Two genes (*CD9* and *TSPAN1*) coded for cell surface proteins of the tetraspanin family, which mediate signal transduction events in cell development, activation, growth and motility. This indicates that the three kidney tumor types can be distinguished by the activation status of specific biological processes. Two of the classifier genes (*GAPD* and *LGALS3*) were represented by independent clones. The gene expression patterns of these clones were very similar, suggesting that different spots on the microarrays give highly reproducible results.

The relative gene expression pattern suggested that the chRCC tumors can be clearly distinguished from the ccRCC and pRCC. With the exception of *GAPD*, there is an almost inverse relationship of expression patterns of all 18 genes between these two groups. In contrast, the distinction between ccRCC and pRCC tumors is not obvious and appears to rely primarily on the expression of the *GAPD* gene, which is up-regulated in ccRCC but down-regulated in pRCC. This may reflect different metabolic activities between these two tumor types. This finding is consistent with a previous report about high levels of glycolytic enzymes in ccRCC samples (29). The importance of *GAPD* as a part of the classifier was supported by its low

expression in tumor 14, which was consistently misclassified in all algorithms. Apart from *GADP*, ccRCC and pRCC tumor types differ mainly by the expression of crystallin alpha B (*CRYAB*), osteopontin (*SPP1*) and tetraspanin 1 (*TSPAN1*).

4.5 Summary

Gene expression profiling provides a potent universal tool for improved molecular diagnosis and prognostic evaluation. This tool should be carefully handled in order to avoid data misinterpretation, wrong conclusions and generalizations. Here, we have demonstrated a way to identify a set of 18 genes that discriminate between the three major types of RCC, providing candidates for a molecular differential diagnosis.

The establishment of a diagnostic tool requires a robust platform, which is easy to handle, highly sensitive and specific. Whether or not microarrays will eventually enter the field of diagnostic applications remains open. However, microarrays are perfectly able to detect gene expression patterns that can subsequently be exploited for diagnosis using different platform technologies. For example, it is conceivable that array-based analyses lead to the identification of a set of highly diagnostic genes that are then used in diagnostic routines using other (e.g. RT-PCR-based) methods. To achieve this the number of target genes needs to be reduced, and highly specific gene expression patterns between the tissue types need to be identified.

We are aware of the fact that the current histopathology-based kidney tumor diagnosis is robust, and molecular methods are not likely to replace the clinical routine diagnosis of kidney cancer in the near future. However, in contrast to routine histopathology, microarray data yield important insights into the molecular differences between kidney tumors: Gene expression patterns can not only confirm current diagnostic procedures, but at the same time reveal further highly promising target genes for a more specific therapy directed against the different kidney tumor types. The prerequisites to achieve this aim are the "druggability" of the targets and the availability of highly specific compounds with low degrees of side effects. To this end, more effort is urgently needed to systematically exploit the huge amount of gene expression data for the therapy of cancer patients. The application of genome-wide microarrays for RCC studies may yield additional genes and ESTs that could be used for the identification of previously unrecognized kidney tumor subgroups and novel therapeutic targets.

Data accessibility

The microarray data reported here were submitted to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) and assigned the accession number E-DKFZ–1.

Acknowledgements

This work was supported by a grant of the German Federal Ministry for Education and Research (BMBF) in the German Genome Project (DHGP), grant number 01KW9911/9. The experiments were performed in accor-

dance with the German ethical requirements and were approved by the ethics commission of the University of Göttingen (3/6/02).

References

1. Fleming S (1998) Genetics of kidney tumours. *Forum (Genova)* **8**: 176–184.
2. Brieger J, Weidt EJ, Schirmacher P, Storkel S, Huber C and Decker HJ (1999) Inverse regulation of vascular endothelial growth factor and VHL tumor suppressor gene in sporadic renal cell carcinomas is correlated with vascular growth: an in vivo study on 29 tumors. *J Mol Med* **77**: 505–510.
3. Takahashi A, Sasaki H, Kim SJ, Kakizoe T, Miyao N, Sugimura T, Terada M and Tsukamoto T (1999) Identification of receptor genes in renal cell carcinoma associated with angiogenesis by differential hybridization technique. *Biochem Biophys Res Commun* **257**: 855–859.
4. Ishikawa J, Maeda S, Umezu K, Sugiyama T and Kamidono S (1990) Amplification and overexpression of the epidermal growth factor receptor gene in human renal-cell carcinoma. *Int J Cancer* 45: 1018–1021.
5. Moch H, Sauter G, Gasser TC, Bubendorf L, Richter J, Presti JC, Jr Waldman FM and Mihatsch MJ (1998) EGF-r gene copy number changes in renal cell carcinoma detected by fluorescence *in situ* hybridization. *J Pathol* **184**: 424–429.
6. Lager DJ, Slagel DD and Palechek PL (1994) The expression of epidermal growth factor receptor and transforming growth factor alpha in renal cell carcinoma. *Mod Pathol* **7**: 544–548.
7. Drabkin HA, Bradley C, Hart I, Bleskan J, Li FP and Patterson D (1985) Translocation of *c-myc* in the hereditary renal cell carcinoma associated with a t(3;8)(p14.2;q24.13) chromosomal translocation. *Proc Natl Acad Sci USA* **82**: 6980–6984.
8. Yao M, Shuin T, Misaki H and Kubota Y (1988) Enhanced expression of *c-myc* and epidermal growth factor receptor (*C-erbB–1*) genes in primary human renal cancer. *Cancer Res* **48**:6753–6757.
9. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP and Sauter G (1999) High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol* **154**: 981–986.
10. Perou CM, Sørlie T, Eisen MB, *et al*. (2000) Molecular portraits of human breast tumours. *Nature* **406**: 747–752.
11. Van't Veer LJ and De Jong D (2002) The microarray way to tailored cancer treatment. *Nat Med* 8: 13–14.
12. Beer DG, Kardia SL, Huang CC, *et al*. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* **8**: 816–824.
13. Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM and Isaacs WB (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* **61**: 4683–4688.
14. Takahashi M, Rhodes DR, Furge KA, Kanayama H, Kagawa S, Haab BB and Teh BT (2001) Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. *Proc Natl Acad Sci USA* **98**: 9754–9759.
15. Rickman DS, Bobek MP, Misek DE, Kuick R, Blaivas M, Kurnit DM, Taylor J, Hanash SM (2001) Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* **61**: 6885–6891.
16. Alizadeh AA, Eisen MB, Davis RE, *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.
17. Armstrong SA, Staunton JE, Silverman LB *et al*. (2002) MLL translocations

specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **30**: 41–47.
18. Bertucci F, Salas S, Eysteries S, *et al*. (2004) Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene* **23**: 1377–1391.
19. Sültmann H, von Heydebreck A, Huber W, *et al*. (2005) Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. *Clin Cancer Res* **11(2)**: 646–655.
20. Huber W, Von Heydebreck A, Sultmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18(Suppl 1)**: S96–S104.
21. Tibshirani R, Hastie T, Narasimhan B and Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 99: 6567–6572.
22. Vapnik V (1999) *The Nature of Statistical Learning Theory*: Springer Verlag, Berlin.
23. Breiman L (2001) Random forests. *Machine Learn J* **45**: 5–32.
24. Hromas R, Broxmeyer HE, Kim C, Nakshatri H, Christopherson K, 2nd Azam M and Hou YH (1999) Cloning of BRAK, a novel divergent CXC chemokine preferentially expressed in normal versus malignant cells. *Biochem Biophys Res Commun* **255**: 703–706.
25. Frederick MJ, Henderson Y, Xu X, Deavers MT, Sahin AA, Wu H, Lewis DE, El-Naggar AK and Clayman GL (2000) In vivo expression of the novel CXC chemokine BRAK in normal and cancerous human tissue. *Am J Pathol* **156**: 1937–1950.
26. Weinschenk T, Gouttefangeas C, Schirle M, *et al*. (2002) Integrated functional genomics approach for the design of patient-individual antitumor vaccines. *Cancer Res* **62**: 5818–5827.
27. Morimoto I, Sasaki Y, Ishida S, Imai K, Tokino T (2002) Identification of the osteopontin gene as a direct target of TP53. *Genes Chromosomes Cancer* **33**: 270–278.
28. Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D and Yeatman TJ (2002) Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *J Natl Cancer Inst* **94**: 513–521.
29. Steinberg P, Storkel S, Oesch F and Thoenes W (1992) Carbohydrate metabolism in human renal clear cell carcinomas. *Lab Invest* **67**: 506–511.

Protocols

CONTENTS

Protocol 4.1: Tissue samples and RNA isolation

Thirty-five kidney tumors (13 ccRCC, 13 pRCC, 9 chRCC), which belonged to a larger study (19) were processed by standard pathology. Immediately after surgery, tumor pieces were subjected to routine histopathological examination as described previously (19). Other tumor pieces were snap-frozen in liquid nitrogen and stored at –80°C. Following homogenization with a Micro-Dismembrator S (Braun Biotech, Melsungen, Germany), total cellular RNA was isolated by the Trizol method (TriFast, peqlab, Erlangen, Germany). RNA quality was checked with the Agilent 2100 bioanalyzer (Agilent Technologies GmbH, Waldbronn, Germany). Only high-quality RNA (28S/18S rRNA and E_{260}/E_{280} ratios close to 2) was used for the experiments.

Protocol 4.2: Microarray experiments

The cDNA microarrays encompassed 1794 clones for oncologically relevant genes and 2314 genes and expressed sequence tags (ESTs) found to be differentially expressed in previous work (19). With further control genes, altogether the microarrays contained 4207 genes and ESTs. Insert DNA was amplified from bacterial clones by PCR using vector-specific primers. The PCR products were precipitated by isopropanol, washed in 70% ethanol, dried and dissolved in spotting buffer consisting of 3 × SSC/1.5 M betaine. Presence of product bands was confirmed by agarose gel electrophoresis (Ready-to-Run, Amersham Pharmacia Biotech, Freiburg, Germany). The DNA was spotted in duplicate onto epoxysilane-coated glass slides (Quantifoil, Jena, Germany) using the Omnigrid (Genemachines, San Carlos, CA) arrayer and SMP3 split pins (Telechem, Sunnyvale, CA). After spotting, microarrays were rehydrated, and DNA was denatured with boiling water prior to washing with 0.2% SDS, water, ethanol, and isopropanol. The arrays were dried with pressurized air.

Protocol 4.3: RNA labeling and hybridization

Ten μg total RNA were mixed with 1 μg $(dT)_{17}$ primer, incubated at 70°C for 10 min and cooled on ice. The labeling reaction was performed in 12.5 μl containing 2.5 μl 5 × RT buffer (Invitrogen, Karlsruhe, Germany), 1.25 μl 0.1 M DTT, 1 μl dNTP mix (5 mM each dATP, dGTP, dTTP), 0.5 μl 3 mM dCTP, 0.5 μl (20 U) RNasin, 0.5 μl 1 mM Cy3- or Cy5-labeled dCTP (Amersham) and 1 μl (100 U) Superscript II reverse transcriptase (Invitrogen). The mixture was incubated for 1 h at 42°C, and the reaction was stopped by addition of 1.25 μl 50 mM EDTA (pH 8). The RNA was removed by hydrolysis with 5 μl 1 M NaOH at 65°C for 10 min, followed by neutralization with 1 μl 5 M acetic acid. Cy3- and Cy5-labeled samples were combined, precipitated with 100 μl isopropanol at –20°C for 30 min and centrifuged at 13,000 *g* for 15 min. The pellets were washed with 70% ethanol, air dried, and dissolved in 30 μl 1 × DIG-Easy hybridization buffer (Roche Diagnostics, Mannheim, Germany), containing 5 × Denhardt's solution and 10 ng μl^{-1} Cot1-DNA (Invitrogen). The sample was heat denatured (65°C, 2 min) and hybridized to the DNA on microarrays in a hybridization chamber (Corning, Acton, MA) overnight at 37°C. Unspecific probe binding was removed by washing with 1 × SSC/0.1% SDS (15 min) and 0.1 × SSC/0.1% SDS (10 min) followed by cleaning with 70% ethanol, 95% ethanol, and isopropanol before drying with pressurized air.

Protocol 4.4: Signal quantification and data analysis

Arrays were scanned with the GenePix 4000B microarray scanner (Axon Instruments Inc., Union City, CA), and spots were quantified using Arrayvision 6.0 software (Imaging Research Inc., St. Catharines, Ontario, Canada). Intensity values for duplicate spots of each cDNA clone were averaged. Background-corrected intensity values were normalized and transformed to generalized log-ratios through the VSN method (20) (see also Chapter 17). All analyses were performed using R (www.r-project.org), and Bioconductor packages (www. bioconductor.org).

Gene expression analysis of differentiating neural progenitor cells – a time course study

5

Ulf Gurok and Ulrike A Nuber

5.1 Introduction

Cellular states are determined by the expression of thousands of genes. Thus, the transition between cell types, for example from a progenitor cell to a differentiated cell, involves changes in the activity of many genes. This transition is not a sudden event, but takes place over a certain time range. The microarray technology allows us to simultaneously measure expression patterns of thousands of genes. To identify gene expression changes underlying the differentiation of neural progenitor cells (NPCs), we performed a time course study using cDNA microarrays. This chapter also addresses certain technical aspects, such as the fast and automated amplification of cDNA probes for the generation of DNA microarrays and the problem of limited biological material.

5.2 The experiment

We prepared cultures of NPCs from mice and differentiated these *in vitro* into more mature cells of the central nervous system: astrocytes, neurons, and oligodendrocytes (*Figure 5.1*) (1). Gene expression changes take place during the differentiation of NPCs and were analyzed with our cDNA microarray platform. We were mainly interested in early changes that take place during the transition from an undifferentiated cell type to a differentiated one. The hybridization scheme (*Figure 5.1*) depicts the cohybridization of cDNA derived from RNA of undifferentiated NPCs as a reference with targets from NPCs that were allowed to differentiate for 1, 2, or 4 days. Bioinformatics tools were used to normalize the obtained microarray data and to identify differentially expressed genes. This data was further subjected to a cluster analysis to deduce onset and dynamics of relevant gene expression changes (see also Chapters 19 and 20).

It is well known that cells change during cultivation. Therefore, we minimized the propagation time *in vitro*. However, this limits the amount of RNA available for analysis. To perform experiments with small amounts of

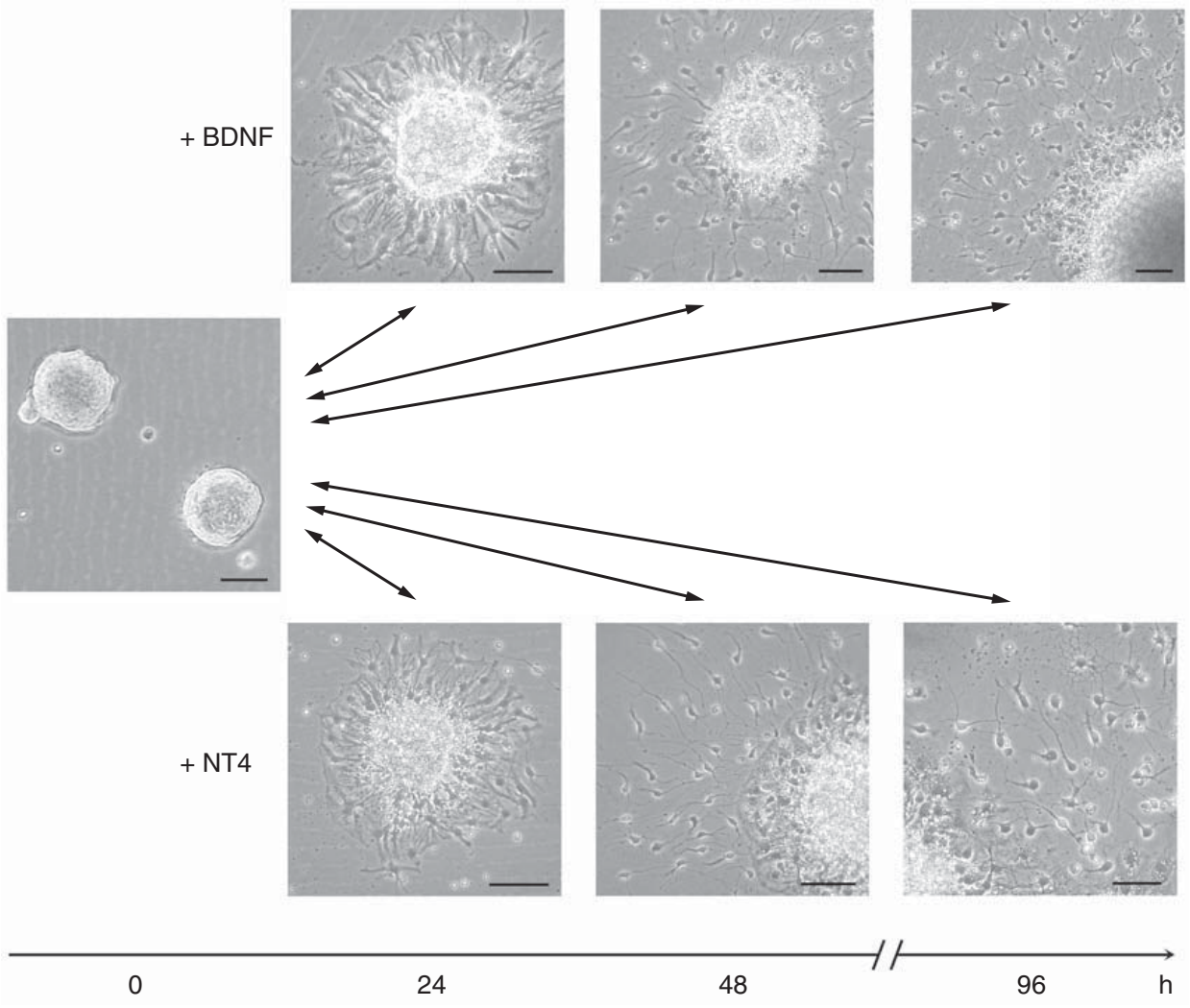


Figure 5.1.

Neural progenitor cell differentiation *in vitro*. The photographs depict morphological changes associated with adhesion, migration, and differentiation of neural progenitor cells within the first 4 days after induction of differentiation. Either neurotrophic growth factor BDNF or NT4 was added. Arrows indicate the microarray hybridization scheme, where RNA from differentiated cells was co-hybridized with RNA from undifferentiated cells as a reference. Scale bar equals 50 μm.

RNA, one can either amplify the starting material or the signal on the array. RNA can be amplified using a number of different protocols. They usually rely on the activity of reverse-transcriptase to generate cDNA, which serves as template for an RNA polymerase. The resulting RNA (aRNA) is then hybridized onto a microarray. A disadvantage is, however, that the amplification procedure potentially introduces a bias and therefore might change the original proportions of different transcripts that one actually wants to measure with a microarray. We decided to abstain from RNA amplification and instead applied a signal amplification method based on the 3DNA dendrimer technology (Genisphere, Hatfield, PA). This method relies on the specific hybridization of highly branched molecules called dendrimers to

the array-bound samples. These structures carry multiple fluorescent labels like Cy3 or Cy5. Therefore, each probe on the array is visualized by many fluorescent molecules thereby increasing the signal intensity and facilitating the detection with the array scanner.

Our protocol describes the preparation of cDNA microarrays, the hybridization of cDNA samples derived from RNA of undifferentiated and differentiated NPCs to the arrays, signal amplification with 3DNA dendrimers, scanning of the arrays, processing of raw data, identification of differentially expressed genes, and a cluster analysis of relevant gene expression changes.

5.3 Summary

We have applied cDNA microarrays containing 13 627 clones to analyze gene expression changes that take place during the *in vitro* differentiation of neural progenitor cells. The rapid production of arrays was facilitated by applying a robotic platform combined with a cooled microtiter plate storage system. To deal with the limited amount of RNA available, we applied a signal amplification method (Genisphere Inc. (see Protocol 5.3)). This allowed us to minimize cultivation time of primary cells, hybridize each sample twice with dyes swapped, and exclude potential biases in relative transcript abundance that might be introduced by an RNA amplification procedure.

Using a variance estimation we determined that a cut off in expression change for each individual clone set at twofold resulted in 2–5% false positives (see Protocol 5.4). This rate of false positives was further reduced by considering only such clones as relevant, whose expression changed more than twofold in two experimental series and represents another data filtering step.

To gain a more sophisticated insight into the molecular processes underlying the differentiation of neural progenitor cells we collected samples at different time points during the experiment. We followed the changes in gene expression in the cells over a 4-day period. This approach gives a more detailed insight into the molecular mechanisms involved in cellular development in our model system. For an individual gene or a group of genes a time course illustrates the onset and dynamics of expression changes that take place after induction of differentiation. A cluster analysis helps to identify such groups of coregulated genes which might also be involved in shared biological processes (see Protocol 5.5).

The goal of microarray experiments like this one is to learn more about gene expression changes that occur during the switch of cell states. Genes that are down-regulated during differentiation are preferentially expressed in the progenitor cells. They might be relevant for maintaining their self-renewing and differentiation capacity. For example, cluster 9 contains genes with the strongest and the earliest expression changes, and most of them encode cell-cycle proteins, indicating that one of the first events in this *in vitro* differentiation is cell cycle exit. In contrast, up-regulated genes are likely relevant for the process of differentiation or necessary for a specific function of the differentiated cell. Also, the timing of gene expression changes is definitely important for proper differentiation. Genes that

change early, especially those related to transcriptional regulation or signal transduction, are likely to have regulatory functions in the differentiation process. Many genes whose expression changes later encode for products that are important in the differentiated cell.

The heterogeneity of the neurosphere culture complicates the evaluation of the microarray results. Changes in expression of individual genes are unlikely to take place in all cells equally. Instead, the fold changes measured can be due to much stronger changes in gene expression in subpopulations of cells. Therefore, it is desirable to attribute individual gene expression changes to certain cell types, which can be done by immunofluorescence using antibodies against corresponding gene products. The purpose of this analysis was to discover genes that are relevant for the maintenance of neural progenitor cells, as well as for the migration and differentiation of their progeny. Having identified candidate genes relevant for these processes and knowing the dynamics of their expression is crucial for further functional analyses that will enhance our understanding of adult neural progenitor cells.

References

1. Gurok U, Steinhoff C, Lipkowitz B, Ropers HH, Scharff C and Nuber UA (2004) Gene expression changes in the course of neural progenitor cell differentiation. *J Neurosci* **24**: 5982–6002.
2. Huber W, von Heydebreck A, Sultmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18(Suppl 1)**: S96–S104.
3. Sabatti C, Karsten SL and Geschwind DH (2002) Thresholding rules for recovering a sparse signal from microarray experiments. *Math Biosci* **176**: 17–34.
4. Fan G, Egles C, Sun Y, Minichiello L, Renger JJ, Klein R, Liu G and Jaenisch R (2000) Knocking the NT4 gene into the BDNF locus rescues BDNF deficient mice and reveals distinct NT4 and BDNF activities. *Nat Neurosci* **3**: 350–357.

Protocols

CONTENTS

Protocol 5.1: Microarray production

The array used in this study contains 13 627 clones. Of these 10 080 are from the arrayTAG clone set (LION Bioscience, Heidelberg, Germany) and are short, sequence-verified murine cDNA clones, which are usually located close to the 3′ end of the respective transcript, but do not contain poly-A sequences. From the resource centre of the German Human Genome Project (RZPD, Berlin, Germany) 3 510 murine cDNA clones representing brain-expressed transcripts were obtained. In addition, 34 plant cDNA sequences were included. LION clones were amplified with LION 3′ and LION 5′ primers, RZPD and plant clones with M13 forward and M13 reverse primers. Clones were kept as inserts in plasmid vectors in *Escherichia coli* glycerol stocks in 96-well plates.

Since cDNA microarrays in general consist of thousands of clones, manual PCR is not applicable. We have used an automated PCR setup consisting of a robotic system from Tecan (Genesis Workstation 200, Tecan, Crailsheim, Germany) combined with a temperature-controlled hotel (automatic incubator cytomat 6002, Heraeus, Hanau, Germany). Template microtiter plates containing bacterial clones as well as empty PCR microtiter plates were stored in the hotel at 4°C (*Figure 5.2B, C*).The process was set up as follows: microtiter plates from the hotel were placed onto the robotic platform (*Figure 5.2A*) by means of an elevator within the hotel and a plate gripper of the robot. Next, PCR master mix (50 µl/well) was distributed and bacterial template from glycerol stocks (2 µl/well) was added by a liquid handling system (*Figure 5.2A*). The finished plates were transported back into the hotel (*Figure 5.2B, C*). They could be stored there or taken out while the process continued. Ready-to-cycle PCR plates were either immediately processed using external PrimusHT multiblock thermal cyclers (MWG Biotech, Munich, Germany; *Figure 5.2D*) or stored at –20°C until PCR amplification.

In addition, this system prepared the PCR products for evaluation with agarose gel electrophoresis and for DNA precipitation. In detail, PCR microtiter plates after PCR amplification as well as empty microtiter plates were placed in the hotel (*Figure 5.2B, C*). Then, each PCR plate along with an empty microtiter plate was transported out of the hotel onto the robotic platform (*Figure 5.2A*). 5 µl of gel-loading buffer was distributed into empty microtiter plates by a liquid handling system and 5 µl of 50 µl PCR product was added. All PCR products in these gel-loading plates were evaluated by agarose gel electrophoresis using the RoboSeq 4204S (MWG Biotech). To the remaining 45 µl of PCR product precipitation mix was added by the liquid handling system (see below). The combination of the Heraeus hotel with the liquid- and plate-handling robotic system from Tecan and the external PCR cyclers from MWG Biotech facilitated high-throughput PCR amplifications of cDNA clones to be used as probes for DNA microarrays.

After addition of precipitation mix (2.5 volumes ethanol, 0.1 volume sodium acetate) by the liquid-handling system, plates were transported back into the hotel. After further steps were performed manually, plates were incubated at –80°C for 30 min, and centrifuged (20 000 *g*, 30 min, 4°C). The pellets were washed with 70% ethanol, centrifuged again (20 000 *g*, 15 min, 4°C), and resuspended in 18 µl 3 × SSC. Eight microliters were transferred into 384-well plates by a Multimek 96/384 (Beckman Coulter, Krefeld, Germany) and stored at –20°C. These PCR products were printed on Corning GAPS II slides by using a robotic spotting device (SDDC-2 MicroArrayer, ESI, Toronto, Canada; ChipWriter Pro, BIORAD) with SMP3 pins from TeleChem International (Sunnyvale, CA). The average spot center-to-center distance was 204 µm.

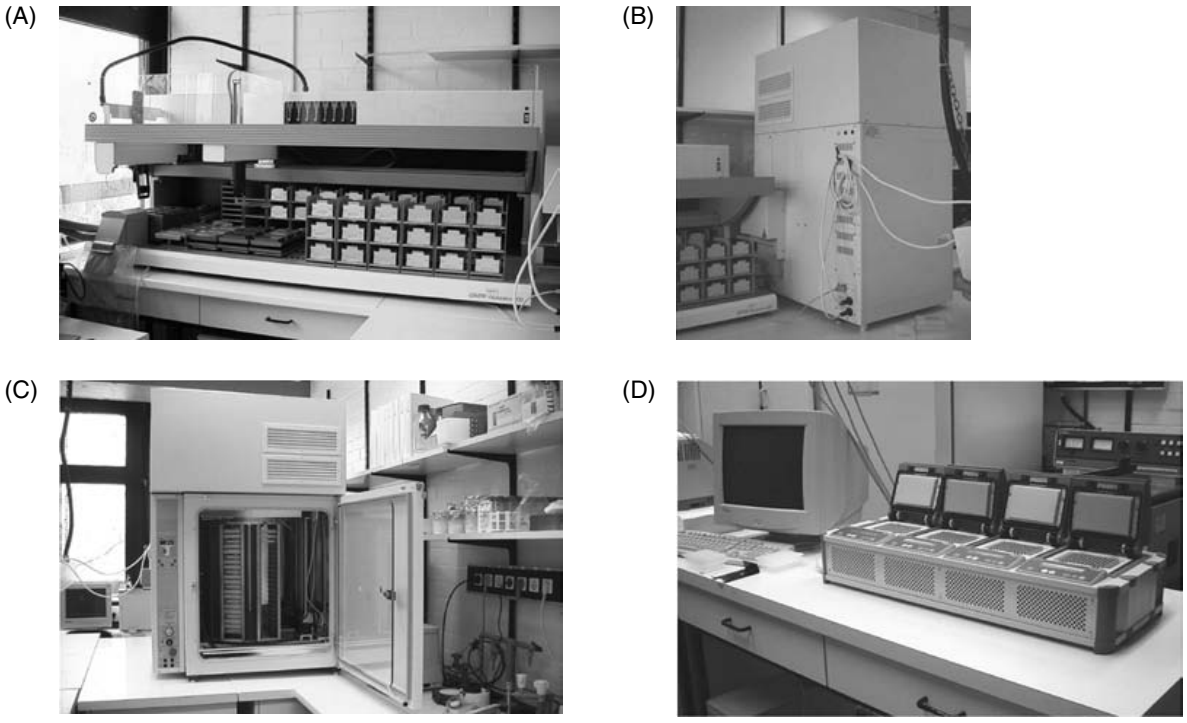


Figure 5.2.

Images of the Genesis Workstation 200 from Tecan (**A**), the automatic incubator cytomat 6002 from Heraeus (**B**, **C**), and the PrimusHT multiblock thermal cyclers from MWG Biotech (**D**), which were used to handle and amplify the cDNA microarray clones.

Table 5.1

PCR reaction for amplification of clones

Reaction mix		Primers	
10 × PerkinElmer PCR buffer	5 µl	LION 5′	AGCGTGGTCGCGGCCGAGGT
dNTPs (1 mM each)	10 µl	LION 3′	TCGAGCGGCCGCCCGGGCAGGT
MPI taq (made inhouse) (10 U/µl)	2 Units	M13 forward	GTAAAACGACGGCCAG
Forward primer (10 µM)	2 µl	M13 reverse	CAGGAAACAGCTATGAC
Reverse primer (10 µM)	2 µl		
Nuclease-free water	ad 50 µl		

Thermal profile	LION primers			M13 primers		
	Temperature	Time	Cycle number	Temperature	Time	Cycle number
Initial denaturation	94°C	3 min	1 ×	94°C	5 min	1 ×
Denaturation	94°C	30 s		94°C	45 s	
Annealing	68°C	30 s	35 ×	54°C	90 s	35 ×
Elongation	72°C	50 s		72°C	2 min	
Final elongation	72°C	10 min	1 ×	72°C	10 min	1 ×

Protocol 5.2: Cell culture and RNA preparation

Cells were prepared and cultivated according to standard protocols from the literature. A detailed description is given in Gurok *et al.* (1). Briefly, the progenitor cells grew in suspension and formed spherical aggregates called neurospheres. Differentiation of neurosphere cells was induced by removing a mitogen epidermal growth factor (EGF), allowing the cells to attach to the dish which was coated with the adhesive substrate poly-L-lysine, and by addition of either brain-derived growth factor (BDNF) or neurotrophin 4 (NT4) to the differentiation medium. Both are neurotrophic factors and support neuronal differentiation. This treatment led to breaking up of the neurosphere aggregates. The cells migrated away from the sphere and changed their morphology (*Figure 5.1*). They also sythesized marker proteins for distinct neural lineages such as βIII-Tubulin or GFAP (not shown).

We took samples of the cells before induction of differentiation and after 24, 48, and 96 h of differentiation. Total RNA was isolated with Trizol reagent (Invitrogen), precipitated with ethanol and resuspended in nuclease-free water. The concentration was determined spectrophotometrically and the quality was checked by agarose gel electrophoresis.

Protocol 5.3: Hybridization, washing and scanning

For every co-hybridization two RNAs were labeled in a reverse transcriptase (RT) reaction, RNA from undifferentiated and from differentiated cells, resulting in three hybridizations each for the BDNF series and the NT4 series (*Figure 5.1*). To account for dye-specific effects every hybridization was done twice, with the dyes exchanged by flipping the RT primers so that every RNA/cDNA was labeled once with the Cy3-specific capture sequence and once with the Cy5-specific sequence. Thus, six arrays were used for each series (NT4 and BDNF), adding up to 12 hybridizations in total.

Before hybridization, the spotted material was rehydrated by holding the slides over hot water until a vapor coating appeared, and quickly dried by placing them on a hot plate (98°C) for 3–5 s. Then the spotted material was crosslinked with the slide's surface by two successive UV crosslinking steps (120 mJ) in a UV Stratalinker 1800 (Stratagene, Amsterdam, The Netherlands). Remaining chemically active sites on the surface were blocked by 15 min incubation in succinic anhydride/sodium borate solution. Afterwards, the slides were briefly washed in ultrapure water and dried by centrifugation (125 *g*, 3 min, room temperature) and 3–5 s incubation on the hot plate (98°C). They were now ready for hybridization.

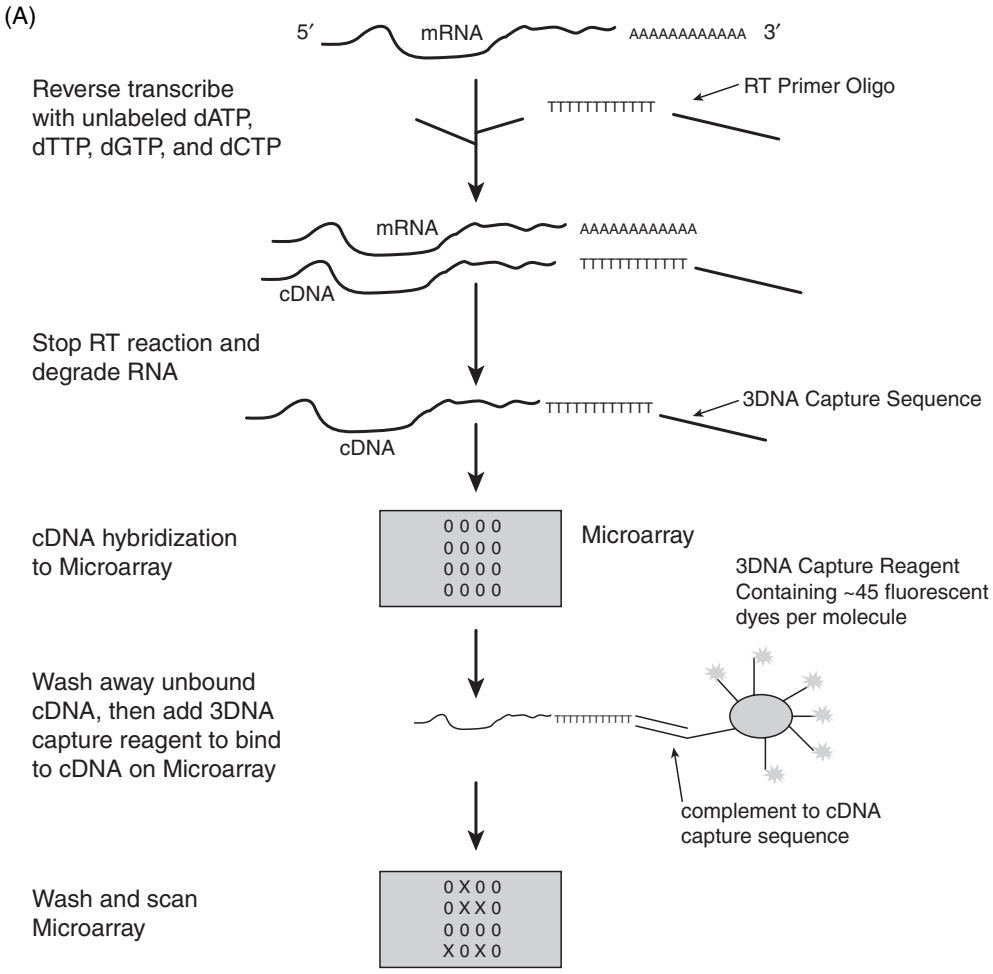
Labeling and hybridization reactions were performed using the 3DNA Array 50 Expression Array Detection Kit (Genisphere, Hatfield, PA; *Figure 5.3*). The labeling kit was used according to manufacturer's instructions. For each labeling reaction, 20 μg of total RNA was deployed. In brief, RNA and RT primers were mixed, heat denatured, and RNase inhibitor was added. Then a reaction mix consisting of the Superscript II RT (Invitrogen), Superscript II reaction buffer, dNTPs, and DTT was pooled with the RNA/RT primer mix. The RT was allowed to react for 2 h at 42°C before the reaction was stopped and the DNA/RNA hybrids were heat denatured.

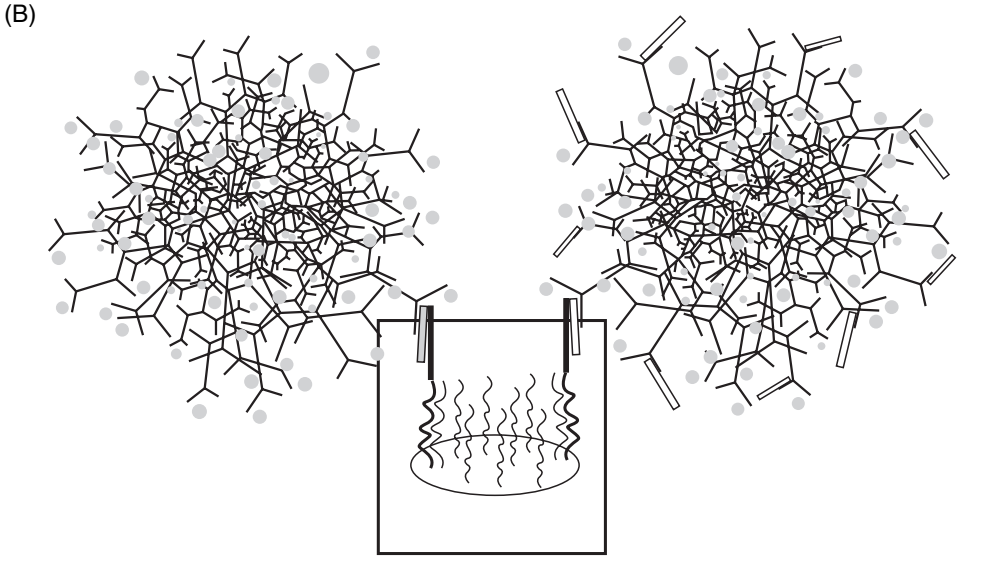
The primers for the reverse-transcription contained a poly-T sequence which bound to the poly-A tail of the mRNA. In addition they included a capture sequence which allowed for discrimination of the two cDNA pools after hybridization of synthesized cDNA to the array.

After cDNA synthesis and incorporation of the capture sequences, the cDNA was concentrated according to the manufacturer's instructions. Briefly, linear acrylamide, NaCl, and ethanol were added to the reactions. The mix was incubated at –20°C for 30 min and then centrifuged at 12,200 *g* at room temperature for 15 min. After aspiration of the supernatant the pellets were washed with 70% ethanol, recentrifuged for 5 min, and dried in a heat block at 65°C for 20 min.

For cDNA hybridization to the array the pellets were carefully resuspended in 10 μl of nuclease-free water and heated to 65°C for 10 min. Then the final hybridization mix was prepared and incubated at 75°C for 10 min and 45°C for 20 min. Meanwhile the array was prewarmed to 45°C for 15 min. The final hybridization mix (cDNA) was mixed, centrifuged briefly, applied to the prewarmed array and covered with a coverslip. The slide was put into a sealed humidified chamber and was incubated in a water bath at 42°C overnight.

The next day the 3DNA capture reagents, that is the dendrimers, were hybridized to the array. First, the array was washed by sequential incubation in 2 × SSC/0.2% SDS for 10 min, in 2 × SSC for 10 min, and in 0.2 × SSC for 10 min at room temperature. To remove remaining liquid from


(A)
5′
mRNA
AAAAAAAAAAAA 3′
Reverse transcribe
with unlabeled dATP,
dTTP, dGTP, and dCTP
RT Primer Oligo
TTTTTTTTTTTT
mRNA
AAAAAAAAAAAA
TTTTTTTTTTTT
cDNA
Stop RT reaction and
degrade RNA
3DNA Capture Sequence
TTTTTTTTTTTT
cDNA
cDNA hybridization
to Microarray
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
Microarray
3DNA Capture Reagent
Containing ~45 fluorescent
dyes per molecule
Wash away unbound
cDNA, then add 3DNA
capture reagent to bind
to cDNA on Microarray
complement to cDNA
capture sequence
Wash and scan
Microarray
0 X 0 0
0 X X 0
0 0 0 0
X 0 X 0


(B)

the slide surface, it was transferred into a slide holder and centrifuged (125 *g*, 3 min). Meanwhile the 3DNA capture reagents were thawed in the dark for 20 min. 3DNA capture reagents and the hybridization buffer were then heated to 55°C for 10 min, and an anti-fade reagent was added to the hybridization buffer. The final hybridization mix (3DNA) was prepared, mixed very carefully, and incubated at 75°C for 10 min followed by an incubation at 45°C for 20 min. The array was again prewarmed as before. The mix was applied to the array and the slides were kept in a dark humidified chamber at 42°C for 3 h. Subsequently, the array was washed as before in 2 × SSC/0.2% SDS for 10 min, 2 × SSC for 10 min, 0.2 × SSC for 10 min, and finally briefly in deionized water. The slide was dried by centrifugation (125 *g*, 3 min) followed by an incubation at 42°C for 5 min. It was stored in a dry and dark box until scanning.

The arrays were scanned with the Affymetrix 428 Array Scanner (Affymetrix, Santa Clara, USA). Fluorescence intensities of Cy3 and Cy5 were measured separately at 532 nm and 635 nm. The photomultiplier tube gain was typically between 40–55 dB for Cy3 and 35–45 dB for Cy5. This ensured that the signal intensity reached saturation in less than 1% of the spots. The resulting images were saved as 16-bit data files in tag image file format (TIFF).

Table 5.2

Final hybridization mix (cDNA)		Final hybridization mix (3DNA)	
Concentrated cDNA	10 µl	Hybridization buffer + Anti-Fade	25.0 µl
Hybridization buffer	22 µl	3DNA capture reagent 1 (Cy3)	2.5 µl
Array 50 dT Blocker	2 µl	3DNA capture reagent 2 (Cy5)	2.5 µl
Cot–1 DNA	2 µl	Nuclease-free water	5.0 µl
		Cot–1 DNA	2.0 µl

Figure 5.3.

Labeling and detection of mRNA with Genisphere dendrimer technology. (**A**) Messenger RNA is reverse transcribed into cDNA using unlabeled nucleotides and a poly-T oligo, that carries a capture sequence. Two such cDNA populations are co-hybridized to the microarray. These are specifically detected by the 3DNA capture reagent. (**B**) The capture reagent contains dendrimers, large molecules composed of DNA strands coupled to fluorescent labels (Cy3 and Cy5, depicted as circles). The dendrimers present sequences complementary to the capture sequences on the cDNA, thereby allowing the detection of cDNAs derived from one RNA pool only. In consequence, a single cDNA molecule attracts approximately 45 fluorescent labels (Cy3 or Cy5) to the microarray, leading to a higher sensitivity of this method compared to the hybridization of cDNA directly labeled during reverse transcription with Cy3- or Cy5-coupled nucleotides.

Protocol 5.4: Data processing

The image files were imported into the Microarray Suite image analysis software (Version 2.0), which runs as an extension of IPLab Spectrum Software (Scanalytics, Fairfax, VA). The software determined the raw spot intensities of Cy3 and Cy5 and performed a local background subtraction. Empty spots and spots carrying plant sequences were excluded from further analysis. Each dye swap experiment was normalized by applying variance stabilization (see also Chapter 17) (2) using the vsn package of bioconductor (http://www.bioconductor.org). Means of normalized log-products and log-ratios of each dye swap experiment pair were used for further analysis. Normalization procedures were performed using R (http://cran.R-project.org).

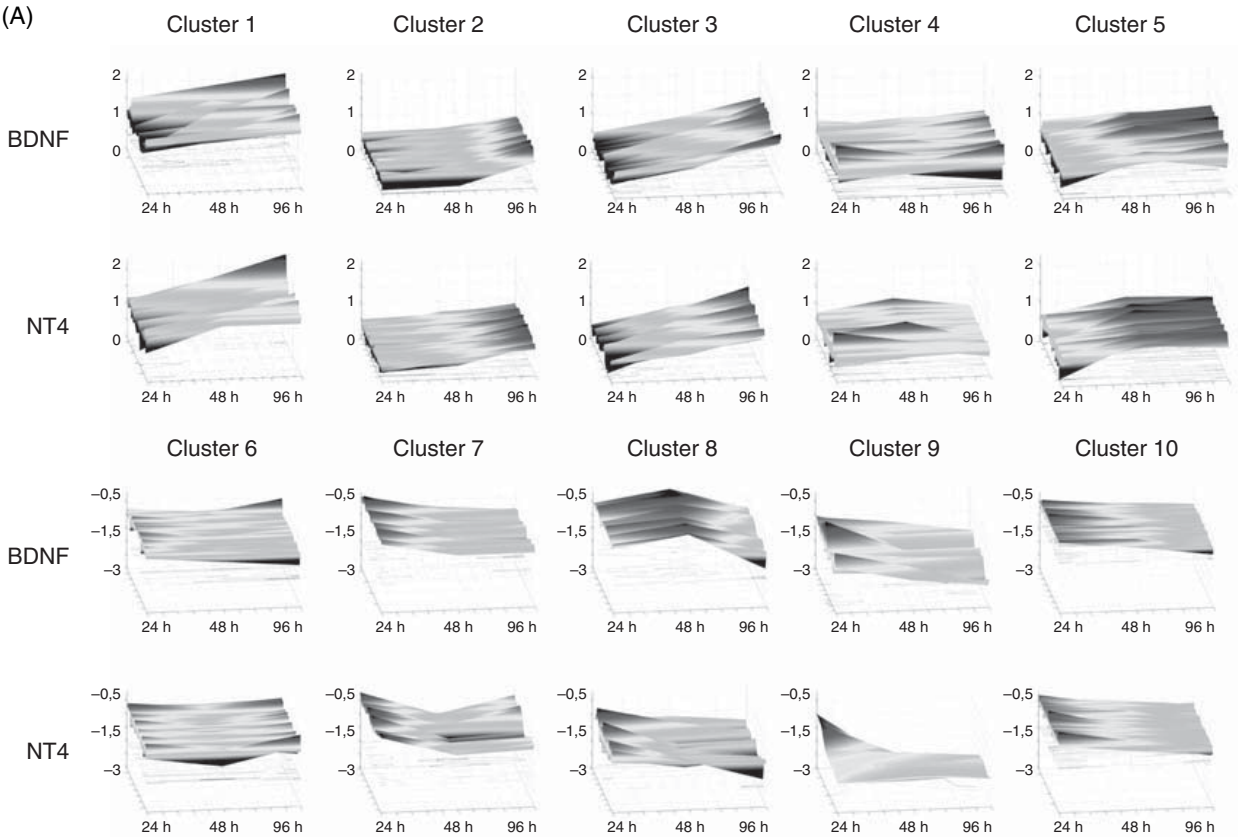
To determine a meaningful cut-off value that designates differentially expressed genes we applied a statistical analysis which considers the small number of biological replicates and aims at minimizing the percentage of false positives. To do so, a variance estimation using a pooled estimate of the variance over all genes of three self-to-self comparisons with RNA from undifferentiated cells was performed. A similar approach has been described by Sabatti *et al.* (3). In order to use a robust variance measurement we determined the median of absolute deviation (MAD) as variance estimator. The MAD in all three independent experiments of self-to-self comparison was very similar: 0.297 ± 0.021. Based on this analysis, we can assume a rate of 2% false positives when applying a universal threshold of 2.17-fold change. A rate of 5% false positives can be assumed when applying a threshold of 1.8-fold. We therefore considered all clones above a 2.0-fold change as relevant. Thus, we can assume a false positive rate of 2–5% at this point of analysis.

To extract clones that are of interest for further analysis we concentrated on all clones, whose expression changed more than twofold in at least one of the three time points of differentiation in both experimental series. These were 722 clones in the BDNF series and 624 in the NT4 series. Since both neurotrophic growth factors target the TrkB receptor, they cause very similar molecular effects in the cells. Their functions *in vivo*, however, do not perfectly overlap (4). To find common biological effects related to neural differentiation we then took the intersection of both lists, amounting to 454 clones. In addition this step further reduced the number of false positives.

Protocol 5.5: Cluster analysis

Of the 454 clones, 441 clones showed a consistent up- or down-regulation at the three differentiation time points were included in a cluster analysis. We clustered the 441 datasets, each representing a single time course, by applying the *k*-means algorithm and using a refined Euclidean distance measure (*Figure 5.4*). This specifically takes into account the time dependence of gene expression changes. We performed *k* means clustering with *k* values ranging from 3 to 15 and found that for our dataset, $k = 10$ is optimal to separate many clearly different dynamics without separating genes with too similar dynamics. The distance was defined as the weighted sum of *k*-means assignment and a similarity of shapes between cluster centers (gradient). The distance measure (D) was defined as follows: $D(x,y) = a\ d1(x,y)+(1-a)\ d2(x,y)$ where a is 0.5, x,y are the two gene profiles to be compared, $d1$ the Euclidean distance between x and y, and $d2$ the gradient of x and y. These calculations and the respective visualization were carried out using MATLAB (Version 6.0.0.88, Release12, MathWorks, MA).

(A)



(B)

Cluster	1	2	3	4	5	6	7	8	9	10	Sum
BDNF	24	51	36	106	54	39	21	15	18	77	441
NT4	22	82	30	74	63	39	16	18	6	91	441
Intersection	21	35	12	55	33	22	3	1	6	61	249

Figure 5.4.

(**A**) Clusters of clones with similar expression changes over time. The dynamic behavior of the 441 clones that are consistently up- or down-regulated in both the BDNF and NT4 series is shown separately. For each series, 10 clusters were found. Clusters 1–5 contain up-regulated, and clusters 6–10 down-regulated clones. The x-axis depicts the three time points of differentiation. The y-axis shows relative fold changes, that is expression changes referred to the undifferentiated state (0 h). These relative numbers are estimated logarithmic fold changes. Note that y-axis values at 24 h which are different from the value 0 indicate that the expression of a clone had already changed between 0 h and 24 h. (**B**) Numbers of clones contained in each cluster of the NT4 and BDNF series.

A microarray-based screening method for known and novel SNPs

6

Ena Wang and Francesco M Marincola

6.1 Introduction

Genetic polymorphism is a hallmark of human biology and the basis for individuality. Although the completion of the human genome project provides the first reference sequence of all human chromosomes the challenge remains to identify and characterize the frequency of deviations from this reference among populations with different ethnic background as well as individuals carrying distinct traits within a population (1). Approximately, 1.8 million polymorphic sites mostly consisting of single nucleotide polymorphisms (SNPs) have been so far discovered throughout the human genome (http://snp.cshl.org;http://www.ncbi.nlm.nih.gov/SNP/). The approximate frequency of SNP occurrence is one per kilobase (2, 3). Detection of SNPs due to genetic variation in a given population (polymorphisms) or subsequent genetic adaptations occurring throughout life (mutations) has gained increasing attention due to the functional implications that SNPs in coding and non-coding regions bear on biological and pathological events. In fact, 25% of the known non-synonymous SNPs could affect the function of the correspondent gene product (4–7). Therefore, detection of SNPs due to genetic variation in a given population may have important implications in the natural history of disease and its response to therapy (8). Yet, for several reasons it is still unclear whether the prevalence of common diseases can be truly attributed, at least in part, to SNPs. The main reason is that the prevalence of SNPs throughout the genome in a given population is known only for few genes as exemplified by the human leukocyte antigen (HLA) complex which has been extensively studied due to the significance that polymorphic sites bear on allo-immunization. A lesson learned from the study of the HLA region is that the number of polymorphisms recognized in a given population is highly correlated with the accuracy and resolution of the method used. Therefore, as an example, with the ever-growing interest in the study of polymorphic sites of genes associated with immune function (9–12), it is likely that the number of recognized variations will continue to grow in the coming years. This rate of discovery will be enhanced by the high-throughput technologies that are continuously developed for SNPs to cover the complexity and heterogeneity of human biology and pathology. While,

experimentally, tools are available for limited population studies, in clinical research a large number of individuals may need to be screened when investigating associations between genetic variation and disease susceptibility or responsiveness to therapy. In such an endeavor, a tool capable of efficiently and economically identifying known and flagging unknown SNPs could dramatically increase the understanding of human pathology making feasible the direct application of genome-wide investigation during the conduct of clinical trials (13).

6.2 High resolution SNP detection methods

High resolution SNP identification relies on sequence-based typing (SBT), which can recognize known SNPs in homo- or heterozygous conditions and spot unknowns. Capillary sequencing makes this method semi-high throughput. However, the majority of SNPs are dispersed across the whole genome, with few exceptions such as the HLA region, at an average 1-kb distance from each other. Since accurate sequencing is limited to about 800–1000 bp per sequencing reaction, SBT is not an efficient method for high throughput sequencing of most genomic regions. Furthermore, the cost of SBT is not affordable for most research facilities. Finally, even sequencing has its own limitations since it does not allow segregating *cis*- to *trans*- ambiguities in heterozygous conditions when more than one SNP is detected in the region being sequenced (14). Pyrosequencing is a newly developed real time quantitative sequencing method that may complement high-throughput SBT since it can resolve *cis*/*trans* ambiguities. Using a programmed nucleotide dispensation order (NDO), pyrosequencing is especially suitable for the detection of SNPs when their frequency is low and for studies using pooled samples (15, 16). The drawback of this method is its cost, the complexity of designing NDOs and the special equipment and reagents necessary.

In recent years, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has emerged as a powerful tool for genotyping SNPs (17–19). The comprehensive utilization of this approach is, however, hampered by the complexity of the technology, the cost and the limited throughput.

6.3 High throughput methods for SNP detection

The ideal SNP detection method should be simple, accurate, inexpensive, and suitable for potential automation to allow genome-wide screening for known and unknown SNPs. Array-based technology is the only approach that, at least in theory, could fulfill most of these requirements. Array technology applied to the detection of SNPs has confirmed most of the common polymorphisms previously identified by conventional techniques and has identified a large number of new SNPs (20).

Array-based allelic discrimination methods can be categorized according to four principles: allele-specific hybridization, allele-specific primer extension, allele-specific oligonucleotide ligation and allele-specific invasive

cleavage (21). These methods utilize known SNP position and their flanking sequence information for the design of allele-specific probes.

6.4 Screening methods for known and unknown SNPs

Most array-based approaches are aimed at the detection of known SNPs (22, 23). To broaden the power of SNP detection to the identification of unknown variants, high-density oligonucleotide arrays have been designed that cover all possible sequence permutations of the genomic region of interest and function as an array sequencing method (24–27). These arrays are composed of oligomer probes overlapping at one nucleotide interval (one nucleotide tiling principle) covering the genomic segment investigated. Each position includes four alternative nucleotides to cover all possible genomic permutations. Thus for a given genomic fragment, the number of oligos needed to identify any possible SNP is equal to the number of base pairs of the fragment times four. Although powerful, this approach is limited by the complexity of its design and the cost of the production and utilization of the chips. In particular, the cost of array preparation is disproportionate for genomic areas with low density of SNPs.

Single base chain extension (SBCE) utilizes the dideoxynucleotide sequence termination reaction principle to interrogate probes downstream of differentially labeled single nucleotide incorporation templates. Using a one nucleotide tailing design, it is possible to achieve results similar to those obtainable with array sequencing using four times less probe for a given genomic region.

A simplified screening tool that could discriminate conserved from polymorphic genomic regions or could identify rare individuals carrying unusual SNPs could dramatically increase the efficacy of allele-specific sequencing or targeted SNP identification.

Designing a simplified SNP screening array

The strategy described here utilizes the well-established allele-specific hybridization method in combination with a chromosome walking approach for the genome-wide screening of known and unknown SNPs. This approach requires four times less oligos than the SBCE method and eight times less oligos than the array sequencing method to investigate a given fragment of genomic DNA. Thus, this strategy should be considered a screening tool applicable to the investigation of unexplored areas of the human genome prior to extensive sequencing expeditions or the construction of high-density oligo arrays. Known SNPs can be identified while genomic regions containing unknown SNPs can be flagged and subsequently annotated by SBT. In this way, large chromosomal segments are screened and few regions where SNPs are present are identified. Therefore, the amount of sequencing required for their definitive characterization is drastically reduced. In addition, as new alleles are identified new allele-specific oligonucleotides can be incrementally added to the array for definitive typing purposes.

The SNP screening array includes two types of oligos: consensus oligos (covering the full length of the genomic fragment investigated) and allele-specific oligos (representing known variants from the empirically selected consensus sequence of the genomic area to be investigated). Both types of oligos are 18-nucleotides long. Consensuses oligos are defined and designed according to the sequence of an arbitrarily selected reference DNA and cover, at a four-nucleotide tiling interval, the full sequence of the genomic area investigated. The SNP position in a 20 nucleotide oligo (approximately the same length of the oligos used in this array) bears minor effects on the hybridization efficiency as compared to the kind of nucleotide change as long as the three outermost positions at either end of the oligo are avoided (28). Therefore, the four-nucleotide tiling design assures that any SNP within the interested region should be identifiable by at least one consensus oligo (*Plate 3*). Furthermore, the overlapping consensus oligonucleotide system provides references consisting of four or five consecutive consensus oligonucleotides with interrogating SNP dynamically positioned with which information derived from the allele-specific oligonucleotides can be compared, facilitating the interpretation and discrimination of allele-specific hybridization in hetero- or homozygous conditions. The fluctuating hybridization pattern of the consensus oligonucleotides surrounding an unknown SNP indicates its existence that could be subsequently confirmed by sequencing limited to the individual carrying the variant and the region of interest (29).

Allele-specific oligonucleotides of the same length as consensus oligonucleotides are selected based on available databases and the SNP interrogation site in the oligo is positioned at the centermost position. With this strategy, the SNP screening array can be used to spot novel while identifying known SNPs.

Detection system

The screening array employs two differentially labeled fluorochromes for proportional hybridization testing. The reference sample is arbitrarily selected and consistently used for all arrays. For instance, a cell line (possibly homozygous for the genomic site of interest) that can be continuously expanded represents a good reference sample. Complete sequence information about the region investigated could be obtained by sequencing the reference cell line and using the sequence to design the consensus oligonucleotides. For larger genomic segments or for arrays covering different chromosomal regions it is most likely that any selected reference would include heterozygous sites. This is acceptable as long as the information is documented and the reference kept constant. This will allow interpretation of the experimentally obtained data. The reference sample exemplified here consists of a cell line with homozygous SNP loci identical to the consensus sequence (*a,a*). Test and reference samples are amplified by PCR followed by *in vitro* transcription to generate single-stranded RNA. Array data are generated from cohybridization of fluorescence-labeled reference (i.e. Cy3, green) and test (i.e. Cy5, red) cDNA samples to consensus and allele-specific oligos, the latter representing known SNPs (variant oligos). Results are represented as log base 2 of the fluorescence intensity ratio

(Log_2Ratio). In diploid organisms, four combinations can occur: (i) homozygosity at a certain locus of the test sample identical to the consensus (*a,a*) (consensus oligos $Log_2Ratio = 0$); (ii) homozygous SNP alleles that differ from the homozygous consensus (*b,b*) (allele-specific and the corresponding consensus oligo $Log_2Ratio > 1$); (iii) heterozygosity with one allele being identical and one allele being different from the consensus (*a,b*) (allele-specific oligo $Log_2Ratio > 1$ while the corresponding consensus oligo $Log_2Ratio < 1$); (iv) heterozygosity with both alleles different from the consensus (*b,c*) (consensus oligo $Log_2Ratio > 1$ and no hybridization to the allele-specific oligo). In regions containing unknown SNPs and, therefore, when no allele-specific oligos were designed to represent the unknown SNP, competitive hybridization occurs only in the consensus oligo between reference and test sample. Because of the perfect complementation of the reference sample to the consensus oligo, exclusive reference sample hybridization indicates the presence of a new SNP at that specific position (*Plate 4*).

Although this conceptually applies to the whole genome, in loci containing more than one polymorphic site, various combinations can simultaneously occur. This approach has been validated using the polymorphic HLA gene complex to exemplify the various combinations (29). Various permutations of homozygosity and heterozygosity have been illustrated and correspondent consensus hybridization that produces complex hybridization patterns highly specific for a particular phenotype could be observed. In these highly polymorphic conditions, each haplotype combination maintains a highly reproducible profile characterized by minimal variance. This allows the creation of "genotypic masks" within narrow ranges of variation to "fingerprint" known haplotype permutations for high-throughput typing of highly polymorphic genes. The power of this strategy in identifying unknown SNPs was analyzed using Relative Operating Characteristics (ROC) analysis which characterizes the performance of a binary classification model across all possible trade-offs between the false negative and false positive classification rates and allows the performance of multiple classification functions to be visualized and compared simultaneously (30). For each 18-mer probe, starting from third base and ending at 16th base, if the test target contained at least one single nucleotide different for the consensus sequence, it was defined as specific region SNP(+); otherwise it was SNP(–). When the test sample is most different from the consensus reference, as for *b,b* and *b,c*, higher accuracy with a sensitivity of 82% and a specificity of 96% was observed. The worst accuracy was noted when test and reference samples were closest as in the case of *a,b* heterozygosity (sensitivity 82% and specificity of 82%). The most informative analysis was, however, provided using data from all the possible combinations since in most common experimental conditions the relationship between test and consensus sample is not known and, therefore, all possible allelic combinations should be expected. In that case an optimal threshold of $Log_2Ratio <$ or equal to -0.62 yielded a sensitivity of 82% and a specificity of 89%. Thus, this strategy may identify four out of five unknown SNPs with 90% accuracy and the highest chance of discriminating false positive results when an *a,b* heterozygous sample is tested.

Genomic DNA amplification for array analysis

High quality and a sufficient quantity of genomic DNA are critical for high throughput approaches. In the case of clinical samples or biopsies where limited material is available, the amount of isolated DNA is in most cases far below the requirement for multiple genomic analyses. Therefore, high fidelity genomic DNA amplification becomes the first challenge for accurate genotyping. Depending on the genotyping method employed, DNA amplification can be allele-specific, gene-specific or whole genome-wide. *Table 6.1* summarizes amplification methods according to their underlying principle and their advantages and disadvantages.

Here, only the T7-based gene-specific amplification method suitable for the SNP screening array analysis is presented. This strategy can be applied to the study of any locus using PCR in combination with *in vitro* transcription. Gene-specific primers flanking the gene of interest within a 1- to 10-kb range can be designed. Multiple primer pairs are needed when the targeted gene is larger than 10 kb or multiple genes are scanned simultaneously. To generate single strand targets, a T7 promoter sequence (5′aaa cga cgg cca gtg aat tgt aat acg act cac tat agg cgc 3′) is attached to the 5′ end of the forward primer for PCR amplification. *In vitro* transcription generates large quantities of linear single strand RNA for fluorescence labeling and hybridization (See Protocl 6.1).

6.5 Summary

The current SNP scanning array represents a potentially powerful and efficient strategy for high-throughput screening of genes for which little is known about their polymorphism. This strategy could also be used to identify mutations in disease genes or for typing known allelic variants of well-characterized genes such as HLA. This, however, would require specialized design of numerous oligos encompassing known variants and supportive software for efficient data interpretation. Various scenarios have been best exemplified by using exon 2 of the HLA-A locus as a model to identify an unknown allele as well as a known allele in *a,a, b,b, a,b* and *b,c* homo and heterozygous conditions (29). However, SNPs occur in the human genome on average every 600–2000 bases (1, 31). Therefore, most genes are characterized by a relatively narrow range of polymorphisms that would allow a relatively simple design of oligo-array chips and interpretation of results. Independently of the genomic region investigated, this strategy can identify unknown variants through observation of disproportionably depressed Log_2Ratios of signals obtained at the position of consensus oligonucleotides. Thus, it may provide a great improvement in the ability to screen different genes for the frequency and location of polymorphic sites, which can be confirmed by site-directed sequencing limited to the region of interest. Thus, the best application of this strategy stems from the clinical need to rapidly segregate genes characterized by the presence or lack of polymorphisms in their coding or regulatory regions that may affect clinical phenotypes. A good example of such application is the screening of cytokines, chemokines and their receptors whose polymorphism(s) have been asso-

Table 6.1. Methods for genomic DNA amplification

Methods	Amplification	Primer used	Advantage	Disadvantage	Reference
SNP-specific amplification	PCR	Allele-specific primers with SNP position at the 3′ end	SNP-specific	One SNP/reaction	
Gene-specific amplification	PCR plus T7 IVT	Gene-specific primer, 5′ primer with attachment of T7	Final products are single strand RNA or cDNA which reduces the complexity of competitive hybridization and enhances specific and efficacy of hybridization	Multiple primers needed for broader coverage of the genome	(29)
Random genome amplification	PCR	Random primer with extended PCR primer sequence	Potential coverage of all SNPs	Low efficiency of amplification and nonspecific artifacts	(35)
Linear T7 genomic amplification	Restriction fragmentation, poly dT tailing and dA-T7 IVT	Oligo dA-T7 primer	Potential coverage of all SNPs, less bias and large quality products	Relative short fragments and incomplete coverage of loci	(36)
Linker-adapter PCR	Restriction fragmentation, adaptor ligation and PCR	Adaptor-specific PCR primer	Potential coverage of all SNPs, validated and commercially available kits, 200–700 bp size.	Missing SNPs in proximity of the restriction site	(37)
ϕ 29 Multiple displacement amplification (ϕ 29 MDA)	ϕ 29 polymerase based random amplification	Phosphorothioate-modified random primer	Low error rate (3×10^{-6}), 99.82 genome coverage	Amplified DNA >10 kb and further amplification and fragmentation are needed	(38, 39)
Degenerate oligonucleotide-primed PCR (DOP-PCR)	PCR	DOP primer CCGACTCGAGNN NNNNATGTGG	One primer, low cost, robust amplification with reduced genome complexity	Only 48% of the validated SNPs could be genotyped and 22% of the predicted products are not amplified.	(40–42)
Primer extension pre-amplification (PEP)	Random PCR	Random mixture of 15 base oligo primers	Simple PCR reaction	78% representative of genome and high error (1×10^{-3})	(43, 44)
OmniPlex amplification	Random DNA fragmentation followed by OmniPlex library generation with universal flanked adaptor followed by PCR	Universal primer GTAATACGACTCA CTATA	One primer, 99.8% concordance in SNP genotyping and 90% representative of original genome	Lower than 10 ng input DNA can reduce representation significantly	(45)

ciated with individual predisposition to immune pathology, survival of transplanted organs and predisposition to cancer (22, 32–34).

References

1. Wang DG, Fan J-B, Siao C-J, *et al.* (1998) Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.
2. Li WH and Sadler LA (1991) Low nucleotide diversity in man. *Genetics* **129**: 513–523.
3. Sachidanandam R, Weissman D, Schmidt SC, *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
4. Cooper DN, Ball EV, and Krawczak M (1998) The human gene mutation database. *Nucleic Acids Res*. **26**: 285–287.
5. Ng PC and Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **12**: 436–446.
6. Collins FS, Brooks LD and Chakravarti AA (1998) DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**: 1229–1231.
7. Schafer AL and Hawkins JR (1998) DNA variation and the future of human genetics. Nature Biotech. **16**: 33-39.
8. Jin P and Wang E (2003) Polymorphism in clinical immunology. From HLA typing to immunogenetic profiling. *J Transl Med* **1**: 8.
9. Howell WM, Calder PC and Grimble RF (2002) Gene polymorphisms, inflammatory diseases and cancer. *Proc Nutr Soc* **61**: 447–456.
10. Haukim N, Bidwell JL, Smith AJP, *et al.* (2002) Cytokine gene polymorphism in human disease: on-line databases, supplement 2. *Genes Immunol* **3**: 313–330.
11. van Sorge NM, van der Pol W-L and van de Winkel JGJ (2003) FCgR polymorphisms: implications for function, disease susceptibility and immunotherapy. **61**: 202.
12. Parham P (2003) Immunogenetics of killer-cell immunoglobulin-like receptors. *Tissue Antigens* **62**: 194–200.
13. Kwok P-Y (2001) Genetic association by whole-genome analysis. *Science* **294**: 2669–2670.
14. Adams SD, Barracchini KC, Chen D, Robbins F, Wang L, Larsen P, Luhm R and Stroncek DF (2004) Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification. *J Transl Med* **2**: 30.
15. Wasson J, Skolnick G, Love-Gregory L and Permutt MA (2002) Assessing allele frequencies of single nucleotide polymorphisms in DNA pools by pyrosequencing technology. *Biotechniques* **32**: 1144–6, 1148, 1150.
16. Gruber JD, Colligan PB and Wolford JK (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing. *Hum Genet* **110**: 395–401.
17. Sauer S and Gut IG (2002) Genotyping single-nucleotide polymorphisms by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci* **782**: 73–87.
18. Krebs S, Medugorac I, Seichter D and Forster M (2003) RNaseCut: a MALDI mass spectrometry-based method for SNP discovery. *Nucleic Acids Res* **31**: e37.
19. Kim S, Shi S, Bonome T, Ulz ME, Edwards JR, Fodstad H, Russo JJ and Ju J (2003) Multiplex genotyping of the human beta2-adrenergic receptor gene using solid-phase capturable dideoxynucleotides and mass spectrometry. *Anal Biochem* **316**: 251–258.

20. Hacia JG, Brody LC, Chee MS, Fodor SP and Collins FS (1996) Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis [see comments]. *Nat Genet* **14**: 441–447.
21. Kwok PY (2001) Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* **2**: 235–258.
22. Turner D, Choudhury F, Reynard M, Railton D and Navarrete C (2002) Typing of multiple single nucleotide polymorphisms in cytokine and receptor genes using SNaPshot. *Hum Immunol* **63**: 508–513.
23. Guo Z, Gatterman MS, Hood L, Hansen JA and Petersdorf EW (2002) Oligonucleotide arrays for high-throughput SNPs detection in the MHC class I genes: HLA-B as a model system. *Genome Res* **12**: 447–457.
24. Chee M, Yang R, Hubbell E, *et al*. (1996) Accessing genetic information with high-density DNA arrays. *Science* **274**: 610–614.
25. Hacia JG (1999) Resequencing and mutational analysis using oligonucleotide microarrays. *Nat Genet* **21**: 42–47.
26. Hacia JG, Sun B, Hunt N, Edgemon K, Mosbrook D, Robbins C, Fodor SP, Tagle DA and Collins FS (1998) Strategies for mutational analysis of the large multi-exon ATM gene using high-density oligonucleotide arrays. *Genome Res* **8**: 1245–1258.
27. Patil N, Berno AJ, Hinds DA, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
28. Lee I, Dombkowski AA and Athey BD (2004) Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. *Nucleic Acids Res* **32**: 681–690.
29. Wang E, Adams S, Zhao Y, Panelli MC, Simon R, Klein H and Marincola FM (2003) A strategy for detection of known and unknown SNP using a minimum number of oligonucleotides. *J Transl Med* **1**: 4.
30. Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science* **240**: 1285–1293.
31. International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
32. Keen LJ (2002) The extent and analysis of cytokine and cytokine receptor gene polymorphism. *Transpl Immunol* **10**: 143–146.
33. McCarron SL, Edwards S, Evans PR, *et al*. (2002) Influence of cytokine gene polymorphism on the development of prostate cancer. *Cancer Res* **62**: 3369–3372.
34. Howell WM, Turner SJ, Bateman AC, and Theaker JM (2001) IL–10 promoter polymorphisms influence tumour development in cutaneous malignant melanoma. *Genes Immunol* **2**: 25–31.
35. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D and DeRisi JL (2002) Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* **99**: 15687–15692.
36. Liu CL, Schreiber SL and Bernstein BE (2003) Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* **4**: 19.
37. Klein CA, Schmidt-Kittler O, Schardt JA, Pantel K, Speicher MR and Riethmuller G (1999) Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc Natl Acad Sci USA* **96**: 4494–4499.
38. Hosono S, Faruqi AF, Dean FB, *et al*. (2003) Unbiased whole-genome amplification directly from clinical samples. *Genome Res* **13**: 954–964.
39. Paez JG, Lin M, Beroukhim R, *et al*. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* **32**: e71.

40. Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA and Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**: 718–725.
41. Jordan B, Charest A, Dowd JF, Blumenstiel JP, Yeh RF, Osman A, Housman DE and Landers JE (2002) Genome complexity reduction for SNP genotyping analysis. *Proc Natl Acad Sci USA* **99**: 2942–2947.
42. Kwok PY (2002) Making 'random amplification' predictable in whole genome analysis. *Trends Biotechnol* **20**: 411–412.
43. Zhang L, Cui X, Schmitt K, Hubert R, Navidi W and Arnheim N (1992) Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci USA* **89**: 5847–5851.
44. Pirker C, Raidl M, Steiner E, *et al*. (2004) Whole genome amplification for CGH analysis: Linker-adapter PCR as the method of choice for difficult and limited samples. *Cytometry* **61A**: 26–34.
45. Barker DL, Hansen MS, Faruqi AF, *et al*. (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res* **14**: 901–907.

Protocol

CONTENTS

Protocol 6.1: Target preparation

MATERIALS

Reagents for genomic DNA isolation.

Depending on the sample source, type and quantity, the genomic DNA isolation method can vary. Select the one optimal for each sample.

PCR reagents:

Hotstar Taq master mix (Qiagen. Cat. no. 203445)

Primers: 15 mM forward and reverse primers in DEPC-treated water

Genomic DNA samples

PCR product quantification and visualization:

Agilent DNA12000 labchip kit (Agilent. Cat. no. 5064-8231)

PCR product precipitation:

7.5 M ammonium acetate

100% ethanol

***In vitro* transcription reagents:**

T7 Megascript kit (Ambion, Inc. Austin, TX. Cat. no. 1334)

Target labeling reagents and material:

Low T dNTP (5 mM dA, dG and dCTP, 2mM dTTP)

1mM Fluorolink Cy3-dUTP and Fluorolink Cy5-dUTP (Amersham Biosciences Corp. Piscataway, NJ. Cat. no. PA53022 and PA55022)

Superscript II RNaseH$^-$ (with 5 $\times$ first strand buffer and 50 mM DTT) (Invitrogen Corp. Carlsbad, CA. Cat. no. 18064-07)

RNasin (20 units/μl) (Promega. Cat. no. N2111)

50 mM EDTA

1 M NaOH

Microbiospin 6 columns (Bio-Rad. Cat. no. 732-6222)

pd(N)6 (Boerhinger Mannheim. Cat. no. 1034731)

1 $\times$ TE

1 M Tris pH 7.5

Microcon YM-30 column (Millipore. Cat. no. 42410).

Hybridization reagents:

50 × Denhardt's blocking solution (Sigma. Cat. no. 2532)

Poly dA $_{40\text{-}60}$ (8 mg ml^{-1}) (Pharmacia. Cat. no. 27-7988-01)

Human Cot I DNA (1 mg ml^{-1}) (Invitrogen. Cat. no. 15279-011)

20 × SSC

10% SDS

Hybridization chambers (Corning. Cat. no. 2551)

Array scanner:

GenePix 4000B scanner (Axon Instrument)

METHODS

PCR reaction

PCR setting for generation of SNP-typing target.

PCR reaction mixture:

1 μl of 5′ T7-primer (15 μM)

1 μl of 3′ primer (15 μM)

10.5 μl of genomic DNA (containing approximately 50–100 ng of genomic DNA)

12.5 μl of hot start mixture

Total volume 25 μl

PCR profile:

95°C for 10min

96°C for 20 s

65°C for 45 s

72°C for 3 min

5 cycles

96°C for 20 s

60°C for 50 s

72°C for 3min

20 cycles

96°C for 20 s

55°C for 1 min

72°C for 3min

9 cycles

Run Agilent DNA chip (*Figure 6.1*).

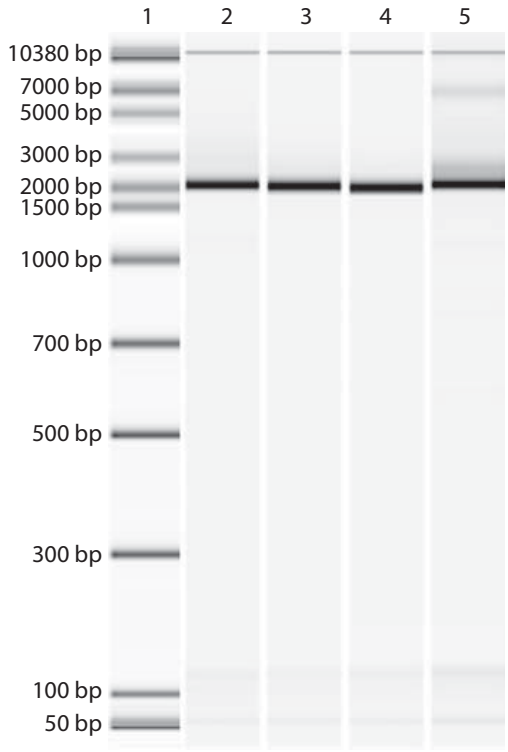


Figure 6.1.

Agilent Bioanalyzer DNA7500 chip analysis. A 2000-bp genomic DNA fragment of HLA A locus spanning exon 1 to exon 6 was PCR amplified using a T7-gene-specific primer pair. Lane 1, molecular weight marker; lane 2–5, genomic DNA fragments amplified from CL013, CL033, CL018 and CL096 cell line genomic DNA.

PCR product precipitation

Add 12.5 μl of 7.5 M ammonium acetate to the PCR product (25μl volume).

Add 100 μl of 100% EtOH.

Centrifuge at 13 000 *g* for 20 min at room temperature to avoid co-precipitation of primers.

Wash with 500 μl of 100% EtOH twice.

Dry pellet completely and then re-suspend in 10 μl of DEPC-treated water.

Check DNA amount by either Agilent Bioanalyzer or spectrophotometer.

In vitro transcription using T7 Megascript Kit

2 μl each of 75mM NTP (A, G, C and UTP)

2 μl reaction buffer

2 μl enzyme mix (RNase inhibitor and T7 phage RNA polymerase)

1 μg of PCR amplified DNA in 8 μl volume

Incubation at 37°C for 6 h.

Purification of amplified RNA

Any manufactured RNA isolation kit can be applied. A monophasic reagent such as TRIzol reagent from Invitrogen (cat. no. 15596026) is exemplified here based on the efficient recovery of aRNA. Other methods for RNA isolation can also be employed.

1. Add 1 ml of TRIzol solution to the transcription reaction. Mix the reagents well by pipetting or gentle vortexing.
2. Add 200 μl chloroform per ml of TRIzol solution. Mix the reagents by inverting the tube for 15 s. Allow the tube to stand at room temperature for 1–2 min.
3. Centrifuge the tube at 10 000 *g* for 15 min at 4°C.
4. Transfer the aqueous phase to a fresh tube and add 500 μl of isopropanol per ml TRIzol reagent.
5. Store the sample at room temperature for 5 min and then centrifuge at 13 000 *g* for 20 min.
6. Wash the pellet twice with 1 ml 70% EtOH.
7. Allow the pellet to dry in air and then dissolve it in 30 μl of DEPC H_2O.
8. Measure the quantity of RNA using the Agilent Bioanalyzer RNA 6000 chip (*Figure 6.2*).

Target labeling by reverse transcription

4 μl First strand buffer

1 μl dN6 primer (8 μg μl^{-1})

2 μl 10 × lowT-dNTP (5 mM A, C and GTP, 2 mM dTTP)

2 μl Cy-dUTP (1 mM Cy3 or Cy5)

2 μl 0.1 M DTT

1 μl RNasin

3 μg amplified RNA in 8 μl DEPC H_2O

Mix well and heat to 70°C for 3 min then cool down to 42°C.

Add 1 μl SSII. Incubate for 30 min at 42°C and add another 1 μl SSII for 40 min at 42°C. Add 2.5 μl 500 mM EDTA and heat to 65°C for 1min. Add 5 μl 1 M NaOH and incubate at 65°C for 15 min to hydrolyze the RNA. Add 12.5 μl 1 M Tris immediately to neutralize the pH. Bring the volume to 70 μl by adding 35 μl of 1 × TE.

Note: The amounts of aRNA used for labeling depend on the size of the array. If the array

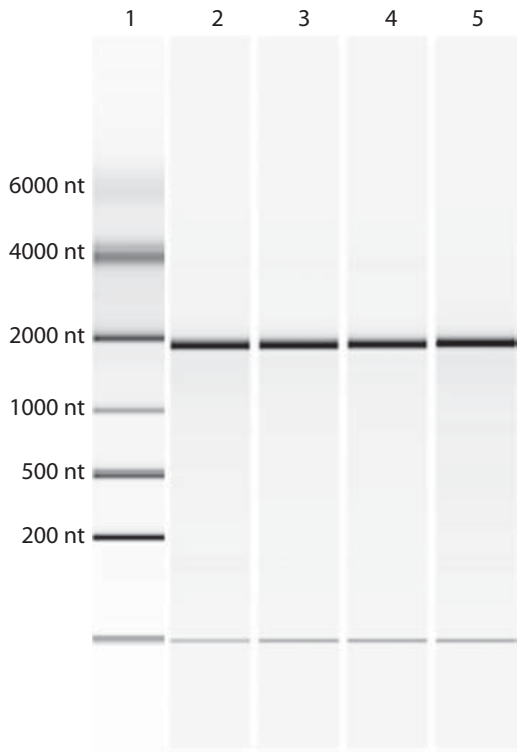


Figure 6.2.

Agilent Bioanalyzer RNA 6000 chip analysis. A 2000-nt RNA fragment corresponding to the HLA A locus from exon 1 to exon 6 was amplified using the T7 Megascript kit. Lane 1, RNA ladder; lane 2–4, amplified RNA using CL013, CL033, CL018 and CL096 cell line genomic DNA fragment as templates.

contains 2000–8000 oligo probes, 3 μg aRNA will be sufficient while a larger chip such as one containing 16–20 k oligo probes will need 6 μg of aRNA. The labeling reaction components do not need to be changed.

Target clean up: Prepare a Bio-6 column and apply the target solution through it according to the manufacturer's instructions. Collect flow through and add 250 μl 1 × TE to it. Concentrate target to around 20 μl using a Microcon YM-30 column.

Hybridization: Combine Cy3-labeled reference sample and Cy5-labeled test target (adjust the color to purple in order to balance the amount of test and reference samples) and then completely dry the sample using a Speedvac. Re-suspend the pellet in 25 μl volume by adding 1 μl 50 × Denhardt's blocking solution, 1 μl poly dA (8 μg/μl), 1 μl yeast tRNA (4 mg/ml), 10 μl Human Cot I DNA (1 mg/ml), 3 μl 20 × SSC, 0.6 μl of 10% SDS and 8.4 μl of DEPC-treated water. Heat the solution for 2 min at 99°C and apply this target mixture to the slide, add a coverslip, place the slide into a humidified hybridization chamber (Corning. Cat. no. 2551), and hybridize at 45°C overnight.

Slide washing:

1. Wash with 2 × SSC + 0.1% SDS to get rid of the cover slip.
2. Wash with 1 × SSC for 1 min.
3. Wash with 0.2 × SSC for 1 min.
4. Wash with 0.05 × SSC for 10 s.
5. Centrifuge slide at 80–100 *g* for 3 min (the slide can be put in a slide rack on a microplate carrier or in a 50-ml conical tube and centrifuged in a swinging-bucket rotor).

Scan slide.

From gene chips to disease chips – new approach in molecular diagnosis of eye diseases

7

Rando Allikmets and Jana Zernant

7.1 Introduction

Inherited retinal degenerations account for a substantial fraction of blindness in children and young adults and represent clinically and genetically heterogeneous disorders. On one end of the genetic spectrum are retinal disease phenotypes associated with one gene. For example, *ABCA4* (*ABCR*) is the causal gene for autosomal recessive (ar) Stargardt macular dystrophy (arSTGD). In addition to arSTGD, at least three more different retinal disease phenotypes; cone-rod dystrophy (arCRD), retinitis pigmentosa (arRP), and age-related macular degeneration (AMD) are caused by mutations in this gene. Due to the size (*ABCA4* contains 50 exons) and a substantial genetic heterogeneity (>450 known mutations), this gene presents an extremely difficult target for genetic analysis and diagnostic applications.

On the other end are 'multigenic' diseases such as RP, where mutations in more than 30 genes can cause the same RP phenotype (estimated prevalence 1:3500), making it impossible to predict the specific gene underlying the disease in a patient based on a clinical examination. For example, the early-onset form of RP, Leber congenital amaurosis (LCA), can be caused by more than 300 mutations in at least six genes, which together account for less than 50% of the disease load. Therefore, it is not surprising that the current management of patients with retinal degenerations relies on clinical examination, electrophysiology and other ancillary tests, since available methodology does not allow for an efficient, comprehensive, and cost-effective genetic screening of patients, who are often left with no specific information on their genotype.

To overcome these limitations, we developed genotyping microarrays for *ABCA4* ('gene array') and for LCA ('disease array'), representing comprehensive and cost-effective screening tools. Arrays were designed utilizing a method called solid-phase minisequencing or arrayed primer extension (APEX), which has been developed for high-throughput detection of nucleotide variations (1, 2). The APEX approach can be successfully applied for the detection of single nucleotide polymorphisms (SNPs), as well as any

deletions and insertions in heterozygous and homozygous patient samples. The designed arrays contain all currently known disease-associated genetic variants (mutations) in *ABCA4* and in all known LCA genes for one-step screening of patients with STGD, CRD, RP and LCA. Both arrays are more than 99% effective in screening for known mutations, can be easily updated with new variants, and are used for highly efficient, accurate, and affordable screening of patients.

In the following chapter, we will summarize the application of APEX technology for genotyping large cohorts of patients with various eye diseases. We will also show how it allows a systematic detection and analysis of genetic variation, which facilitates proper diagnosis, results in more precise prognosis of the disease progression, helps in genetic counseling for family members and, eventually, allows the suggestion of emerging therapeutic options.

7.2 APEX – arrayed primer extension

APEX is a rapid solid-phase genotyping method that combines the efficiency of a microarray-based assay with Sanger sequencing. In APEX, a DNA microarray (a 'chip') of sequence- (mutation-) specific detection oligonucleotides is used to determine the genotypes in a sample DNA. For each position of interest (e.g. a variable site/SNP) in the sample DNA, two 25-mer oligonucleotides (primers) are synthesized according to the wildtype sequence in both sense and antisense directions. The primers are usually designed with their 3′ ends immediately adjacent to the site of interest. The oligonucleotides are arrayed and attached to an amino-activated glass surface via an amino linker at their 5′ end with an automated arrayer.

The sample DNA is PCR-amplified in a single or multiplex reaction. All PCR products to be applied to one chip are pooled and purified together. The size of PCR products is not important, because all PCR products will be fragmented before APEX reaction to an optimal size of around 100–200 bp (for subsequent hybridization reaction) by replacing a fraction of dTTP by dUTP in the amplification mix, followed by treatment with thermolabile uracil-N-glycosylase (UNG; Epicentre Technologies, Madison, WI). UNG is highly specific to uracil bases in the DNA; the extent of fragmentation can be, therefore, controlled by the fraction of dUTP incorporation during PCR. The APEX reaction is reliable only if no dNTPs are carried over from the amplification mix, so the dNTP leftover is removed enzymatically by shrimp alkaline phosphatase, in a one-step reaction together with the UNG treatment (3).

Fragmented and heat-denatured PCR product-mix is applied to the chip together with fluorescently labeled ddNTPs (each of the four ddNTPs has a different label) and Thermo Sequenase™ DNA Polymerase (Amersham Biosciences, Piscataway, NJ). During a 15-min hybridization at 58°C, the target sample DNA fragments anneal to the detection primers on the chip immediately adjacent to the queried nucleotide. DNA polymerase extends the 3′ end of the primer with a labeled nucleotide analog complementary to the nucleotide of interest resulting in identification of one specific base in the target sequence (*Figure 7.1*). Covalent bonds between the oligonucleotides on the chip and the labeled terminator nucleotides allow a

stringent washing of the arrays after hybridization, to minimize the background (4). The signals are acquired by Genorama™ QuattroImager (Asper Biotech, Ltd.) and Image Pro Plus™ software (Media Cybernetics, Silver Spring, MD) and the genotypes are identified by Genorama™ Basecaller genotyping software (Asper Biotech, Ltd.; *Figure 7.2*).

An advantage of APEX, compared to purely hybridization-based technologies, is that all nucleotides of interest are identified with optimal discrimination at the same reaction conditions. APEX approach can be successfully applied to the detection of SNPs as well as deletions and insertions in hetero- and homozygous patient samples (*Figure 7.2*). APEX, performed in a single array format allows for at least one order of magnitude higher discrimination power between genotypes as compared to techniques that are purely hybridization-based (5). APEX technology, as described in this chapter, was developed and is currently provided by Asper Biotech, Ltd, Tartu, Estonia.

7.3 Application A – the gene array for *ABCA4*-associated retinal dystrophies

Several laboratories independently described *ABCA4 (ABCR)* in 1997 as the causal gene for Stargardt disease (STGD1, MIM 248200) (6–8). STGD1 is usually a juvenile-onset macular dystrophy associated with rapid central visual impairment, progressive bilateral atrophy of the foveal retinal

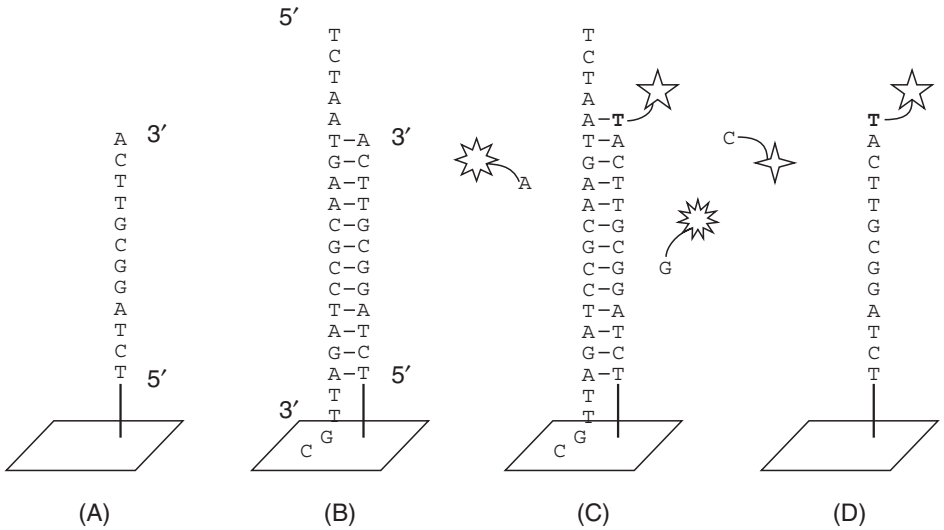


Figure 7.1.

Principle of APEX. (**A**) Oligonucleotides are arrayed on a glass slide via their 5′ end; (**B**) complementary fragment of PCR-amplified sample DNA is annealed to oligos; (**C**) sequence-specific single nucleotide extension of the 3′ ends of primers with dye-labeled nucleotide analogs (ddNTPs) by DNA polymerase; (**D**) sample DNA fragments and not incorporated ddNTPs are washed off followed by signal detection. The dye-labeled nucleotide, T (shown in bold), bound to the oligonucleotide on the slide is the nucleotide being typed.

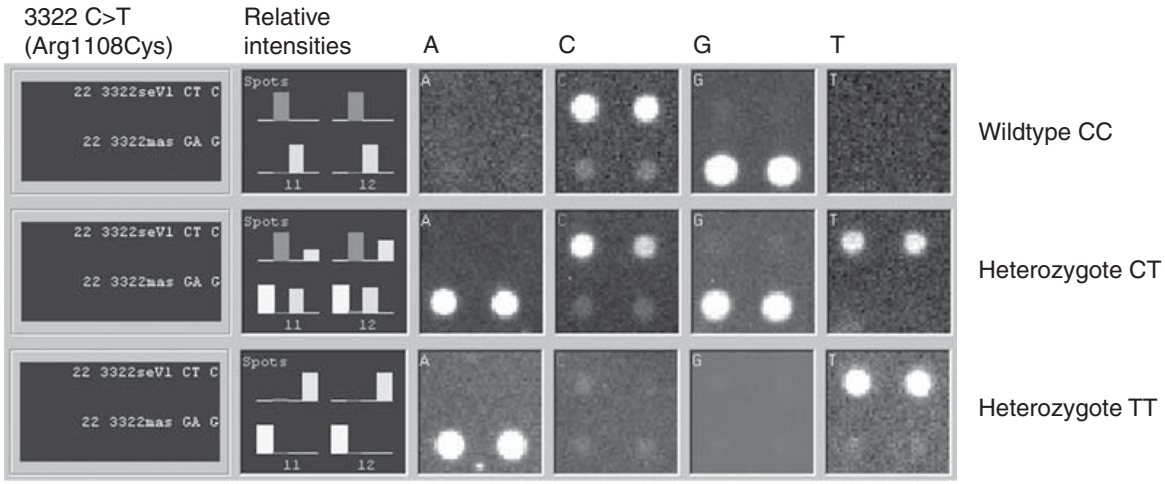


Figure 7.2.

Three possible different genotypes of the same ABCA4 variant on the ABCR400 chip detected by Genorama™ Basecaller genotyping software. The software compares fluorescence intensities (shown as bars in the second cell from left) of four different labels in each spot pair and translates them into the presence or absence of nucleotide(s) in the given position on the array. Every position is queried from both strands, the nucleotide(s) in sense and antisense strand appear as duplicate spots in the upper and lower row of the software window, respectively.

pigment epithelium, and the frequent appearance of yellowish flecks around the macula and/or in the central and near-peripheral areas of the retina. Subsequently, *ABCA4* mutations were identified and co-segregated with retinal dystrophies of substantially different phenotypes, such as autosomal recessive cone-rod dystrophy (arCRD) (9, 10) and atypical autosomal recessive retinitis pigmentosa (arRP, RP19) (9, 11, 12).

Disease-associated *ABCA4* alleles have shown an extraordinary heterogeneity (6, 13–17). Currently over 450 disease-associated *ABCA4* variants have been identified (R. Allikmets and J. Zernant, unpublished data), allowing comparison of this gene to one of the best-known members of the ABC superfamily, *CFTR*, encoding the cystic fibrosis transmembrane conductance regulator (18). What makes *ABCA4* a more difficult diagnostic target than *CFTR* is that the most frequent disease-associated *ABCA4* alleles, for example G1961E, G863A/delG863, and A1038V, have each been described in only around 10% of STGD patients in a distinct population, whereas the delF508 allele of *CFTR* accounts for close to 70% of all cystic fibrosis alleles (19).

Allelic heterogeneity has substantially complicated genetic analyses of *ABCA4*-associated retinal disease. Even in the case of STGD1, where the role of the *ABCA4* gene is indisputable, the mutation detection rate has ranged from around 25% (15, 20) to around 55–60% (13, 14, 17, 21). In each of these studies, conventional mutation detection techniques such as single strand conformational polymorphism (SSCP), heteroduplex analysis, and denaturing gradient gel electrophoresis (DGGE) were applied. Direct sequencing, which is still considered the 'gold standard' of all mutation

detection techniques, enabled a somewhat higher percentage of disease-associated alleles to be identified, from 66% to 80% (22, 23).

To overcome these challenges and to generate a high-throughput, cost-effective screening tool, we developed the ABCA4 genotyping microarray (24). By systematic analysis of all published, reported, and communicated data, we compiled the most comprehensive database of *ABCA4* variants, where only those sequence changes currently considered disease-associated exceed 400. By design, we included on this chip all variants from the coding region of *ABCA4* and adjacent intronic sequences. The overall efficiency of the array was enhanced by designing primers with mismatched or modified bases for several variants where *ABCA4* sequence presented additional challenges, that is hairpin loops, repeats, etc. Currently, from more than 400 variants only three (<1%) remain undetected by the last version of the chip; 93% of all variants are detected from both strands, whereas around 7% are detected reliably from one strand.

The array was validated on an extensive cohort of 136 confirmed STGD samples, which we had previously screened by SSCP and/or heteroduplex analyses (13). The initial SSCP screening had detected 55% of all disease-associated alleles. The microarray screening detected numerous additional alleles, bringing the total to more than 70% of all disease-associated alleles (24). Further evaluation of the ABCR400 array by screening several previously not analyzed STGD patient cohorts of diverse ethnicity (European American, Italian, Dutch, Hungarian and Slovenian) is summarized as *Table 7.1*. The screening efficiency of the ABCR400 microarray was remarkably similar in all six cohorts, yielding from around 53% to 60% of all possible disease-associated *ABCA4* alleles (*Table 7.1*).

Table 7.1. Screening efficiency of the ABCR400 array on several cohorts

STGD patient cohort	Chromosomes analyzed	Disease-associated alleles (%)	Allele distribution[a]
North America 1	300	158 (52.7%)	2–34% 1–39% 0–27%
Italy	62	34 (54.8%)	2–33% 1–45% 0–22%
Hungary	72	40 (55.6%)	2–47% 1–18% 0–35%
North America 2	40	24 (60%)	2–45% 1–30% 0–25%
The Netherlands	36	20 (55.6%)	2–39% 1–33% 0–28%
Slovenia	28	15 (53.6%)	2–29% 1–50% 0–21%

[a] Shows the percentage of the screened patients with both disease-alleles found (2), with one allele found (1) or no alleles found (0).

The allele distribution was also similar across all cohorts (*Table 7.1*, last column). We detected both disease-associated alleles in between 29% and 45% patients (average 36.6%). The fraction of patients with no apparent STGD alleles detected ranged from 21% to 35% (average 26.4%), most likely indicating inclusion of phenocopies, which cannot be avoided completely due to the selection methods. More robust results were obtained on smaller, carefully characterized, cohorts derived from a single clinical source (i.e., Italian), or cohorts with less allelic heterogeneity (Hungarian).

In summary, the ABCR400 array alone determined 55–65% of all possible disease-associated *ABCA4* alleles and, in combination with SSCP analysis, 70–78% of disease-associated alleleles in random cohorts of Stargardt disease patients. These results suggest that: (i) the ABCAY400 array is an efficient screening tool for known variants; and (ii) its efficiency for screening patient populations with STGD is comparable to direct sequencing. The ABCA4 array supplies two major applications: (i) (pre-)screening of all patients with suspected *ABCA4*-associated retinal pathology; including diagnostic screening of patients with Stargardt disease and cone-rod dystrophy; and (ii) high throughput, cost-efficient, and single-standard screening of large cohorts in case-control association studies, for example, for the AMD complex trait.

7.4 Application B – the 'disease array' for a genetically heterogeneous disorder (LCA)

Leber congenital amaurosis was named after the German ophthalmologist Theodor von Leber who in 1869 first described severe visual loss present at birth accompanied by nystagmus, sluggish pupillary reaction and pigmentary retinopathy. A detailed description of LCA-defining clinical signs has been extensively presented in many reviews; albeit a severe and early-onset disease, LCA nevertheless presents with variable expression, which can be sometimes explained by molecular genetic findings (see below). Difficulties with the clinical classification of LCA cases were most prominently demonstrated in a study where 30 out of 75 patients had been initially misdiagnosed (25). All the above further emphasizes the importance of a comprehensive molecular genetic analysis in addition to a thorough clinical evaluation.

The six known LCA genes and their protein products reveal extensive heterogeneity. Together, variants in aryl hydrocarbon receptor-interacting protein-like 1 (*AIPL1*) (26), Crumbs homolog 1 (*CRB1*) (27), cone-rod homeobox (*CRX*) (28), guanylate cyclase 2D (*GUCY2D* or *retGC*) (29), retinal pigment epithelium-specific 65-kDa protein (*RPE65*) (30, 31) and retinitis pigmentosa GTPase regulator-interacting protein 1 (*RPGRIP1*) (32) account for less than 50% of all LCA cases.The other loci implicated in LCA include *LCA3* on 14q24 (33), *LCA5* on 6q11-q16 (34, 35) and *LCA9* on 1p36 (36).

Reports from various laboratories vary widely in the percentage of the disease load assigned to each gene (*Table 7.2*), which may be due to different ascertainment criteria, differences in screening methods and their sensitivity, differences in the ethnic composition of screened cohorts or, most likely, a combination of all the above. One way to alleviate this

Table 7.2. Composition and screening efficiency of the LCA 'disease array'

LCA gene	Exons	Amplicons[a]	Known mutations	Mutation frequency in LCA (%)[b]	LCA array (%)
AIPL1	6	6	25	5.8 (38)	7.8
CRB1	12	13	68	9.0–13.5 (27, 39)	5.4
CRX	3	3	29	2.0 (28), 2.8 (40)	1.5
GUCY2D	20	14	67	6.0 (41, 42), 20.3 (43)	11.7
RPE65	14	10	81	6.8 (41), 8.2 (44), 11.4 (45), 15.6 (46)	2.4
RPGRIP1	24	18	32	5.3 (32), 5.6 (47)	4.9
Total	**79**	**64**	**302**		**33.7**

[a] 'Amplicons' shows the number of PCR products needed to amplify for each gene in order to screen for all known mutations.
[b] The percentage of each gene in the total disease load. Numbers in parentheses refer to the References.

problem is by comprehensive screening of the same, large patient cohort for allelic variation with an efficient methodology. In order to completely screen the six genes, one would need to analyze close to 80 amplicons (*Table 7.2*), which is a labor-intensive and expensive task, especially if applied simultaneously to hundreds of patients. To overcome these limitations, we designed a genotyping microarray (disease chip) for LCA, which includes all of the more than 300 variants currently described in the six genes, allowing for the detection of all known LCA-associated variants (mutations) in any DNA sample in one simple reaction (37). Every known disease-associated sequence change described in all LCA genes, and a small selection of common polymorphisms (for haplotype analysis), was included on the chip via sequence-specific oligonucleotides. The array was validated by screening around 100 confirmed LCA patients with known mutations and the efficiency of the chip was determined by screening more than 200 LCA cases from three independently ascertained cohorts, followed by segregation analyses in families, if applicable. The microarray is more than 99% effective in determining the existing genetic variation and yielded at least one disease-associated allele in about one-third of LCA patients (*Table 7.2*). This fraction will grow as new genes and mutations will be added to the chip.

Screening with the LCA array resulted in an additional intriguing finding: more than two (expected) variants were detected in a substantial fraction of patients, suggesting a multi-allelic inheritance or a modifier effect from more than one gene. In support of this hypothesis, the third allele segregated with a more severe disease phenotype in several families (Allikmets *et al.*, unpublished observation). In summary, the LCA genotyping microarray is a robust, comprehensive, and cost-effective screening tool, representing the first 'disease chip'. Simultaneous screening for all known LCA-associated variants in large LCA cohorts allows a systematic detection and analysis of genetic variation leading to an exact molecular diagnosis, which is often helpful in predicting the disease progression and facilitates selecting patients for clinical trials.

7.5 Summary

We have demonstrated on specific examples, that the APEX method is applicable for designing reliable, high-throughput and affordable genotyping tools for medical genetics. These arrays allow efficient detection of all known genetic variation underlying pathology in a gene or a group of genes. APEX arrays are especially useful in the following applications:

(i) They can be efficiently applied in an average academic laboratory with a limited budget, since they require only a relatively moderate investment.
(ii) They are cost-effective in situations where a few hundred SNPs have to be screened in several hundreds to several thousands of patients.
(iii) They are irreplaceable in precise diagnostic applications, where *every* known mutation, including deletions, insertions, and so on, *has* to be detected by an assay.

Diagnostic tools, similar to the ABCR400 and LCA arrays, should be made available for all genes involved in the entire range of eye diseases. For example, an array for retinitis pigmentosa, including the entire genetic variation in more than 30 known RP loci, would substantially enhance our diagnostic capabilities. More importantly, it would allow a precise determination of the causal genetic defect(s) hopefully followed by the suggestion of a therapeutic option in the very near future.

Acknowledgements

The authors sincerely appreciate support and advice from many colleagues and collaborators, especially from Asper Biotech – Drs. N. Tõnisson, K. Jaakson, and M. Külm. This study was supported, in part, by NIH Grant EY13435 and Research to Prevent Blindness.

References

1. Syvanen AC, Sajantila and A, Lukka M (1993) Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing. *Am J Hum Genet* **52:** 46–59.
2. Syvanen AC, Aalto-Setala K, Harju L, Kontula K and Soderlund H (1990) A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8:** 684–692.
3. Kurg A, Tonisson N, Georgiou I, Shumaker J, Tollett J and Metspalu A (2000) Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet Test* **4:** 1–7.
4. Tonisson N, Zernant J, Kurg A, Pavel H, Slavin G, Roomere H, Meiel A, Hainaut P and Metspalu A (2002) Evaluating the arrayed primer extension resequencing assay of TP53 tumor suppressor gene. *Proc Natl Acad Sci USA* **99:** 5503–5508.
5. Pastinen T, Kurg A, Metspalu A, Peltonen L and Syvanen AC (1997) Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res* **7:** 606–614.
6. Allikmets R, Singh N, Sun H, *et al.* (1997) A photoreceptor cell-specific ATP-binding transporter gene (ABCAY) is mutated in recessive Stargardt macular dystrophy. *Nat Genet* **15:** 236–246.

7. Azarian SM and Travis GH (1997) The photoreceptor rim protein is an ABC transporter encoded by the gene for recessive Stargardt's disease (ABCR). *FEBS Lett* **409:** 247–252.
8. Illing M, Molday LL and Molday RS (1997) The 220-kDa rim protein of retinal rod outer segments is a member of the ABC transporter superfamily. *J Biol Chem* **272:** 10303–10310.
9. Cremers FP, van de Pol DJ, van Driel M, *et al.* (1998) Autosomal recessive retinitis pigmentosa and cone-rod dystrophy caused by splice site mutations in the Stargardt's disease gene ABCR. *Hum Mol Genet* **7:** 355–362.
10. Maugeri A, Klevering BJ, Rohrschneider K, Blankenagel A, Brunner HG, Deutman AF, Hoyng CB and Cremers FP (2000) Mutations in the *ABCA4* (ABCR) gene are the major cause of autosomal recessive cone-rod dystrophy. *Am J Hum Genet* **67:** 960–966.
11. Martinez-Mir A, Paloma E, Allikmets R, Ayuso C, del Rio T, Dean M, Vilageliu L, Gonzalez-Duarte R and Balcells S (1998) Retinitis pigmentosa caused by a homozygous mutation in the Stargardt disease gene ABCR. *Nat Genet* **18:** 11–12.
12. Rozet JM, Gerber S, Ghazi I, Perrault I, Ducroq D, Souied E, Cabot A, Dufier JL, Munnich A and Kaplan J (1999) Mutations of the retinal specific ATP binding transporter gene (ABCR) in a single family segregating both autosomal recessive retinitis pigmentosa RP19 and Stargardt disease: evidence of clinical heterogeneity at this locus. *J Med Genet* **36:** 447–451.
13. Lewis RA, Shroyer NF, Singh N, Allikmets R, Hutchinson A, Li Y, Lupski JR, Leppert M and Dean M (1999) Genotype/phenotype analysis of a photoreceptor-specific ATP-binding cassette transporter gene, ABCR, in Stargardt disease. *Am J Hum Genet* **64:** 422–434.
14. Maugeri A, van Driel MA, van de Pol DJ, *et al.* (1999) The 2588G—>C mutation in the ABCAY gene is a mild frequent founder mutation in the Western European population and allows the classification of ABCR mutations in patients with Stargardt disease. *Am J Hum Genet* **64:** 1024–1035.
15. Fishman GA, Stone EM, Grover S, Derlacki DJ, Haines HL and Hockey RR (1999) Variation of clinical expression in patients with Stargardt dystrophy and sequence variations in the ABCR gene. *Arch Ophthalmol* **117:** 504–510.
16. Fumagalli A, Ferrari M, Soriani N, *et al.* (2001) Mutational scanning of the ABCR gene with double-gradient denaturing-gradient gel electrophoresis (DG-DGGE) in Italian Stargardt disease patients. *Hum Genet* **109:** 326–338.
17. Rivera A, White K, Stohr H, *et al.* (2000) A comprehensive survey of sequence variation in the *ABCA4* (ABCR) gene in Stargardt disease and age-related macular degeneration. *Am J Hum Genet* **67:** 800–813.
18. Riordan JR, Rommens JM, Kerem B, *et al.* (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245:** 1066–1073.
19. Zielenski J and Tsui LC (1995) Cystic fibrosis: genotypic and phenotypic variations. *Annu Rev Genet* **29:** 777–807.
20. Webster AR, Heon E, Lotery AJ, *et al.* (2001) An analysis of allelic variation in the *ABCA4* gene. *Invest Ophthalmol Vis Sci* **42:** 1179–1189.
21. Simonelli F, Testa F, de Crecchio G, Rinaldi E, Hutchinson A, Atkinson A, Dean M, D'Urso M and Allikmets R (2000) New ABCR mutations and clinical phenotype in Italian patients with Stargardt disease. *Invest Ophthalmol Vis Sci* **41:** 892–897.
22. Shroyer NF, Lewis RA, Yatsenko AN, Wensel TG and Lupski JR (2001) Cosegregation and functional analysis of mutant ABCR (*ABCA4*) alleles in families that manifest both Stargardt disease and age-related macular degeneration. *Hum Mol Genet* **10:** 2671–2678.

23. Yatsenko AN, Shroyer NF, Lewis RA and Lupski JR (2001) Late-onset Stargardt disease is associated with missense mutations that map outside known functional regions of ABCR (*ABCA4*). *Hum Genet* **108:** 346–355.
24. Jaakson K, Zernant J, Kulm M, *et al.* (2003) Genotyping microarray (gene chip) for the ABCR (*ABCA4*) gene. *Hum Mutat* **22:** 395–403.
25. Lambert SR, Kriss A, Taylor D, Coffey R and Pembrey M (1989) Follow-up and diagnostic reappraisal of 75 patients with Leber's congenital amaurosis. *Am J Ophthalmol* **107:** 624–631.
26. Sohocki MM, Bowne SJ, Sullivan LS, *et al.* (2000) Mutations in a new photoreceptor-pineal gene on 17p cause Leber congenital amaurosis. *Nat Genet* **24:** 79–83.
27. den Hollander AI, Heckenlively JR, van den Born LI, *et al.* (2001) Leber congenital amaurosis and retinitis pigmentosa with Coats-like exudative vasculopathy are associated with mutations in the *crumbs homologue 1* (*CRB1*) gene. *Am J Hum Genet* **69:** 198–203.
28. Freund CL, Gregory-Evans CY, Furukawa T, *et al.* (1997) Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (*CRX*) essential for maintenance of the photoreceptor. *Cell* **91:** 543–553.
29. Perrault I, Rozet JM, Calvas P, *et al.* (1996) Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis. *Nat Genet* **14:** 461–464.
30. Gu SM, Thompson DA, Srikumari CR, *et al.* (1997) Mutations in *RPE65* cause autosomal recessive childhood-onset severe retinal dystrophy. *Nat Genet* **17:** 194–197.
31. Marlhens F, Bareil C, Griffoin JM, *et al.* (1997) Mutations in *RPE65* cause Leber's congenital amaurosis. *Nat Genet* **17:** 139–141.
32. Dryja TP, Adams SM, Grimsby JL, McGee TL, Hong DH, Li T, Andreasson S and Berson EL (2001) Null *RPGRIP1* alleles in patients with Leber congenital amaurosis. *Am J Hum Genet* **68:** 1295–1298.
33. Stockton DW, Lewis RA, Abboud EB, Al-Rajhi A, Jabak M, Anderson KL and Lupski JR (1998) A novel locus for Leber congenital amaurosis on chromosome 14q24. *Hum Genet* **103:** 328–333.
34. Dharmaraj S, Li Y, Robitaille JM, Silva E, Zhu D, Mitchell TN, Maltby LP, Baffoe-Bonnie AB and Maumenee IH (2000) A novel locus for Leber congenital amaurosis maps to chromosome 6q. *Am J Hum Genet* **66:** 319–326.
35. Mohamed MD, Topping NC, Jafri H, Raashed Y, McKibbin MA and Inglehearn CF (2003) Progression of phenotype in Leber's congenital amaurosis with a mutation at the *LCA5* locus. *Br J Ophthalmol* **87:** 473–475.
36. Keen TJ, Mohamed MD, McKibbin M, Rashid Y, Jafri H, Maumenee IH and Inglehearn CF (2003) Identification of a locus (LCA9) for Leber's congenital amaurosis on chromosome 1p36. *Eur J Hum Genet* **11:** 420–423.
37. Allikmets R, Zernant J, Kulm M, *et al.* (2003) Genotyping microarray (disease chip) for Leber congenital amaurosis [ARVO Abstract]. *Invest Ophthalmol Vis Sci* **44:** 2851.
38. Sohocki MM, Perrault I, Leroy BP, *et al.* (2000) Prevalence of AIPL1 mutations in inherited retinal degenerative disease. *Mol Genet Metabol* **70:** 142–150.
39. den Hollander AI, ten Brink JB, de Kok YJ, *et al.* (1999) Mutations in a human homologue of *Drosophila crumbs* cause retinitis pigmentosa (RP12). *Nat Genet* **23:** 217–221.
40. Swain PK, Chen S, Wang QL, *et al.* (1997) Mutations in the cone-rod homeobox gene are associated with the cone-rod dystrophy photoreceptor degeneration. *Neuron* **19:** 1329–1336.
41. Lotery AJ, Namperumalsamy P, Jacobson SG, *et al.* (2000) Mutation analysis of 3 genes in patients with Leber congenital amaurosis. *Arch Ophthalmol* **118:** 538–543.

42. Dharmaraj SR, Silva ER, Pina AL, *et al*. (2000) Mutational analysis and clinical correlation in Leber congenital amaurosis. *Ophthalm Genet* **21:** 135–150.
43. Perrault I, Rozet JM, Gerber S, *et al*. (2000) Spectrum of retGC1 mutations in Leber's congenital amaurosis. *Eur J Hum Genet* **8:** 578–582.
44. Simovich MJ, Miller B, Ezzeldin H, Kirkland BT, McLeod G, Fulmer C, Nathans J, Jacobson SG and Pittler SJ (2001) Four novel mutations in the *RPE65* gene in patients with Leber congenital amaurosis. *Hum Mutat* **18:** 164.
45. Thompson DA, Gyurus P, Fleischer LL, *et al*. (2000) Genetics and phenotypes of RPE65 mutations in inherited retinal degeneration. *Invest Ophthalmol Vis Sci* **41:** 4293-4299.
46. Morimura H, Fishman GA, Grover SA, Fulton AB, Berson EL and Dryja TP (1998) Mutations in the *RPE65* gene in patients with autosomal recessive retinitis pigmentosa or leber congenital amaurosis. *Proc Natl Acad Sci USA* **95:** 3088–3093.
47. Gerber S,, Perrault I, Hanein S, *et al*. (2001) Complete exon-intron structure of the RPGR-interacting protein (*RPGRIP1*) gene allows the identification of mutations underlying Leber congenital amaurosis. *Eur J Hum Genet* **9:** 561–571.

Protocol

CONTENTS

Protocol 7.1: Template preparation

POLYMERASE CHAIN REACTION (PCR)

1. Prepare a PCR premix by combining the following reagents in 15 μl (final concentrations are given):

   MilliQ water

   1 × PCR buffer (Solis BioDyne, Estonia)

   2.5 mM magnesium chloride

   0.2 mM dNTP (20% of the dTTP fraction substituted by dUTP)

   15 pmol of forward and reverse primer

   20 ng of genomic DNA

   1 U Taq DNA polymerase (Solis BioDyne, Estonia).

2. Cycling conditions in the thermal cycler: denaturing at 95°C for 12 minutes, followed by 26 cycles of denaturation at 95°C for 15 s, stepdown annealing at 68°C/–0.5°C per cycle for 20 s, and extension at 72°C for 45 s, with final extension at 72°C for 7 min.

3. Check the result by running 1/10 of each PCR reaction on a horizontal 1% agarose gel.

PCR PRODUCT PURIFICATION AND TREATMENT WITH URACIL N-GLYCOSYLATE (UNG) AND SHRIMP ALKALINE PHOSPHATASE (SAP)

1. Pool tested PCR products for one chip and purify 5–10 μg of the PCR product mix using a PCR product purification column (General Biosystem, South Korea). Elute the products from the column in 24 μl of MilliQ water.

2. Prepare the UNG-SAP reaction in 30 μl:

   1 × UNG buffer

   2 U thermolabile UNG

   1 U SAP

   24 μl of purified PCR products

   Incubate at 37°C for 1 h.

3. Check the efficiency of fragmentation by running 1/10 of UNG-SAP reaction on a horizontal 1% agarose gel after heating at 95°C for 10 min.

ARRAYED PRIMER EXTENSION (APEX)

1. Place the DNA microarray slide in a slide holder and rinse as follows:

   95°C distilled water for 30 s

   100 mM sodium hydroxide for 10 min

   95°C distilled water for 30 s, twice.

2. Denature and fragment the purified and UNG-SAP-treated PCR product mix at 95°C for 10 min after adding ThermoSequenase DNA Polymerase reaction buffer (1× final concentration).

3. Prepare the APEX reaction in 35 μl:

   Denatured and fragmented PCR products with 1× reaction buffer

   1.4 μM of each fluorescently labeled ddNTP: Texas Red-ddATP, fluoresein-ddGTP (Amersham Biosciences), Cy3-ddCTP, Cy5-ddUTP (NEN)

   4 U ThermoSequenase DNA Polymerase.

4. Apply the reaction mixture to a microarray slide, cover with a coverslip and incubate in a hybidization chamber at 58°C for 15 min.

5. Stop the reaction by washing the slide three times at 95°C in MilliQ water.

6. Read the slide with the Genorama™ QuattroImager and analyze the sequence variants by using Genorama™ Basecaller genotyping software (Asper, Ltd., *Figure 7.2*).

8 Multiplexed SNP genotyping using allele-specific primer extension on microarrays

Juha Saharinen, Pekka Ellonen, Janna Saarela and Leena Peltonen

8.1 Introduction

Systematic sequencing of the genomic DNA of multiple individuals from different populations has produced detailed information of a high number of single nucleotide variations across the human genome (1, 2). The single nucleotide polymorphisms (SNPs) are excellent genetic markers; when compared to the repeat polymorphisms, SNPs are more stable and evenly distributed across the genome (3). Currently over 9 million SNPs in the human genome are deposited to various databases, such as NCBI, dbSNP, HGVBase and the SNP Consortium (4–6). However, despite the overwhelming amount of identified SNPs in databases, only a fraction of them have been carefully validated and their allele frequency information in various populations determined (7, 8). *Table 8.1* presents some major SNP databases and validation efforts.

Table 8.1. SNP databases and large scale SNP validation effort

Database	Internet URL	Type	SNPs	Validated SNPs
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/	Non-profit	10.1 M	5.1 M
International HapMap Project	http://www.hapmap.org		1.0 M	1.0 M
Human Genome Variation Database	http://hgvbase.cgb.ki.se/	Non-profit	2.9 M	2.9 M
Celera Discovery System	http://www.celera.com/	Commercial	8.3 M	4.2 M
JSNP Database	http://snp.ims.u-tokyo.ac.jp/	Non-profit	0.2 M	0.2 M
Sequenom RealSNP	https://www.realsnp.com/	Commercial	5.4 M	0.22 M
Seattle SNPs	http://pga.mbt.washington.edu	Non-profit	25 676	25 676
Perlegen Genotype Data	http://genome.perlegen.com	Non-profit	1.5 M	1.5 M

SNP genotyping is most frequently used in applications involving fine mapping of specific genomic loci, for example in disease gene mapping projects and candidate gene association studies (reviewed in 9, 10). SNP markers have also been used to characterize the allelic diversity of specific genes in various pharmacogenomics projects (reviewed in 11–17) as well as in paternity testing and forensics (18, 19). Most of the SNPs are biallelic (two alternative nucleotides are known to exist for a given position) and thus the information content for SNP markers is much less than for microsatellite markers. This creates problems in the collection of meiotic information in linkage or association studies and full genome scans with SNPs have not been a realistic option. Even for most informative SNPs, typically three SNPs are needed to produce allelic information comparable to multiallelic markers and thus the option of SNP-based genome scans even in family samples has not been reasonable (20, 21). However, efficiency for high throughput SNP genotyping and accumulating information of the linkage disequilibrium intervals in different parts of the genome (the HapMap project) have made the SNP-based genome-wide association studies of human diseases an attractive option (22–24).

Suitability of microarrays as a genotyping platform

DNA microarray technology offers several advantages over classical homogenous laboratory assays. Thousands of probes or samples can be placed on a small microarray slide, thus facilitating multiplexed assays and decreased reagent costs due to the small reaction volumes. In multiplexed assays, several SNP loci are amplified simultaneously, and therefore the consumption of the often precious samples is reduced, allowing the investigator to run multiple analyses and thus gain more data. Ideal genotyping assays would have the throughput and suitability for efficient automation, parallelizing of the assay, low price of produced genotypes, robust and reliable allele calling and a high feasibility for data storage and transfer. Many intrinsic properties of microarrays make them suitable for massive genotyping projects. SNP genotyping microarrays can be manufactured different ways, including *in situ* synthesis or immobilization of the locus-/allele-specific oligonucleotides on the array and by using tag arrays, which act as hybridization partners for the allele-/locus-specific oligonucleotides, tailed with sequence complementary to the tag (28–31). Commercial microarrays for SNP genotyping are available from different vendors and include: Affymetrix GenFlex Tag Array and GeneChip Human Mapping Sets (http://www.affymetrix.com), Asper Biotech APEX (http://www.asperbio.com), Beckman Coulter SNPstream (http://www.beckman.com) and Illumina BeadChip (http://www.illumina.com).

Allele-specific primer extension on microarrays

Allele-specific primer extension is a method for the detection of a number of short allelic variations such as SNPs or mutations in a DNA sample and is well suited to be used in a multiplexed fashion. Multiplexed allele-specific primer extension on microarrays is achieved by simultaneous amplification of numerous loci in multiplexed polymerase chain reaction followed by a

hybridization step where the processed samples are applied on the microarray surface. In the hybridization reaction, the sample molecules carrying variant alleles are captured by the allele-specific oligonucleotides on the microarray surface. The principle of the allele-specific primer extension is illustrated in *Figure 8.1C, D, E*.

In the allele-specific primer extension presented here, a reverse transcriptase is used to extend the matched primer bound to the template in the presence of fluorescent nucleotides. Only a single fluorescent dye is used, and distinction between different alleles is produced by the complementarity of the 3′ nucleotide of the allele-specific oligonucleotide (ASO) attached to the array.

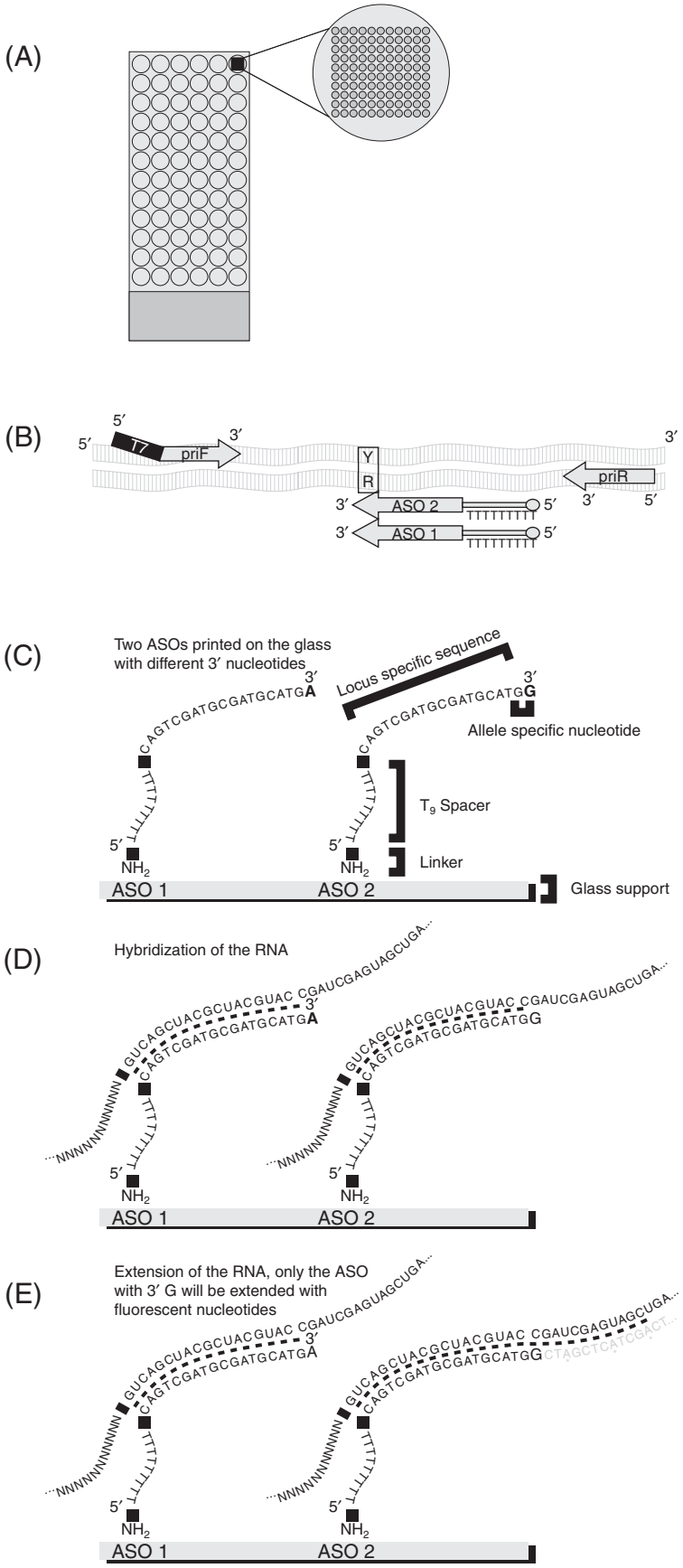
Allele-specific primer extension is a flexible genotyping method for medium throughput applications allowing detection of any kind of nucleotide variation, including insertions/deletions. The only limiting factor for allele-specific primer extension assay design is that ASOs must be designed to be specific for the locus of interest. Allele-specific primer extension has no limitations on nucleotide variants to be detected whereas many other methods suffer from this limitation in terms of single dye chemistries. While the multiplexing level of the PCR is limited, it can be extended by the pooling of several multiplexed PCR products. Since the invention of the method by Pastinen and Syvänen (30), the use of allele-specific primer extension on microarrays has been reported in disease gene mapping, mutation carrier screening and in supplementary paternity testing. Here we demonstrate how researchers are able to design and analyze an SNP microarray of choice rapidly and efficiently without spending time in extensive optimization efforts.

8.2 Practical approach on microarray based allele-specific primer extension

There are multiple ways to successfully accomplish the multiplexed allele-specific primer extension assays using microarray format. In the following section we describe the protocols used in our laboratory, which have been proven to be robust and suitable for the multiplexed SNP genotyping assays for medium throughput projects. The different steps and estimated time span in the multiplexed allele-specific primer extension genotyping assays are schematically presented in *Figure 8.2*.

Manufacturing microarrays for allele-specific primer extension

In genotyping microarrays, a probe is hybridized to a single sample (or to a pooled sample mixture), unlike in gene expression two-color arrays, in which two samples compete in the hybridization reaction. The genotyping microarrays can be used as regular microarrays, where the whole array surface is being used to monitor the hybridization of one sample. Alternatively, the array surface can be divided into subarrays, where each subarray is hybridized with a different sample (see *Figure 8.1A* for the array-of-arrays layout). In the latter case the same set of allele-specific oligos are printed on each subarray. We routinely use duplicate or even triplicate spots on each subarray, to guarantee the reliability of genotyping. Due to the


(A)
(B)
5′
T7
3′
priF
Y
R
priR
ASO 2
ASO 1
(C)
Two ASOs printed on the glass with different 3′ nucleotides
Locus specific sequence
Allele specific nucleotide
T9 Spacer
Linker
Glass support
ASO 1
ASO 2
(D)
Hybridization of the RNA
(E)
Extension of the RNA, only the ASO with 3′ G will be extended with fluorescent nucleotides

standardized automation, it is conventional to use the same well-to-well spacing as in the regular 384-well plates.

The design of arrays starts with the design of oligonucleotides for each bi-allelic SNP to be genotyped. PCR primers are designed for each amplicon as described in *Figure 8.1B*. For each SNP two ASOs are designed which define the alleles in the SNP locus. Each ASO comprises three structural elements (see *Figure 8.1C*). The first element of an ASO is an amine group at the 5′ terminus, which covalently binds the oligonucleotide to the chemically activated slide surface. The amine group is followed by a spacer sequence (for example TTT TTT TTT), which provides physical distance from the slide surface. The third element is a locus-specific sequence followed by the allele-specific sequence, which detects the variant alleles in the sample. The locus-specific sequence of an ASO is 16 to 22 nucleotides in length resulting in a homogenous melting temperature for all ASOs present in the hybridization assay. In the design of each ASO stringent physical parameters are followed, such as avoiding long homonucleotide repeats, secondary structures and self-dimerization. These parameters can be estimated by computer software, for example Oligonucleotide properties calculator (http://www.basic.nwu.edu/biotools/oligocalc.html). The two ASOs required for each SNP differ only in their 3′ nucleotides, which define the two alleles to be detected.

Microarray slides, carrying the ASOs on their surface, can be manufactured in a variety of ways. Contact printing is one of the most commonly used methods. In contact printing or ‘spotting’ a robotic

Figure 8.1.

Array-of-array layout, ASO oligonucleotides and PCR primers, allele-specific primer extension reaction. (**A**) Array-of-array layout. Each genotyped sample is applied to a subarray. The subarrays are spaced according to the well-to-well distances of 384-well plates. The subarrays are formed with tape-gridding or with a histological wax pen. All the subarrays are identical, containing the ASOs, two for each SNP locus and typically duplicated in order to increase the assay reliability.
(**B**) Localization of the different oligonucleotides around the SNP locus. The PCR primers (priF and priR) are designed to amplify a region of 100–200 bps around the SNP to be genotyped. In the forward primer (priF), a T7 promoter sequence is added to the 5′ end. The two ASOs, targeting the SNP (here Y, i.e. C or T alleles in the forward strand), correspond to a sequence in the upper strand.
(**C**) Composition of the ASOs. The ASO contains an amino-group at its 5′ end, required for the covalent attachment to the array surface. A linker region, typically T9, is required to provide a physical distance from the locus-specific sequence to the slide surface and to enhance the flexibility of the oligonucleotide. The two ASOs are identical in their sequence, except for their 3′ terminal nucleotide (either G or A, complementary to the SNP alleles). (**D**) Hybridization of the amplicon to the ASOs. The PCR-amplified region, containing the SNP to be genotyped, is transcribed to RNA and hybridized with the ASOs on the array. (**E**) Allele-specific primer extension. Depending on the alleles, either ASO1, ASO2 (homozygote) or both (heterozygote) is/are extended in the reverse-transcription reaction. Fluorescent nucleotides are incorporated to the extension product, and are required for the subsequent detection.

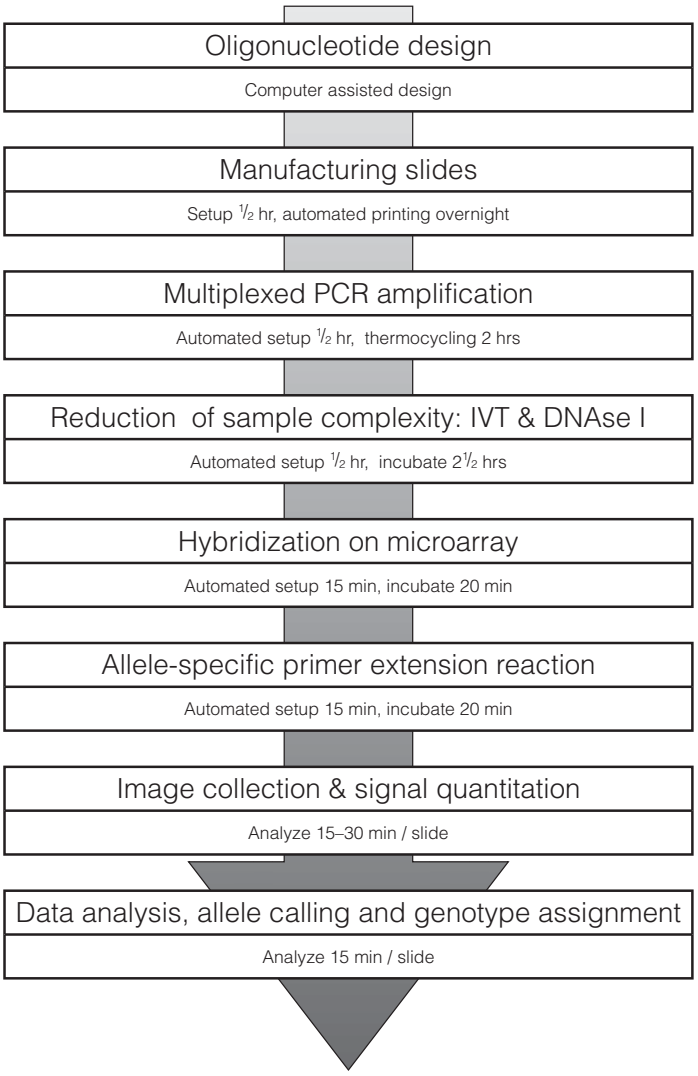


Figure 8.2.

Flow chart of the allele-specific primer extension genotyping and data analysis. Most of the steps are carried out in multi-well plates using pipetting robotics. (i) Once the SNPs to be genotyped are identified, the primers are designed for the locus, as described in *Figure 8.1B* and *Table 8.2*. (ii) The slides are manufactured with a robotic arrayer and the printing is done overnight, depending on the amount of arrays to be produced. (iii) The loci containing SNPs to be genotyped are multiplex PCR-amplified from the samples. PCR reactions are performed with touchdown annealing in about 2 h. (iv) In order to prevent self-pairing of the PCR products in the subsequent hybridization step, the PCR products are transcribed to RNA by T7 polymerase, followed by DNA template degradation using DNaseI. These reactions take about 2,5 h. (v) Hybridization of the RNAs to the ASOs is carried out on the microarray in a humid chamber and takes around 0.5 h. (vi) The allele-specific primer extension is carried out using reverse-transcriptase, for example MMLV-RT. The allele-specific extension incorporates the fluorescent nucleotides into the extension product. (vii) The fluorescent emission is detected using a standard microarray scanning instrument, producing an image, which is then quantitated. (viii) The quantitated image data is background-subtracted and normalized. The allele calling is done by clustering methods, followed by genotype assignments.

arrayer is used to transfer small volumes of ASOs from a microtiter plate onto a microscopic slide. These slides are aminosilane-coated for covalent binding with ASOs (36). The arrayer dips a quill pin into ASO solution containing 20 μM of oligonucleotide in a 1 × Micro Spotting Solution (ArrayIt Microarray Technology) and subsequently moves the pin over the slide. When all spots are printed for a given ASO, the pin is washed in an ultrasonic water bath washing station and vacuum-dried to prevent carry-over contamination between ASO spots. By repeating the cycle of dipping, printing and washing, the arrayer builds an array-of-arrays layout. With contact printing hundreds of spots can be replicated from a single dip. The produced spots are 100–500 μm in diameter depending on the size of the pin and surface chemistry used. The temperature and humidity of the printing unit also affect the printing process and extra care should be paid to control for this.

Sample preparation: PCR of the DNA samples

Selected genomic regions containing the SNPs to be monitored are amplified by PCR to yield a sufficient amount of DNA molecules for microarray-based detection of the SNP genotypes. PCR amplification of the sample is performed in a multiplexed fashion, that is all primer pairs are amplified simultaneously in a single-tube reaction, each primer pair producing a 100- to 200-nucleotide-long amplicon for the SNP locus. A feasible level of PCR multiplexing is up to 20 SNP loci in a single reaction. This can be extended by pooling different multiplex PCR products. The growing complexity of the oligonucleotide mixture in the multiplex PCR reactions typically results in around 80% of successful genotyping assays giving distinct genotype clusters in data analysis. Success rate is expected to decrease as the level of multiplexing increases.

Successful multiplexed PCR assay requires careful primer design, which takes into account uniform melting temperature and amplicon length for all primers. Parameters for oligonucleotide selection are shown in *Table 8.2*. In order to prevent mishybridization of sample DNA and oligonucleotides during the polymerase chain reaction, two different actions are taken. Firstly, a mispriming library containing known repetitive elements of the human genome, such as Alu repeats, is utilized, preventing the primers from targeting any known repetitive sequences. Secondly, cross binding to other targets in the multiplex PCR is prevented by including all the other multiplexed loci sequences and primers in the primer design process. This second step is then iterated for all the loci in the same multiplex PCR design. The multiplex PCR primer design system is accessible on our website at http://apps.bioinfo.helsinki.fi/mpd, where the actual underlying primer design algorithm is the Primer3 program (37). Each PCR product contains a T7 RNA polymerase promoter sequence (TAA TAC GAC TCA CTA TAG GGA GA) introduced by a T7-tagged forward primer, needed later for *in vitro* transcription, which is introduced by a tailing of the 5′ end of the PCR primer on the opposite strand of the ASOs (see *Figure 8.1B*).

Multiplexed PCR reactions are carried out in a microtiter plate format in a reaction volume of 5–20 μl using 1–20 ng of DNA as a template. Thermocycling is performed in a touchdown manner where the annealing

Table 8.2. PCR primer design parameters with Primer3

	Minimum value	Optimum value	Maximum value
PCR Primer size	18 nt	20 nt	23 nt
PCR Primer T_m value[a]	58°C	60°C	62°C
Primer maximum GC content		55%	
ASO oligonucleotide length	16	18	22
ASO oligonucleotide T_m	43	45	N/A

[a] T_m values calculated using the Nearest Neighbor method

temperature is decreased by 0.5–1°C during the first few cycles, which produces few specific copies of amplicons at the optimal annealing temperature. After touchdown cycling a final amplification is performed at the lowest annealing temperature of the primers in the assay.

Improving the specificity of the hybridization: reduction of template complexity

In order to avoid self–pairing of the PCR products, they are not directly used for hybridization with the ASOs. Rather both the specificity of the hybridization as well as number of target molecules is increased by transcription of the PCR products to single-stranded RNA molecules, using the T7 promoter sequence tailed on the forward PCR primer (see above). *In vitro* transcription is performed in a 4-µl reaction volume containing 2.0 µl of PCR template, 0.85 × T7 reaction buffer, 6.17 mM of each deoxyribonucleotides, 8.65 mM of DTT and 0.35 µl of T7 RNA polymerase solution (modified from Ampliscribe T7 High Yield Transcription Kit, Epicentre Biotechnologies).

The transcription reaction is followed by the degradation of the PCR products by the addition of 1.0 µl of DNaseI solution containing 0.1 U of DNAseI in 1 × T7 reaction buffer. All the enzymatic steps are easy to automate and can be carried out in a microtiter plate format in a reaction volume as low as 5 µl. This results in RNA target molecules that act as a template for the extension of the spotted ASOs.

Hybridization of samples and allele-specific primer extension

In the hybridization step each sub-array on the slide is covered with a droplet of transcribed ssRNA sample and incubated in a humid chamber at 42°C for 20 min. To prevent contamination of adjacent sub-arrays, sample wells are formed using special tape grids or a histological wax pen (Pap Pen, Daido Sangyo Co., Ltd, Japan). In the hybridization the complementary ssRNA molecules anneal to ASOs on the microarray surface. After incubation the slide is washed in buffer containing 0.5 × TE, 0.3 M NaCl and 0.1% Triton X–100, rinsed in distilled water and dried by pressurized air.

For the allele-specific primer extension each sub-array is covered with 2.0 µl of primer extension cocktail containing 2 U Moloney Murine Leukemia Virus reverse transcriptase (MMLV-RT), 10 mM of DTT, 1 µM of

Cy5-labeled dCTP and dUTP nucleotides, 1.0 μM of dATP and dGTP, 0.46 M trehalose and 8% glycerol in 1 × MMLV-RT reaction buffer. The slide is subsequently incubated at 52°C for 20 min. High incubation temperature enhances allelic discrimination in the allele-specific primer extension reaction. Trehalose and glycerol are used to stabilize the polymerase in the extension reaction performed above the optimal temperature for MMLV-RT.

During the primer extension reaction MMLV-RT polymerizes the ASOs having complete hybridization with the ssRNA, including the crucial 3′ nucleotide with deoxynucleotides in the cocktail, simultaneously introducing fluorescent nucleotides. If the ssRNA sample has a mismatch with the 3′ nucleotide of the ASO, the primer extension is suppressed. However, often the allelic discrimination of the primer extension reaction is not complete and some residual extension can take place, which needs to be compensated for by the data analysis. In a sample that is homozygous for a given SNP only one of the two ASOs is fluorescently labeled, whilst in a heterozygous sample both ASOs are labeled. After primer extension, the microarray slide is washed in buffer containing 0.5 × TE, 0.3 M NaCl and 0.1% Triton X–100, rinsed in distilled water and dried with pressurized air.

Image collection of the microarray slide and signal quantification

A digital image of the microarray slide is obtained by a microarray scanner with CCD detector (Scan Array 4000 laser scanner, GSI Lumonics/Packard Bioscience). The scanner measures the emitted fluorescence of the excited ASO spots on the microarray surface and produces a corresponding digital image. Usually a 16-bit TIFF image is used to store a high dynamic range of values per pixel.

The signal intensities of the microarray spots are quantified by image analysis software. The basis for signal quantification is to identify the spot location in the image, define its borders and morphology and quantify the signal and background intensities as well as other parameters. The simplest form of the quantitative spot analysis consists of defining the center of the spot and measurement of the signal within a given radius. This approach is hampered by the fact that contact-printed spots seldom are perfect circles and there might be differences in size and morphology between different ASO spots. The reliability of the allelic discrimination can be increased by utilizing an internal hybridization control oligonucleotide printed within each ASO spot. This can be accomplished by printing an equal amount of a control oligonucleotide to each ASO spot and respectively adding 5′-phosphorylated Cy3-labeled oligonucleotide, complementary to the control oligonucleotide on the array, to the primer extension cocktail. The emitted signal from the internal control is acquired using a different wavelength to the ASOs and is used to normalize the ASO signal.

8.3 Data analysis – allele calling and genotype assignment

The raw signals from the image quantification process are used to derive the allele calls and finally to assign the genotypes. The quantification data is similar to the numerical data typically collected from gene expression

arrays, containing noise from different sources, like the hybridization specificity, ASO printing anomalies and chemical residues, affecting the image. The results from the data analysis are depicted in *Figure 8.3*.

The data analysis starts with data normalization, where we have used standard log transformation of background-subtracted signal intensities. Next we calculate the mean of the summed intensities from signals obtained for both ASOs for a given marker in all samples and exclude outlier ASOs differing from the mean more than certain times the standard deviation, for example more than 2 S.D.. This procedure is able to filter out non-amplified samples as well as extremities of the signal intensities, usually due to non-specific fluorescence signals.

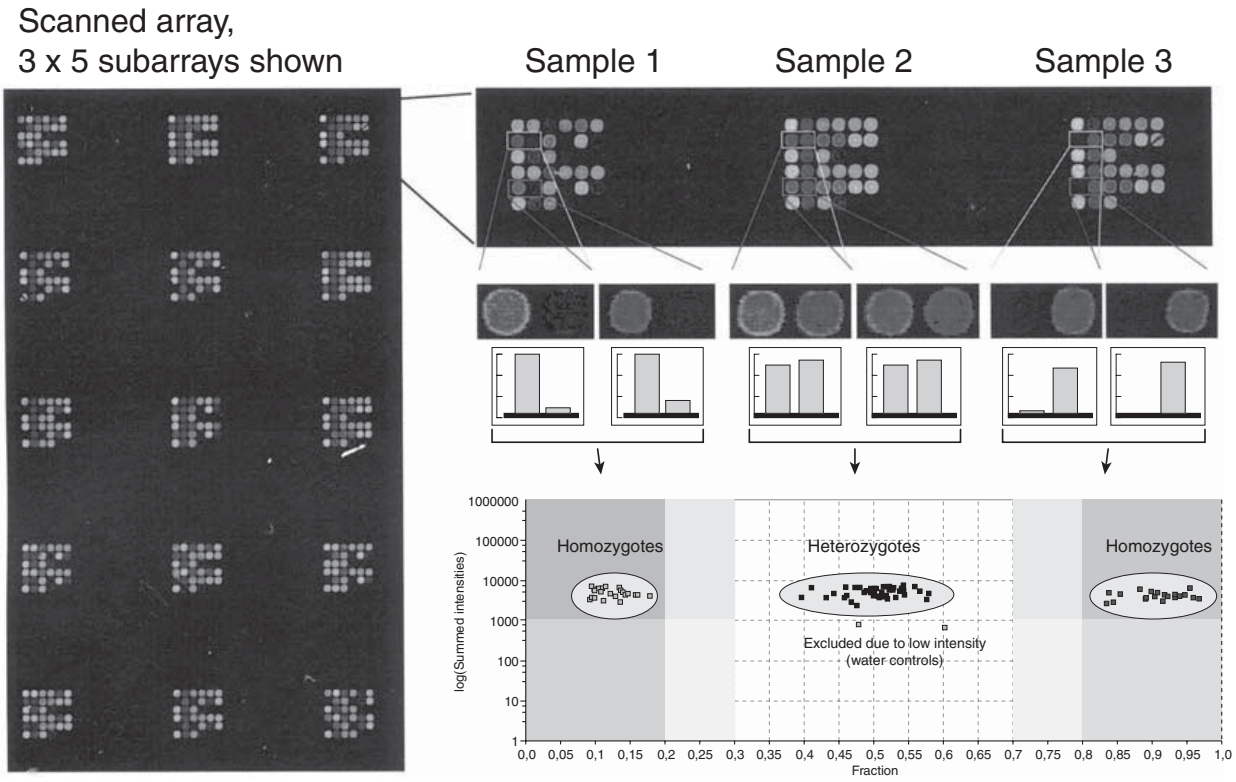


Figure 8.3.

Analysis of the image data, allele calling and genotype assignment. The data shown is from a multiplex genotyping assay of 20 SNPs. Clockwise from the left: The scanned image of a fraction of the genotyping array, indicating the arrays-of-array layout. Each subarray represents an independent sample and each SNP locus is represented by the two spots which contain the ASO1 or ASO2 oligonucleotide. In the three enlarged subarrays, duplicate spots per SNP allele are used, increasing the reliability of genotyping. The spots are quantitated and by using the intensity fractions, the spot pairs are clustered into three distinct genotypes (both homozygotes and heterozygotes). The clustering is confirmed by checking, for example that duplicate spots are within the same cluster. In the graph, the x-axis is the clustered fraction of the spot pairs' background-subtracted intensities and the y-axis is the logarithm of the summed intensities of the spot pairs. Water controls are routinely used and distinguished by low summed intensity values. Possible outlier spot pairs, with too high or low summed intensity values or intensity fraction values between the clusters are excluded and not genotyped.

Optimally the validated data should get organized to three distinct classes, representing the two homozygotes and the heterozygote samples. We typically use clustering methods, such as a modified version of the *k*-means clustering from the one-dimensional signal intensity fraction data. We set $k = 3$ and pre-assign the cluster centroids to 0.2, 0.5 and 0.8 fraction values. We also optimize the clustering so that replicates of the same sample are to be assigned to the same cluster, if possible. Replicate samples having discrepancies in their cluster assignments will not be assigned a genotype, unless the researcher decides to manually exclude the conflicting samples. Usually the clusters converge easily even in the situation where the fraction values of the cluster centroids are heavily skewed to either end of the fraction scale. This makes the clustering approach superior to static assignments of genotypes based just on the intensity fraction values. The clustering can be further directed by using reference samples, for which the genotypes are already known, as well as no template control, for reduction of the error due to the unspecific fluorescence emission.

In order to decrease false genotyping assignments, we next calculate distances between the cluster centroids as well as standard deviation of the samples from the cluster centroids. We use this information to set uncertainty areas between the cluster centroids and all samples in these regions will be excluded from the allele calling, because of the reduced probability of correct cluster assignment and thus increased possibility for a genotyping error.

As the next quality control step we calculate the standard Hardy-Weinberg distributions and use a Chi-Square test in order to evaluate the likelihood for the observed genotyping assignments. Finally we enter all genotyping data to a database, where we check the Mendelian inheritance rules of the samples, if this information is available.

All data analysis steps described here are implemented in SNPSnapper (http://www.bioinfo.helsinki.fi/snpsnapper/), a software specially designed for both allele—specific primer extension and minisequencing in our laboratory (Saharinen *et al.*, manuscript in preparation). SNPSnapper also displays all the data in various dynamic graphs and allows manual intervention in each step and provides the original scanned array image, for example for rejection of conflicting sample replicates. Finally the data is stored in a relational database and can be exported, for example in linkage files to downstream analysis programs.

8.4 Summary

The microarray format has been proven to be a successful tool for multiplexed SNP genotyping, providing medium to high throughput. With a basic level of laboratory automation it is feasible to produce around 960 genotypes for 96 samples in 8 h, depending on the multiplexing level of the assay. In comparison to other currently used genotyping technologies, allele-specific primer extension on microarrays is typically quickly adaptable for novel SNP markers and has very low limitations on the context of the genotyped locus. Whilst the genotyping throughput is not as high as in most robust technologies, the method is highly applicable for smaller-scale projects, involving for example rapid custom candidate gene

genotyping. The special advantage of the method is that it does not require expensive investments on instrumentation.

Since multiple fluorescent nucleotides are added in the extension reaction, the amount of emitted fluorescence is higher than in minisequencing where only a single fluorescent nucleotide is added to the detection. In the minisequencing reaction, the detection primer bound to the array is only extended by a single nucleotide using fluorescently labeled dideoxy nucleotides. By using four different fluorophores for the four ddNTPs, all different alleles can be detected.

The interpretation of the allele signal is often easier with the allele-specific primer extension chemistry. When compared to allele-specific primer extension, minisequencing chemistry however doubles the amount of information that can be retrieved from the same number of probes on the array. With rare SNPs having more than two alleles, this difference is even greater.

Together with the current high-quality microarray technologies and intelligent allele-calling and genotyping software, the reliability of the produced genotypes is high, which is of utmost importance for the downstream analysis of the genotype information.

References

1. Sachidanandam R, Weissman D, Schmidt SC, *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
2. Venter JC, Adams MD, Myers EW, *et al.* (2001) The sequence of the human genome. *Science* **291**: 1304–1351.
3. Marth G, Schuler G, Yeh R, *et al.* (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci USA* **100**: 376–381.
4. Wheeler DL, Church DM, Edgar R, *et al.* (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* **32**: D35–D40.
5. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H and Brookes AJ (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res* **32**: D516–D519.
6. Thorisson GA and Stein LD (2003) The SNP Consortium website: past, present and future. *Nucleic Acids Res* **31**: 124–127.
7. Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Shi MM, Cantor CR and Braun A (2004) Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Res* **14**: 1664–1668.
8. Jiang R, Duan J, Windemuth A, Stephens JC, Judson R and Xu C (2003) Genome-wide evaluation of the public SNP databases. *Pharmacogenomics* **4**: 779–789.
9. Chanock S (2001) Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis Markers* **17**: 89–98.
10. Daly AK (2003) Candidate gene case-control studies. *Pharmacogenomics* **4**: 127–139.
11. Stephens JC (1999) Single-nucleotide polymorphisms, haplotypes, and their relevance to pharmacogenetics. *Mol Diagnost* **4**: 309–317.
12. McCarthy JJ and Hilfiker R (2000) The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat Biotechnol* **18**: 505–508.

13. Ring HZ and Kroetz DL (2002) Candidate gene approach for pharmacogenetic studies. *Pharmacogenomics* **3**: 47–56.
14. Guzey C and Spigset O (2002) Genotyping of drug targets: a method to predict adverse drug reactions? *Drug Safety* **25**: 553–560.
15. Roses AD (2002) Genome-based pharmacogenetics and the pharmaceutical industry. *Nat Rev Drug Discov* **1**: 541–549.
16. McLeod HL and Yu J (2003) Cancer pharmacogenomics: SNPs, chips, and the individual patient. *Cancer Invest* **21**: 630–640.
17. Twyman RM and Primrose SB (2003) Techniques patents for SNP genotyping. *Pharmacogenomics* **4**: 67–79.
18. Inagaki S, Yamamoto Y, Doi Y, *et al.* (2004) A new 39-plex analysis method for SNPs including 15 blood group loci. *Forens Sci Int* **144**: 45–57.
19. Frudakis T, Venkateswarlu K, Thomas MJ, *et al.* (2003) A classifier for the SNP-based inference of ancestry. *J Forens Sci* **48**: 771–782.
20. Reich DE, Cargill M, Bolk S, *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.
21. Weiss KM and Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? *Nat Genet* **26**: 151–157.
22. Liu T, Johnson JA, Casella G and Wu R (2004) Sequencing complex diseases with HapMap. *Genetics* **168**: 503–511.
23. International HapMap Project (2003) The International HapMap Project. *Nature* **426**: 789–796.
24. Cardon LR and Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet* **19**: 135–140.
25. Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2**: 930–942.
26. Chen X and Sullivan PF (2003) Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *Pharmacogenomics J* **3**: 77–96.
27. Kwok PY and Chen X (2003) Detection of single nucleotide polymorphisms. *Curr Issues Mol Biol* **5**: 43–60.
28. Fan JB, Chen X, Halushka MK, *et al.* (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res* **10**: 853–860.
29. Hirschhorn JN, Sklar P, Lindblad-Toh K, *et al.* (2000) SBE-TAGS: an array-based method for efficient single-nucleotide polymorphism genotyping. *Proc Natl Acad Sci USA* **97**: 12164–12169.
30. Pastinen T, Raitio M, Lindroos K, Tainola P, Peltonen L and Syvanen AC (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res* **10**: 1031–1042.
31. Lindroos K, Liljedahl U, Raitio M and Syvanen AC (2001) Minisequencing on oligonucleotide microarrays: comparison of immobilisation chemistries. *Nucleic Acids Res* **29**: E69–E79.
32. Pastinen T, Kurg A, Metspalu A, Peltonen L and Syvanen AC (1997) Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res* **7**: 606–614.
33. Kurg A, Tonisson N, Georgiou I, Shumaker J, Tollett J and Metspalu A (2000) Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet Test* **4**: 1–7.
34. Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S and Syvanen AC (2003) Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic Acids Res* **31**: e129.
35. Bell PA, Chaturvedi S, Gelfand CA, *et al.* (2002) SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Biotechniques* **Suppl**: 70–72, 74, 76–77.

36. Guo Z, Guilfoyle RA, Thiel AJ, Wang R and Smith LM (1994) Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res* **22**: 5456–5465.
37. Rozen S and Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365–386.

Profiling the *Arabidopsis* transcriptome

9

Lars Hennig

9.1 Introduction

During recent years, *Arabidopsis thaliana* (Thale cress) has become the most important model species for plant physiology and genetics. In 2000, the *Arabidopsis* genome was the first plant genome to be sequenced making it the third eukaryote genome to be completed (1). Therefore, it was not surprising that *Arabidopsis* became the main model plant for plant functional genomics as well. Several microarrays were developed to probe the *Arabidopsis* transcriptome, and the Affymetrix AG and ATH1 GeneChip® arrays are currently the most widely used.

Development of the first *Arabidopsis* microarrays was driven by community needs. In the US, NSF supported the development of a spotted microarray with around 11 000 cDNAs (2). In parallel, the Novartis Agriculture Discovery Institute, Inc. (NADII) and Affymetrix together developed the first *Arabidopsis* GeneChip® array. This AG GeneChip® array contains around 8300 probe sets (3) and in 2000 Affymetrix made this microarray publicly available. Again supported by the NSF, the Institute for Genomic Research (TIGR) and Affymetrix developed a second *Arabidopsis* GeneChip® array. Because this ATH1 GeneChip® array contains more than 22 000 probe sets, that is it probes nearly every *Arabidopsis* gene, it is commonly referred to as a 'full genome microarray' (4). Importantly, a systematic comparison of AG and ATH1 microarrays showed that results were consistent between both microarray generations (5).

In the meantime additional *Arabidopsis* microarrays were developed by both academic consortia and commercial service providers. The EU-supported CAGE consortium constructed the CATMA microarray, which contains nearly 20 000 cDNAs (http://www.catma.org/). The Agilent *Arabidopsis* 3 Oligo Microarray Kit uses 60-nucleotide oligomers to probe 40 000 *Arabidopsis* transcripts (including non-coding transcripts). Qiagen Operon offers a set of nearly 30 000 70-nucleotide oligomers that has been used to study flower development in *Arabidopsis* (6). In addition to *Arabidopsis*, microarrays for other plant species are entering the market as well. They include barley, grape and soybean Affymetrix GeneChip® arrays, oligonucleotide sets for grape, *Medicago* and peach from Operon and rice oligonucleotide microarrays from Agilent.

Here, I will describe the typical work-flow of an RNA profiling experiment in *Arabidopsis* using Affymetrix GeneChip® arrays.

9.2 MIAME/Plant – documentation of the experiment

MIAME (Minimum Information About a Microarray Experiment, see Chapter 22) is a standard that aims at providing a conceptual structure for the core information to be captured from most microarray experiments (7). The MIAME standard is very useful for the annotation of labeling and hybridization procedures, measurement data and array design. MIAME/Plant aims to extend the MIAME standard and to establish a list of controlled vocabularies for plant microarray experiments. MIAME/Plant is an extended list of plant experimental description terms from:

- experimental design,
- growth protocol,
- extraction protocol,
- genotype,
- starting material,
- developmental stage,
- plant organs.

For details on MIAME/Plant see the white paper (8).

9.3 RNA extraction

Different RNA extraction protocols work, but we usually use TRIZOL-based extraction followed by a clean-up on RNeasy microspin columns. If possible, we treat all aqueous solutions with 0.1% DEPC overnight before autoclaving.

9.4 Labeling

Labeling involves the synthesis of double-stranded cDNA from total RNA followed by *in vitro* transcription (IVT). The quality of the T7-$(T)_{24}$ primer is critical for the success of the whole experiment. It is essential that the primer is PAGE or HPLC purified, and we recommend verifying the quality of this primer before embarking on further experiments. The quality of the primer can be controlled for example by cDNA synthesis followed by IVT. During the IVT reaction, biotin-labeled cRNA transcripts are produced by a T7 RNA polymerase-catalyzed reaction in the presence of biotin-labeled CTP and UTP nucleotides. Use Qiagen RNeasy columns for purification of IVT samples. Do not use phenol/chloroform extraction to purify biotinylated samples. Compare also the Affymetrix white papers (9).

9.5 Hybridization

The Eukaryotic Hybridization control mix contains non-eukaryotic transcripts, which serve as controls for hybridization quality and array performance. A synthetic control oligonucleotide (B2) provides alignment signals used by the scanner software to position the grid over the array image. Addition of the Eukaryotic Hybridization control mix to the hybridization cocktail is not compulsory, but the B2 control oligonucleotide must be added to every cRNA sample to be hybridized. In this

protocol we use 15 μg of cRNA for the hybridization of standard GeneChip® arrays. Use only RNAse-free plasticware (Eppendorf tubes, pipette tips) and DEPC-treated water. All the buffers used in this protocol must be sterile-filtered (0.22 μm filter).

9.6 Washing, staining and scanning

After hybridization, the probe array is subjected to a series of washes in the fluidics station. Stringent and non-stringent washes are specifically optimized for each probe array type. The hybridized and washed probe arrays are next stained with strepavidin-phycoerythrin conjugate. Please check the fluidics protocol(s) required for the array type you are using on the information sheet provided for each Affymetrix probe array type.

The Affymetrix GeneChip® fluidics station is used for array washing, staining and signal amplification. Place water, washing buffers, SAPE solution and antibody solution in the fluidics station. Refer to the user manual for handling the fluidics station.

After completion of the wash program, check the probe array window for air bubbles. To remove air bubbles, insert a clean pipette tip into the upper septum of the array. Keep the array in a vertical position and pipette 200 μl non-stringent wash buffer into the array using the lower septum of the array. Pipette another 150 μl buffer into the array (extra buffer will come out through the pipette tip attached to the upper septum). Keep the probe arrays without air bubbles at 4°C in the dark until scanning. Refer to the instructions of the scanner and the operating software for scanning. Each complete array image is stored in a separate raw data file (*.dat). The GeneChip® operating software analyzes the image files and derives a single intensity value for each probe cell of an array. These values are contained in the cell intensity (*.cel) file.

9.7 Data pre-processing and data analysis

For preprocessing of Affymetrix GeneChip® microarray data various algorithms exist, for example MAS, RMA and GCRMA (10–12). For a detailed discussion of normalization algorithms see Chapter 17. See Chapter 18 for approaches to detect differentially expressed genes, Chapter 19 for clustering algorithms and Chapter 20 for approaches to detect patterns in time course series. Several online tools are available to analyze microarray data from plants (*Table 9.1*). Data from completed microarray experiments should be submitted to public data repositories (e.g. ArrayExpress, GEO) but also to plant specific microarray databases (e.g. Genevestigator, NASCArrays, TAIR).

9.8 Useful tips

These are only suggestions, but they can make the procedures easier.

1. Have at least 30 μg total RNA for each sample before you start; even if labeling of one sample causes problems, you can still repeat the labeling using the remaining RNA.

Table 9.1. Public tools for analyzing *Arabidopsis* microarray data

Name	Application	Link	Ref.
TAIR GO	GO-classification	http://www.arabidopsis.org/tools/bulk/go/index.jsp	(16)
TAIR Aracyc	Display of expression data on a metabolic map	http://www.arabidopsis.org:1555/expression.html	(17)
TAIR promoter analysis	Identification of enriched promoter motifs	http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp	(16)
TAIR Chromosome Map Tool	Mapping genes on chromosomes	http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp	(16)
Genevestigator	Analysis of expression patterns during development and stress	https://www.genevestigator.ethz.ch	(14)
MapMan	Data visualization	http://gabi.rzpd.de/projects/MapMan	(18)

2. For difficult tissue like seeds use a borate buffer method (13).
3. Use a double-labeling kit (e.g. from Affymetrix or Ambion) if you have limited amounts of RNA to start with (works with as little as 50 ng total RNA).
4. For better recovery, pass RNA-containing solutions twice over RNeasy columns and elute twice (first with 30 µl, then with 20 µl water).
5. Store wash buffers at 4°C, but leave at room temperature over night before use to avoid air bubbles in the fluidics station.

9.9 Summary

Although several competing microarray platforms are available for transcriptional profiling of *Arabidopsis thaliana*, Affymetrix GeneChip® ATH1 microarrays are certainly among the most powerful. This is in part due to the widespread availability of Affymetrix systems in laboratories and service centers. Because ATH1 microarrays are commercially available and the technology is very robust, researchers do not need to spend time on technology development but can focus on their primary goal – research. Moreover, the use of standardized protocols and microarrays has enabled novel tools for meta-analysis of independent experiments from various groups. One such tool from our own lab, Genevestigator (14), has proven to be extremely popular in the field of plant science. Nonetheless, there are disadvantages of the ATH1 microarrays. First, all disadvantages common to any oligonucleotide array apply to ATH1 arrays. Second, the *Arabidopsis* genome contains more than 29 700 annotated genes, but only 22 000 genes are probed by the ATH1 array. Finally, many non-coding RNAs are generated from the *Arabidopsis* genomes (15). These RNAs often have important regulatory roles but are usually not probed by the ATH1 array. However,

full-genome tiling arrays will likely become available soon (15) eliminating many of the major current limitations.

Acknowledgement

I would like to thank Nicole Schönrock, ETH Zürich, and John Okyere, the Nottingham Arabidopsis Stock Center, for comments on the manuscript.

References

1. Arabidopsis Genome Initiatives AG (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana Nature* **408**: 796–815.
2. Wisman E and Ohlrogge J (2000) Arabidopsis microarray service facilities. *Plant Physiol* **124**: 1468–1471.
3. Zhu T and Wang X (2000) Large-scale profiling of the *Arabidopsis* transcriptome. *Plant Physiol* **124**: 1472–1476.
4. Redman JC, Haas BJ, Tanimoto G and Town CD (2004) Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. *Plant J* **38**: 545–561.
5. Hennig L, Menges M, Murray JAH and Gruissem W (2003) *Arabidopsis* transcript profiling on Affymetrix genechip arrays. *Plant Mol Biol* **54**: 457–465.
6. Wellmer F, Riechmann JL, Alves-Ferreira M and Meyerowitz EM (2004) Genome-wide analysis of spatial gene expression in *Arabidopsis* flowers. *Plant Cell* **16**: 1314–1326.
7. Brazma A, Hingamp P, Quackenbush J, *et al*. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat Genet* **29**: 365–371.
8. MIAME Group (2004) Minimum Information About a Microarray Experiment – MIAME for plant genomics (MIAME/Plant). http://arabidopsis.info/info/miame.html.
9. Allymetrix (2004) GeneChip® expression analysis technical manual. http://www.affymetrix.com/support/technical/manual/expression_manual.affx.
10. Liu WM, Mei R, Ryder TB, *et al*. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**: 1593–1599.
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264.
12. Wu Z, Irizarry RA, Gentleman R, Murillo FM and Spencer F (2003) *A Model Based Background Adjustment for Oligonucleotide Expression Arrays Technical Report* John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD.
13. Wan CY and Wilkins TA (1994) A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum L.*). *Anal Biochem* **223**: 7–12.
14. Zimmermann P, Hirsch-Hoffmann M, Hennig L and Gruissem W (2004) Genevestigator. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol* **136**: 2621–2632.
15. Yamada K, Lim J, Dale JM, *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**: 842–846.
16. Rhee SY, Beavis W, Berardini TZ, *et al*. (2003) The *Arabidopsis* information resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res* **31**: 224–228.

17. Mueller LA, Zhang P and Rhee SY (2003) Aracyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol* **132**: 453–460.
18. Thimm O, Bläsing O, Gibon Y, *et al.* (2004) Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* **37**: 914–939.

Protocols

CONTENTS

Protocol 9.1: RNA extraction

PRECAUTIONS

1. Wear gloves and use RNase-free tubes and pipette tips.
2. Water used in the protocol is molecular biology grade (nuclease-free) water.

MATERIALS

Reagents and kits

- Trizol (Invitrogen)
- RNeasy RNA purification kit (Qiagen)

METHODS

1. Grind 100 mg of tissue in liquid nitrogen to a fine powder and transfer into an Eppendorf tube.
2. Add 1 ml Trizol and continue according to manufacturer's protocol.
3. Purify the RNA on RNeasy microspin columns according to manufacturer's protocol.
4. Quality control: dilute sample 1:100 in 10 mM TRIS (pH 7.5). Measure absorbance at 260 nm and 280 nm. The ratio abs_{260nm}:abs_{280nm} should be between 1.9 and 2.1. Run 1 μg of RNA on a 1% agarose gel. Nuclear and plastid ribosomal RNA should be visible as distinct bands. Alternatively, the Agilent Bioanalyzer 2100 (lab-on-a-chip) microcapillary system can be used to assess RNA integrity. For best results use only non-degraded RNA of high purity.

Protocol 9.2: Labeling

MATERIALS

Reagents and kits

- 5 × fragmentation buffer (200 mM Tris-acetate, pH 8.1, 500 mM KOAc, 150 mM MgOAc)
- T7-oligodT (GGC CAG TGA ATT GTA ATA CGA CTC ACT ATA GGG AGG CGG-$(dT)_{24}$)
- SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen)
- Phase Lock Gels (Eppendorf)
- BioArray High Yield IVT kit (ENZO)
- Alternatively: MEGAscript T7 *in vitro* transcription kit (Ambion) or GeneChip® IVT Labeling Kit (Affymetrix).

METHODS

cDNA synthesis

1. Use 15 μg total RNA for first and second strand cDNA synthesis according to manufacturer's protocol.
2. Clean up the cDNA using Phase Lock Gel tubes according to manufacturer's protocol.
3. Generate labeled cDNA by IVT using the ENZO BioArray kit according to manufacturer's protocol.
4. Clean up the cRNA with Qiagen RNeasy columns according to manufacturer's protocol.
5. Quantify the cRNA: measure Abs_{260nm} of cRNA (typically diluted 1/50 to 1/100) to determine the yield. When using total RNA as the starting material, it is necessary to calculate *an adjusted cRNA yield* to correct the carryover of unlabeled total RNA. Using an estimate of 100% carryover, use the following formula to determine the adjusted cRNA yield:

   adjusted cRNA yield = RNA_m – (total RNA_i)

   RNA_m = amount of cRNA measured after IVT (μg)

   RNA_i = starting amount of total RNA (μg)

Note: use the adjusted cRNA yield when calculating the amount of cRNA needed for fragmentation and array hybridization.

6. Check 1 μg of the purified transcripts on a 1% agarose gel. Transcript lengths should range from 0.5 to 2 kb.

7. Fragmentation of the cRNA. Mix the following:

   16 μg cRNA (Note: use adjusted cRNA yield.)

   8 μl 5 × fragmentation buffer

   RNAse-free H_2O to 40 μl

   Incubate at 95°C for 35 min. Store at –20°C. Check 2.5 μl (=1 μg) on a 1% agarose gel. The size of the cRNA should be reduced to 100 bp.

Protocol 9.3: Hybridization

PRECAUTIONS

1. Avoid fingerprints on the array cartridge window as these may interfere with scanning. If fingerprints are present, these should be cleaned with soft paper and ethanol.
2. Use only powder-free gloves to minimize introduction of powder particles into the sample, buffers, or array cartridges.

MATERIALS

Reagents

- Eukaryotic Hybridization control mix (Affymetrix)
- B2 control oligonucleotide (Affymetrix)
- Herring sperm DNA
- Acetylated BSA
- 12 × MES stock, pH 6.5–6.7 (1 l, do not autoclave)

  70.4 g MES free acid monohydrate

  193.3 g MES sodium salt
- 2 × Hybridization buffer (50 ml)

  8.3 ml of 12 × MES

  17.7 ml of 5 M NaCl

  4.0 ml of 0.5 M EDTA

  0.1 ml of 10% Tween-20

  Mix and adjust volume to 50 ml. Filter through a 0.2-μm filter.

METHODS

1. Equilibrate the probe array to room temperature immediately before use (probe arrays should be stored at 4°C).
2. Pre-hybridize the probe array with 1 × hybridization buffer. Keep the array upside down. Insert a clean small pipette tip into the top septum of the array to allow venting of air from the chamber inside the probe array. Pipette 200 μl of 1 × hybridization buffer into the array through the bottom septum. Incubate for at least 10 min at 45°C with 60 r.p.m. rotation (GeneChip® hybridization oven).

3. Preparing the hybridization target: mix the following components in a RNAse-free 1.5-ml Eppendorf tube:

15 μg of fragmented cRNA	37.5 μl
20 × Eukaryotic Hybridization control mix	15 μl
3 nM B2 control oligonucleotide	5 μl
Herring sperm DNA (10 mg/ml)	3 μl
Acetylated BSA (50 mg/ml)	3 μl
2 × Hybridization buffer	150 μl
RNAse-free H_2O to final volume of 300 μl	86.5 μl

4. Denature the hybridization cocktail at 99°C for 5 min.
5. Incubate the hybridization cocktail at 45°C for 5 min.
6. Spin the samples at maximum speed in an Eppendorf centrifuge for 5 min to remove any insoluble material from the hybridization cocktail.
7. Remove the pre-treatment solution from the pre-hybridized probe arrays and add 200–230 μl of the hybridization cocktail into the probe array. A small air bubble inside the probe array is needed for proper mixing of the hybridization cocktail during the hybridization. Seal the septa with small 8-mm paper stickers to prevent any loss of hybridization cocktail during the incubation.
8. Hybridize for 16 h at 45°C with 60 r.p.m. rotation in the hybridization oven.

Protocol 9.4: Washing, staining and scanning

PRECAUTIONS

1. SAPE is light-sensitive and must be stored at +4°C in the dark.
2. Never freeze SAPE solution.
3. Keep SAPE-staining cocktail in colored Eppendorf tubes.
4. Always prepare staining cocktail freshly before use.

MATERIALS

Reagents

- 20 × SSPE (pH 7.4) (3 M NaCl, 0.2 M NaH_2PO_4, 20 mM EDTA)
- Stringent wash buffer (800 ml):
  - 66.6 ml of 12 × MES
  - 4.2 ml of 5 M NaCl
  - 0.8 ml of 10% Tween-20
- Non-stringent wash buffer (800 ml):
  - 240 ml of 20 × SSPE
  - 0.8 ml of 10% Tween-20
- 2 × Stain buffer (250 ml):
  - 41.7 ml of 12 × MES
  - 92.5 ml of 5 M NaCl
  - 2.5 ml of 10% Tween-20
- Acetylated BSA
- Herring sperm DNA
- R-phycoerythrin streptavidin (Molecular Probes)
- Goat IgG (10 mg/ml in PBS, pH 7.2. Store at 4°C)
- Biotinylated anti-streptavidin (0.5 mg/ml in DEPC-water. Store at 4°C.)

METHODS

1. Remove the hybridization cocktail from the array into a new RNAse-free tube and store it at –20°C. The same hybridization cocktail can be used again (denature the cocktail at 99°C for 5 min before each use).
2. Fill the probe array manually with 250 μl of non-stringent wash buffer. The probe array can be stored up to 3 h at 4°C in the dark before proceeding with the washing and staining.
3. Turn on the power for the hardware (note the order recommended by the manufacturer). Open scanner-operating software (e.g. GCOS).
4. Prepare staining cocktails.

SAPE solution (1200 μl per array)

600 μl 2 × stain buffer

540 μl RNAse-free water

48 μl acetylated BSA (50 mg/ml)

12 μl SAPE (1mg/ml)

Mix well and divide into two aliquots of 600 μl each.

Antibody solution (600 μl per array)

300 μl 2 × stain buffer

266.4 μl RNAse-free water

24 μl acetylated BSA (50 mg/ml)

6 μl goat IgG (10 mg/ml)

3.6 μl biotinylated antibody (0.5 mg/ml)

Mix well.

Affymetrix GeneChip analyses – the impact of RNA quality

10

Ludger Klein-Hitpass and Tarik Möröy

10.1 Introduction

According to the steeply increasing number of reports in the literature Affymetrix DNA oligonucleotide arrays (GeneChips) have gained considerable acceptance in the research community. The latest version of human GeneChips, U133A Plus 2.0, representing approximately 38 500 transcripts on a single chip allows genome-wide expression profiling in a very convenient setting. In contrast to cDNA and oligonucleotide arrays from different manufacturers, Affymetrix GeneChips measure transcripts by a set of multiple probes (called a probe set), which usually consists of 11 probe pairs. Each probe pair contains two 25mer oligonucleotide probes, a perfect match (PM) oligonucleotide that represents part of the cDNA sequence of interest and a mismatch (MM) oligonucleotide, which is identical to the PM oligo except for a mismatch mutation at the central position. In the Affymetrix image analysis method implemented in the MAS5.0 and GCOS software, the PM-MM signal differences of the 11 probe pairs of a probe set are converted into a single probe set signal value, which is a measure of the abundance of the transcript. In addition, for each probe set the software estimates the reliability of the measurement resulting in a detection call (present, absent, or marginal) based on a significance analysis of the PM-MM differences.

To minimize discrepancies due to varying sample preparations, hybridization conditions, staining intensities or probe array lots, the software provides several normalization options, which can be applied to the datasets from different arrays. The recommended procedure for datasets with relatively little expression differences is called 'global scaling'. During global scaling, the software examines all probes on the array to compute a trimmed mean signal. Then, a scaling factor is calculated and applied to each signal on the array to standardize the trimmed mean of the array to a user-specified target signal. Another option, called 'selected probe sets scaling', computes the trimmed mean signal of selected probe sets to derive a scaling factor that is again applied to all probes and adjusts the trimmed mean signal of the selected probe sets to the target signal value specified by the user for all arrays of a given dataset. Selected probe set scaling is more appropriate and provides more

accurate signal measurements, if differences between samples are relatively high. However, it requires a set of transcripts known to be equally abundant in all samples, such as a fixed amount of external controls spiked into a constant amount of starting material.

Most probe sets cover a target sequence of about 300 bases in length located within the region of 600 bases proximal to the 3′-end of the transcripts, mostly in non-translated regions, where distinction of transcripts encoded by highly homologous gene families is facilitated. To select these probe set sequences, information from multiple public domain databases as well as proprietary information of Affymetrix is used. In cases where database entries suggest the occurrence of alternative splicing or polyadenylation, multiple probe sets covering different regions of a gene can be present on the GeneChip arrays. Information on each probe set, including the sequences interrogated by the probe pairs and the sequences of the oligonucleotides chosen, has been made accessible to the public via an internet platform called NetAffx Analysis Center (www.affymetrix.com).

To be analyzed on GeneChips, mRNA molecules contained in total RNA samples are amplified and labeled by a standardized and widely used protocol (*Figure 10.1A*), which involves conversion of mRNA molecules into double-stranded cDNA using an oligo$(dT)_{21}$ primer with a T7 RNA polymerase promoter tag for first strand synthesis. After second strand cDNA synthesis, subsequent *in vitro* transcription by T7 RNA polymerase yields biotinylated anti-sense copy RNA, termed cRNA target, which is sufficient for the hybridization of several GeneChip arrays. Starting with a perfectly intact RNA sample, the resulting cRNA target should ideally represent full length anti-sense mRNA sequences. However, degradation of the RNA during preparation or storage can lead to truncated or cleaved mRNA molecules that cause premature stops of oligo(dT)-primed reverse transcription. Hence, mRNA sequences located 5′ of cleavage sites are not converted into cDNA and a 3′-biased cRNA target is obtained (*Figure 10.1B*). To be able to estimate whether such a 3′-bias has occurred during cRNA preparation and to monitor the quality of the RNA used to prepare a cRNA target, Affymetrix expression GeneChips provide a number of probe sets, which interrogate 5′, middle, and 3′ parts of two transcripts of the housekeeping genes, *Gapdh* and *β-Actin*, which are ubiquitously expressed at relatively high levels. Other 5′, middle, and 3′ probe sets detect different parts of various externally added intact polyA$^+$-spikes, which can be used as external normalization controls and serve to monitor cDNA- and cRNA-synthesis steps. cRNA targets derived from high quality RNA samples display 3′ to 5′ probe-set signal ratios for *Gapdh* and *β-Actin* close to 1.0, whereas cRNA targets from degraded starting material exhibit increased ratios. With a few exceptions, where increased 3′/5′-ratios of housekeeping genes may truly reflect a regulated cellular process such as apoptosis, increased 3′/5′-ratios result mostly from incomplete inactivation and removal of endogenous ribonucleases during cell homogenization and RNA extraction, contamination with exogenous ribonucleases or spontaneous cleavage, which is frequently observed in purified RNA samples that have been subjected to multiple freeze and thaw cycles.

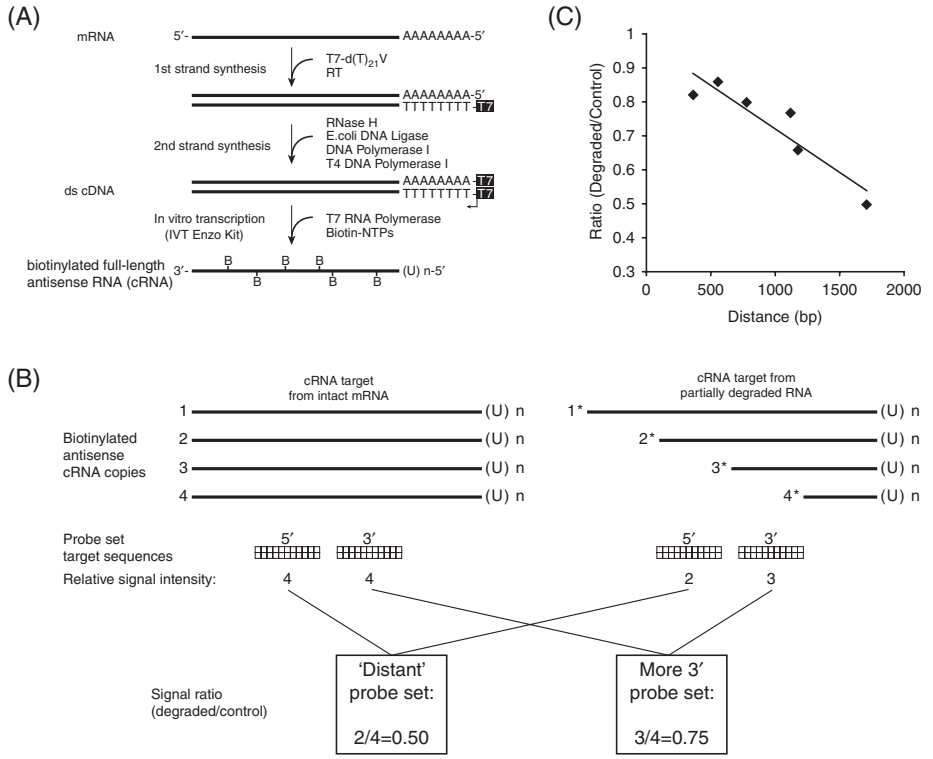


Figure 10.1.

Differential representation of 5′ mRNA sequences in cRNA targets from partially degraded samples. (**A**) Intact mRNAs are converted by the standard labeling method into biotinylated full-length anti-sense cRNA. (**B**) Schematic representation of anti-sense cRNA copies of a specific gene generated from an intact (left panel) and a partially degraded RNA sample (right panel). Depending on the position of the cleavage sites in the mRNA molecules, copies lacking variable parts of the 5′ region are generated (marked 2 to 4). A probe set interrogating more 3′ located sequences would detect copies 1 to 3, whereas a more 5′ probe set could detect only 1 to 2, resulting in a variable degree of under-estimation and increased signal ratios when compared to intact samples. cRNA copy 4 is non-productive with respect to both the 5′ and 3′ probe sets, while 3 is non-productive only with respect to the 5′ probe set. (**C**) Graph showing the relationship of the measured signals in intact and degraded total RNA and the distance between the 5′-end of the probe-set sequences and the polyA end of the transcript. The ratio of signals observed in degraded and control RNA for the AFFX-*Gapdh* and *β-Actin* 5′, middle, and 3′ probe sets is plotted against the distance between the 5′-end of the probe set and the start of the polyA tract. Accession numbers for human *Gapdh* and *β-Actin* full-length sequences are M33197 and X00351, respectively.

Most GeneChip users agree that array data obtained from RNA samples showing 3′/5′-ratios for *β-Actin* greater than 3.0 should be treated with special caution and must not be compared with array data from intact control samples, since the differential representation of 5′ mRNA sequences might introduce a significant error. While most microarray lab units perform mandatory RNA quality checks by analyzing RNA samples on the Agilent BioAnalyzer or by gel electrophoresis to exclude very poor RNA preparations from further processing, a considerable variation of 3′/5′-ratios

is still observed in many projects and experiments. Moreover, because some tissues or cells contain high amounts of ribonucleases or require more time for RNA extraction than others, systematic differences in RNA qualities of samples to be compared can sometimes hardly be avoided.

10.2 Aim and experimental design

Since the GeneChip probe set sequences are strongly biased towards the 3′-end of the published sequences, it is implicated that under-representation of 5′ mRNA sequences occurring in cRNA targets from moderately degraded RNA samples may not affect data quality in most experiments. However, a systematical experimental analysis to what extent this might create false positive targets is lacking. To determine the impact of degradation on microarray data, aliquots of an intact RNA sample prepared from HeLa cells were treated in a controlled fashion by heat in the presence of divalent ions, which results in random cleavage of RNA molecules. After chilling, equal amounts of a set of intact poly(A)$^+$ RNAs were spiked into partially degraded as well as untreated control samples. These spiked-in RNAs, which can be measured on the arrays, served to monitor various steps of the enzymatic conversion into cRNA and, importantly, as external normalization controls using the selected probe set scaling option. All targets were prepared and analyzed by hybridization to Affymetrix HG-U133A arrays containing more than 22 000 probe sets. Since the overall sequence content in degraded and control RNA samples remained identical, genes that were identified by statistical analysis to be significantly up- or down-regulated in the degraded samples represent false positive targets, which are solely due to the introduced 3′-bias in the cRNA target from partially cleaved RNA. In addition, we compared the effect of two different array normalization procedures, global scaling and selected probe set scaling to spiked-in controls, on the rate of false up- and down-regulated transcripts.

10.3 Statistics of RNA and array quality parameters

As indicated in *Table 10.1*, yields of cRNA from control and partially degraded RNA samples were highly similar (p =0.92, paired t test), suggesting that cDNA synthesis and *in vitro* transcription efficiency were not impaired by the pretreatment. For control RNAs, mean 3′/5′-ratios of 0.80±0.03 and 1.24±0.05 for *Gapdh* and *β-Actin* transcripts, respectively, were determined, confirming high RNA integrity. Partially cleaved RNAs had 3′/5′-ratios of 0.99±0.04 (*Gapdh*) and 2.18±0.26 (*β-Actin*), corresponding to a 1.25-fold (*Gapdh*) and 1.75-fold (*β-Actin*) increase (p <0.001). Other array parameters, such as noise and background average, were not significantly different between intact and partially degraded samples. Moreover, despite the clearly reduced RNA quality, there was no change in the overall percentage of probe sets called present by the detection algorithm of MAS5.0, regardless of whether degraded or intact RNA was used for target preparation. In summary, all parameters suggested that the quality of both sets of cRNA samples and GeneChip arrays was very comparable and that the extent of 3′-bias introduced into the cRNAs derived from partially cleaved RNAs was clearly within the range that

Table 10.1. Summary of sample and array characteristics

Variable (applied scaling)	Control RNAs	Degraded RNAs	Ratio (Degraded/ Control)	*t* test *p*-value
No. of arrays	6	6		
cRNA yield	33.7±2.8 µg	33.5±5.1 µg	0.99	0.920
3′/5′-Ratio *Gapdh*	0.80±0.03	0.99±0.04	1.25	<0.001
3′/5′-Ratio *β-Actin*	1.24±0.05	2.18±0.26	1.75	<0.001
Noise	2.95±0.12	2.98±0.29	1.01	0.858
Background	63.3±4.1	63.1±6.5	1.00	0.937
Percent present calls	53.6±1.1	54.2±0.6	1.01	0.283
Mean signal (raw)	355.5±46.7	293.4±25.4	0.83	0.022
Trimmed mean signal (raw)	275.8±38.7	225.8±21.5	0.82	0.025
Cumulative signal/10^6 (raw)	7.9±1.04	6.5±0.57	0.82	0.022
Scale factor (global)	4.27±0.63	5.17±0.48	1.21	0.020
Mean signal (global)	1495.8±23.7	1507.9±24.4	1.01	0.402
Cumulative signal/10^6 (global)	33.3±0.53	33.6±0.54	1.01	0.402
Scale factor (spike)	0.67±0.10	0.65±0.06	0.97	0.937
Mean signal (spike)	234.7±3.7	189.8±5.7	0.81	<0.001
Trimmed mean signal (raw)	181.9±4.0	145.9±3.4	0.80	<0.001
Cumulative signal/10^6 (spike)	5.2±0.1	4.2±0.1	0.81	<0.001

Starting with three different pools of control and degraded RNA samples, cDNAs were generated in duplicate from each pool and used in IVT labeling reactions to generate six cRNA targets in each group, which were hybridized to HG-U133A arrays. Data are mean ±S.D. Array images were analyzed in MAS5.0 in three different ways, to yield unadjusted raw, globally scaled and spike-mask-scaled signals. User-specified target intensities during global and spike-mask scaling were 1000. *t* test *p*-values were determined using the SSPS software.

would permit a reasonable comparison with the control samples according to general recommendations.

10.4 Comparison of signal measures computed by different array normalization procedures in control and degraded samples

Signal calculation for each array image was performed in MAS5.0 in three different ways. First, signals were calculated by omitting any normalization procedure, yielding unadjusted signal intensities (raw signals). Second, images were re-analyzed using the all probe sets scaling option (globally scaled signals). Third, signals were computed using the selected probe sets scaling option with the help of a mask file, which combines all the 27 available probe sets for the spiked-in control RNAs (spike mask scaled signal). As the external spike controls were present in equal amounts in all RNA samples, the latter scaling procedure ensures that the normalization step corrects for any unknown variables but not for systematic parameters linked to the 3′-bias in cRNA targets prepared from partially degraded RNA. Thus, signal intensities obtained by this method should reflect the genuine hybridization signals of the labeled cRNAs more precisely than the raw signals or the signals computed by global scaling.

Statistical analysis of the raw signal data revealed that mean (83%), trimmed mean (82%), and cumulative raw signals (83%) were significantly decreased in degraded RNAs (*Table 10.1*). Scale factors generated by scaling to the spike mask were highly similar in degraded and control samples, demonstrating that the spiked-in controls were indeed equally abundant in all samples. Consistent with the observations already made at the level of raw signals, resulting mean (81%), trimmed mean (80%), and cumulative signals (81%) after the spiked-in control probe-set scaling proved to be highly significantly decreased in degraded samples ($p < 0.001$). A detailed analysis of the size distribution of the mean raw and spike-mask-scaled probe-set signals showed a general downshift to lower intensities in degraded samples (*Figure 10.2A, B*). In contrast, size distributions of globally scaled signals of control and degraded samples did not correctly reflect

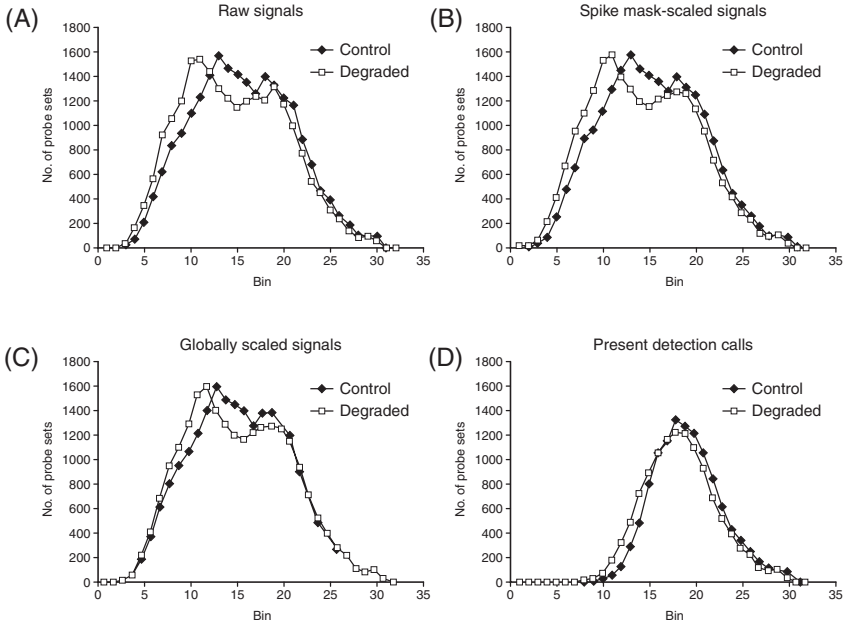


Figure 10.2.

The effects of various array normalization methods on the distribution of resulting signal measures in control and partially cleaved RNA samples. (**A**) Unadjusted mean raw signal measures of control ($n = 6$) and degraded samples ($n = 6$) were logarithmically transformed (base 2) and sorted into bins of equal size (0.5). Increasing bin numbers represent higher signal bins. (**B**) Signal measures resulting from spike-mask scaling are represented as in (**A**). Signals from degraded samples show a general downshift due to the presence of partial, non-productive cRNA copies, which is also evident in the distribution of the raw signals. (**C**) Signal measures resulting from global scaling are shown as in (**A**). Note, that this procedure arithmetically compensates some of the signal drop observed in degraded samples. Thus, signal measures of degraded samples do not accurately reflect hybridization signals any more. (**D**) Spike-mask-scaled signals of 12 118 reliably measured probe sets (≥50% present detection calls in the 12 arrays hybridized) were binned as shown in (**C**). The resulting distributions show that the peaks observed at lower signal intensities in (**A**)–(**C**) represent probe sets which received absent and/or marginal detection calls in the majority of analyses.

this signal drop any more (compare *Figure 10.2B* and *C*), as the difference is largely compensated by correspondingly increased scale factors (p =0.02). Because all arrays were taken from the same batch and very little intra-group variation was observed, raw as well as spike-mask-scaled data indicate a significant drop in signal intensities on arrays hybridized with targets from partially cleaved RNA samples. Since all arrays were hybridized with equal amounts of labeled cRNA, we conclude that due to the partial cleavage mimicking RNA degradation and the 3′-biased amplification procedure, there is a considerably increased fraction of cRNA molecules present in targets prepared from the degraded starting material that does not generate hybridization signals. As indicated in *Figure 10.1B*, such non-productive or less productive molecules are generated during the cDNA and cRNA synthesis steps, when RNA cleavage occurred at a position located within or 3′ of the sequences represented by the corresponding probe set.

10.5 SAM of degraded versus control RNA

Using Significance Analysis of Microarrays (SAM), a statistical procedure specifically designed for analysis of large microarray datasets (1), the number of probe sets displaying significantly different signals in control and degraded samples was first determined in the dataset derived by scaling to the external spike-in control RNAs, which reflects the 'most straight' measurement of the hybridization signals. Prior to statistical analysis, all probe sets, which received less than six (50%) present detection calls in the 12 array analyses, were eliminated from the dataset in order to reduce most of the technical noise present in the lower signal range. Since the detection call algorithm implemented in MAS5.0 is independent of the scaling procedure and because there was no significant difference in the percentage of present detection calls in control and degraded samples (*Table 10.1* and *Figure 10.2D*), this step was not expected to introduce any further bias.

At a median false discovery rate of 0.1%, 6358 (52.5%) of the remaining 12 118 probe sets were called significant by SAM (*Table 10.2*). The majority (96.7%) of the significant probe sets indicated a down-regulation (true under-representation) of the corresponding genes in cRNAs from partially degraded samples, clearly reflecting the signal drop observed on the arrays hybridized with the more 3′-biased targets. At a fold-change cut-off of 2.0, which is commonly used in microarray data mining, as many as 1067 probe sets displayed significantly reduced (down) and only 10 up-regulated (up) signals. Due to the experimental design, true up-regulation of transcripts was not expected to occur in the degraded samples. Thus, the up-regulated probe sets might largely reflect the number of genes occurring by chance at the given level of significance of SAM statistics (false discovery rate 0.1%). One-hundred and thirty-four probe sets, that is 0.6% of all probe sets present on the HG-U133A array, showed more than fourfold reduced signals, while nine probe sets displayed even more than 10-fold reduced hybridization signals in the more 3′-biased samples.

Compared to the dataset obtained from the spike-mask-scaled images, SAM analysis of the signals derived by all probe set scaling revealed a lower number of significant probe sets (3073), whereas a markedly higher fraction of probe sets with apparently increased signals in degraded samples

Table 10.2. Statistical analysis of microarray data

Fold-change cut-off	Scaling to pA+spike mask			Global scaling		
	No. of probe sets	Up	Down	No. of probe sets	Up	Down
No	6358	207	6151	3073	1315	1758
1.5	2926	69	2857	1610	409	1201
2.0	1077	10	1067	560	46	514
3.0	300	2	298	165	4	161
4.0	134	0	134	80	2	78
6.0	38	0	38	22	0	22
8.0	16	0	16	11	0	11
10.0	9	0	9	4	0	4
12.0	4	0	4	0	0	0

Microarray data sets obtained after global scaling or scaling to the polyA+spike mask were analyzed using Significance Analysis of Microarrays (SAM) (1). Signals were log2 transformed prior to analysis. Of the 22 283 probe sets present on the HG-U133A array, only those receiving six or more present detection calls in the Affymetrix single array analysis were selected for SAM (12 118 probe sets). Unpaired analysis was performed with 1000 permutations of the data set. Probe sets called significant at a median false discovery rate (FDR) of 0.1% were selected by adjusting the delta parameter. Data indicate the number of genes passing the indicated fold-change cut-off. Down/up: down- or up-regulated in degraded as compared to control RNA sample.

was evident at all fold-change cut-offs up to fourfold (*Table 10.2*). At the twofold cut-off, 514 probe sets were identified that indicated down-regulation of the corresponding transcripts, while the number of probe sets with up-regulated signals (46) identified by SAM clearly exceeded the number of probe sets expected at the given false discovery rate of 0.1% (12.1). Thus, we conclude that global scaling of arrays derived from degraded and intact RNA samples introduces an increased risk for false up-regulated probe-sets signals in degraded samples and results in a smaller number of probe sets displaying reduced signals in degraded samples.

10.6 Summary

A heat fragmentation procedure introducing random cleavage sites into the total RNA was chosen in this study to generate partially cleaved or degraded RNA samples displaying less than twofold increased 3′/5′-signal ratios for *β-Actin*, in order to achieve a degree of 3′-bias that is within the range frequently observed in RNAs isolated from patient material, but generally judged as being acceptable in Affymetrix GeneChip studies. The results presented here show that those cRNA targets displaying 3′/5′-ratios for *β-Actin* of 2.18±0.26 yield lower hybridization signals reaching only approximately 81% of those obtained with targets derived from intact RNA exhibiting ratios of 1.24±0.05. This difference was associated with the presence of an increased fraction of incomplete and non-productive anti-sense cRNA copies in targets from degraded RNA samples. While this result *per se* is astonishing, such an intensity drop could be well tolerated and be

corrected perfectly by global scaling, if the percentage of non-productive cRNA molecules was identical for each individual transcript measured. However, this is clearly not the case, since SAM analysis of our data showed that the degree of under-estimation in cRNA targets from degraded RNA as measured by individual probe sets is highly variable and reaches ratios greater than 10-fold (*Table 10.2*). The increased fraction of non-productive cRNA molecules present in targets from degraded samples is generated from mRNA templates, where cleavage occurred within or at a position located 3′ of the sequences represented by the corresponding probe set (*Figure 10.1B*). If heat-induced or enzymatic cleavage occurs in a random fashion, then the fraction of non-productive cRNA molecules increases with the distance between the 5′-end of the sequences interrogated by the probe set and the polyA tract of the individual transcript. Indeed, as exemplified by the signals observed on 3′, middle, and 5′ probe sets for *Gapdh* and *β-Actin*, the degree of transcript under-estimation in degraded samples is lowest on the most 5′ probe set and increases with the distance from the polyA site in a linear fashion (*Figure 10.1C*), proving that more distant probe sets are especially vulnerable to signal under-estimation in degraded samples.

Assuming that the linear relationship holds true for all transcripts, the degree of under-estimation observed in the experiment described here can be used to extrapolate the true distance between probe set target sequence and the functional polyA site in the used HeLa cell line. Examination of a number of more than 10-fold under-estimated probe sets in degraded samples indeed revealed examples where the extrapolated distance of more than 3 kb was confirmed by sequence data. In many cases, such distant probe sets have been designed on purpose to allow detection of a shorter transcript variant with a more 5′ polyA site. However, in a number of cases, sequence data in the public databases suggested that these highly under-estimated probe sets were located within the region of 600 bp proximal to the 3′-end of the known sequence. Such discrepancies could point to cases where the cDNA sequences deposited in the public domain databases are incomplete at the 3′-end, resulting in probe set selection that is not as 3′-biased as intended by Affymetrix' probe selection process.

As revealed by the SAM analysis of the dataset obtained by selected probe set scaling of the approximately 12 000 reliably measured transcripts in this study using HG-U133A arrays, more than 1000 probe sets showed greater than twofold reduced signals in the partially cleaved RNAs. Thus, direct comparison analysis of two samples or statistical evaluation of groups of samples with non-matching RNA quality bears an enormously high risk to measure falsely down-regulated transcripts in partially degraded RNAs, which even further increases as the gap between the 3′/5′-ratios of the samples to be compared widens (data not shown). A one-sided increase of the fold-change cut-off for down-regulated genes in the data mining process could help to greatly reduce the number of false positives in such poor versus good RNA comparisons. However, this does not represent an ideal solution, as it will also eliminate many true positive targets from the lists of down-regulated transcripts. Similarly, while the number of falsely down-regulated targets in degraded samples benefits from global scaling normalization due to the arithmetical compensation of the signal shift (compare *Figure 10.2B* and *C*), this procedure also cannot rescue such data,

as the rate of truly positive down-regulated genes is concomitantly reduced. Moreover, resulting adjusted signals on probe sets exhibiting far below average distances from the polyA site represent slightly over-estimated (over-scaled) values, explaining the increased incidence of false up calls in the SAM analysis comparing globally scaled arrays of degraded versus control samples. Clearly, knowledge of all the probe sets exhibiting above-average distance to the 3′-end of the transcript would allow us to put flags on error-prone probe sets and to more selectively sort out false positive targets occurring in poor versus good RNA comparisons. While the study presented in this chapter identified a substantial number of such probe sets on the HG-U133A array, it is obvious that this information cannot be readily transferred to HG-U133A data derived from other cells or tissues, due to differences in alternative splicing, usage of polyA sites and possible genomic alterations. However, to allow some basic risk assessment, it would be helpful if Affymetrix would integrate information regarding the distance of each probe set to the most 3′ located known polyA site into the annotation table that is provided for each GeneChip type. Moreover, as more and more GeneChip datasets are deposited in public databases, such as GEO (www.ncbi.nlm.nih.gov/geo), it appears feasible to recognize probe sets bearing a high risk for under-representation by statistical evaluation of excellent versus poor quality RNAs (based on the reported 3′/5′-ratios of the samples) in additional cell types and tissues.

In summary, our study reveals a highly variable and previously underestimated potential for erroneous measurement of transcript abundance by individual probe sets in GeneChip analyses of partially degraded RNAs. While Affymetrix GeneChips were used in this study, it is important to state that any type of oligonucleotide or cDNA array is prone to the described problem, if the chosen oligonucleotides or cDNA fragments display variable distances to the polyA end of the individual transcripts and preparation of the hybridization samples involves oligo(dT)-primed reverse transcription. Direct labeling of total RNA through cross-linking or the use of random primers during first strand cDNA synthesis might yield less 3′-biased targets from degraded samples, but require more material and capture the bulk of non-polyadenylated RNAs, resulting in strongly decreased sensitivity. Since there is currently no mRNA selective target preparation procedure in sight that avoids the introduction of 3′-bias in partially degraded RNAs, the newly designed Affymetrix GeneChip, Human Genome X3P, which contains more strongly 3′-biased probe set sequences than HG-U133A, might represent an improvement that eliminates at least part of the problem associated with the analysis of partially degraded RNA samples.

Acknowledgements

This study has been supported by a grant from the German National Genome Network Research Cancer Network (NGFN/BMBF). We thank Nadine Esser and Adriane Parchatka for excellent technical assistance.

References

1. Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98(9)**: 5116–5121.
2. Baugh LR, Hill AA, Brown EL and Hunter CP (2001 Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* **29(5)**: E29
3. Dürig J, Nückel H, Hüttmann A, *et al.* (2003) Expression of ribosomal and translation-associated genes is correlated with a favorable clinical course in chronic lymphocytic leukemia. *Blood* **101(7)**: 2748–2755.

Protocol

CONTENTS

Protocol 10.1: Affymetrix GeneChip analyses

HeLa cell total RNA was prepared according to the manufacturer's recommendation using TRIzol (Stratagene) and dissolved in nuclease-free water (Ambion). For each time point, two 40-μl RNA aliquots containing 34 μg of total RNA were each combined with 10 μl of 5 × RNA fragmentation buffer (200 mM Tris-acetate, pH 8.1, 500 mM KOAc, 150 mM MgOAc), heated for 70, 80, or 90 s at 90°C in a thermocycler and immediately chilled. Control reactions were kept on ice. After addition of 1 μl of a mixture of five different polyadenylated spike RNAs (DapX, LysX, PheX, ThrX, TrpnX) in staggered concentrations, duplicate control and heat-treated samples were pooled and repurified on RNeasy mini columns (Qiagen), including an on-column DNase I digestion step. RNA was eluted in water and quantified by OD260 measurement. As analyzed by non-denaturing agarose gel electrophoreses, RNAs subjected to heat fragmentation showed obvious signs of degradation, for example a clearly decreased ratio of the 28S and 18S ribosomal RNA bands, while controls appeared to represent intact RNA.

Synthesis of double strand cDNA was performed in duplicate with 13.5 μg of total RNA from control or progressively degraded samples as described previously by using an anchored T7-oligo-d$(T)_{21}$-V primer (5′-GCATTAGCGGCCGCGAAATTAATACGACTCACTATAGGGAGA$(T)_{21}$V-3′, MWG Biotech, Ebersberg, Germany) for first strand synthesis (2, 3). cDNAs were purified by phenol/chloroform/IAA/PLG extraction, precipitated and transcribed for 16 h at 37°C in 50-μl reactions containing 40 mM Tris-HCl pH 8.0, 16 mM $MgCl_2$, 2 mM spermidine, 5 mM DTT, 1.1% PEG 20 000, 4 mM GTP and ATP, 1.4 mM UTP and CTP, 0.6 mM Biotin–11-CTP and Biotin–11-UTP, 40 U pyrophosphatase, 50 U RNase inhibitor, and 1.5 μg T7 RNA polymerase. Biotinylated cRNAs were purified on RNeasy mini columns (Qiagen, Hilden) and quantified by OD260 measurement. Gel electrophoretic analysis of cRNA samples revealed a detectable shift to smaller average sizes in degraded samples, as expected (not shown). After heat fragmentation of cRNA samples to an average size of 100 to 200 nucleotides, hybridization of Affymetrix HG-U133A arrays with 10 μg of cRNA each from control (n =6) and samples degraded for 70 s (n =2), 80 s (n =2) or 90 s (n =2) was performed, followed by washing, staining, and scanning as recommended by the manufacturer (Affymetrix Expression Analysis Technical Manual, 2000). All 12 arrays used in this study were taken from the same batch (lot no. 3 000 016).

11 Molecular karyotyping by means of array CGH: linking gene dosage alterations to disease phenotypes

Joris Veltman and Lisenka Vissers

11.1 Introduction

Chromosome banding is one of the most widely used techniques in routine cytogenetics (1) and has been invaluable in the search for chromosomal aberrations causally related to, for example, congenital mental retardation and malformation syndromes. Conceptual and technical developments in molecular cytogenetics are now enhancing the resolving power of conventional chromosome analysis techniques from the megabase to the kilobase level. Tools that have mediated these developments include (i) the generation of genome-wide clone resources integrated into the finished human genome sequence, (ii) the development of high-throughput microarray platforms, and (iii) the optimization of comparative genomic hybridization protocols and data analysis systems. Together, these developments have accumulated in a so-called 'molecular karyotyping' technology that allows the sensitive and specific detection of single copy number changes of submicroscopic chromosomal regions throughout the entire human genome.

Array-based comparative genomic hybridization (CGH) builds upon previously well-established CGH procedures (2, 3) using differentially labeled test and reference DNAs to be co-hybridized to cloned genomic fragments with known physical locations, in a microarray format (4, 5). In comparison with conventional CGH, the array format provides a higher resolution, a higher dynamic range, and better possibilities for automation. In addition, it allows for direct linking of copy number alterations to known genomic sequences.

11.2 Array preparation, labeling, hybridization and data analysis

Many of the basic procedures followed in microarray-based genome profiling are similar, if not identical, to those followed in expression profiling, including the use of specialized microarray equipment and data-analysis tools. Since microarray-based expression profiling has been well established over the last decade, much can be learned from the technical advances made in this area. However, there are also distinct differences such as target and probe complexity, stability of DNA over RNA, the presence of repetitive DNA and the need to identify single copy number alterations in genome profiling.

Specifically, the array CGH procedure includes the following steps. First, large-insert clones such as BACs are obtained from a supplier of clone libraries such as BACPAC Resources at the Children's Hospital Oakland Research Institute (http://bacpac.chori.org/). Then, small amounts of clone DNA are amplified by either degenerate oligonucleotide-primed (DOP) PCR (6) or ligation-mediated PCR (7) in order to obtain sufficient quantities needed for spotting (~1 μg/μl). Next, these PCR products are spotted onto glass slides coated with substrates such as aminosilane using microarray robots equipped with high-precision printing pins. Depending on the amount of clones to be spotted and the space available on the microarray slide, clones can either be spotted once per array or in replicate. Repeated spotting of the same clone on an array increases the precision of the measurements if the spot intensities are averaged, and allows for a detailed statistical analysis of the quality of the experiments. Patient and control DNAs (100 ng–1 μg) are usually labeled with either Cy3 or Cy5-dUTP using random priming and are subsequently hybridized onto the microarray in a solution containing an excess of Cot1-DNA to block repetitive sequences. Hybridizations can either be performed manually under a coverslip, in a gasket with gentle rocking or, automatically using commercially available hybridization stations. These automated hybridization stations allow for an active hybridization process, thereby improving the reproducibility as well as reducing the actual hybridization time, which increases throughput. The hybridized DNAs are detected through the two different fluorochromes using standard microarray scanning equipment with either a scanning confocal laser or a charge coupled device (CCD) camera-based reader, followed by spot identification using commercially or freely available software packages.

The increase in data obtained through high-density arrays requires standardized storage systems as well as thorough statistical tools, similar to those required for microarray-based gene expression profiling (8, 9). Owing to the complicated process of producing and hybridizing spotted microarrays, a certain degree of systematic variation does exist in the data produced. Normalization of microarray data is used to eliminate such systematic variation and, therefore, represents an important preprocessing step in the analysis of almost all microarray data. One of the most frequently applied normalization procedures in microarray-based expression studies is the locally weighted scatterplot smoothing (LOWESS). Several laboratories have successfully introduced this procedure in array

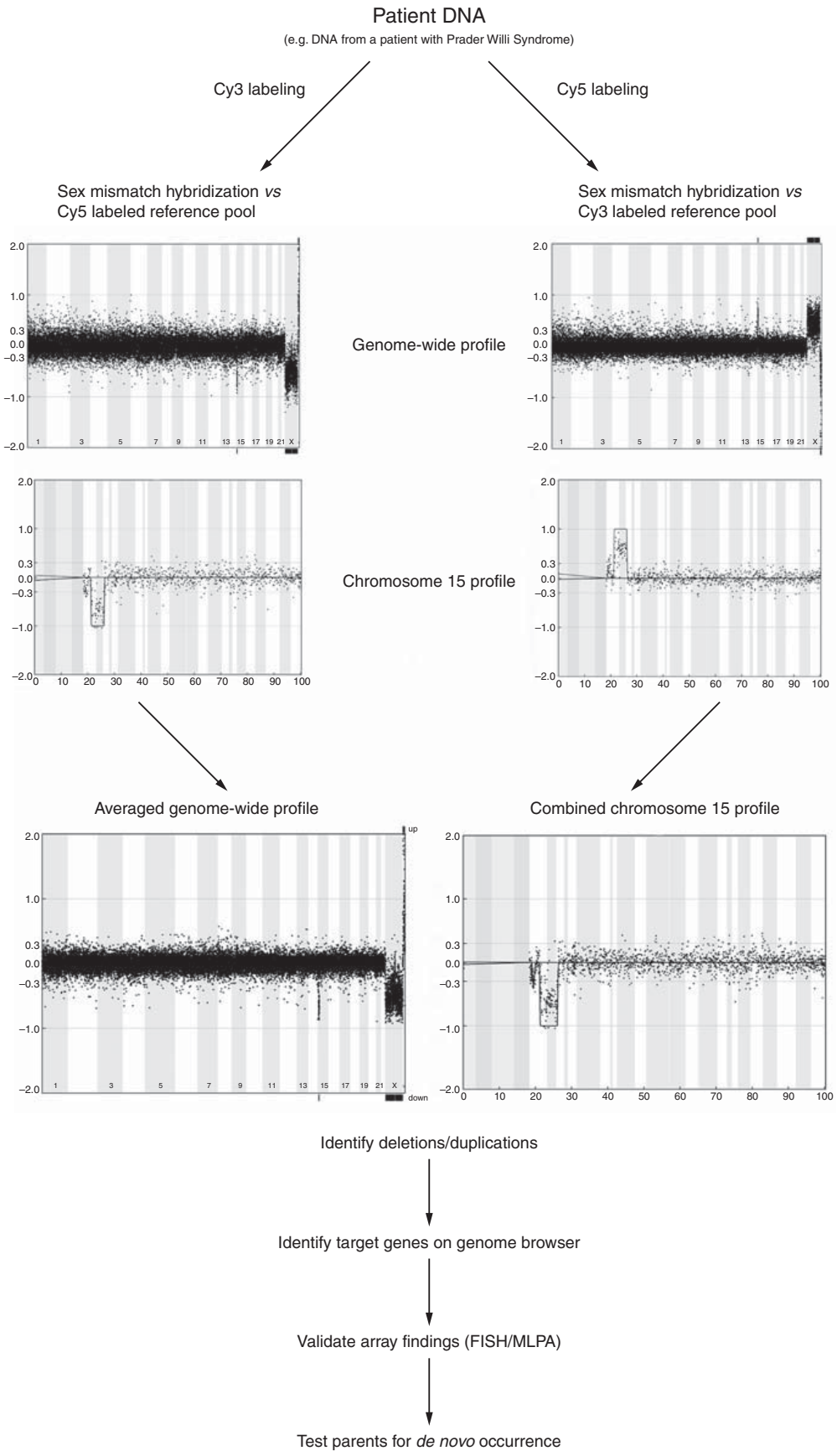
CGH applications (10–12). After data normalization, automated statistical procedures are required for reliable detection of genomic copy number changes. One such algorithm is the Hidden Markov Model (HMM) which, in our hands, is not only suited for distinguishing genuine copy number changes from random microarray noise, but also for precisely localizing the start- and end-points of each copy number alteration (unpublished results; see *Figure 11.1*). Finally, digitized intensity differences in the hybridization patterns of the DNAs onto the cloned fragments can be interpreted as copy number differences between the test and reference genomes.

This technique, once established and validated, allows high-throughput DNA copy number screening with a resolution limited only by the size of the clone fragments used (currently ~100 kb using BAC arrays).

11.3 Molecular karyotyping in clinical genetics

The microarray format allows for a virtually unlimited flexibility in the choice and amount of genomic fragments to be studied. Many laboratories have started their array CGH studies using low-density custom-made arrays consisting of probes specifically targeted at selected genomic regions. Examples of these are arrays targeting all subtelomeric regions (13, 14), arrays targeting regions known to be involved in microdeletion or microduplication syndromes (10, 15–19), or arrays targeting other chromosomal regions of interest (20–24). These low-density arrays are relatively easy to design and implement and, in contrast to high-density arrays, interpretation can be carried out without extensive statistical expertise. The ultimate power of the microarray-based approaches, however, lies in the genome-wide copy number assessment of patient samples without any *a priori* knowledge of the genomic regions involved. To this end, high-density arrays have been constructed with the aim of performing such genome-wide analyses, initially with a resolution of 0.75 to 1.4 megabases (11, 12, 25, 26) and, more recently, with a 50–100 kilobase tiling resolution using approximately 32 000 clones (27). In addition to home-made arrays, there are now several companies that offer microarrays for genomic profiling, either for genome-wide analyses or for more targeted approaches (28–30).

Array-based genome-wide copy number screening is expected to have a profound impact on the diagnosis and genetic counseling of patients with congenital mental retardation and malformation syndromes. Its resolution for identifying chromosomal abnormalities reaches far beyond the detection limit of routine chromosome banding techniques (reviews: 1, 31, 32). In addition, the integration of BAC clones into genome browsers like those from Ensemble and UCSC allows for a direct inspection of candidate genes affected by the copy number alterations, which will be of major help in explaining the phenotype. The first attempts to implement these technologies into a routine diagnostic setting are currently ongoing. Initial studies using 1-Mb genome-wide BAC arrays (11, 33) have indicated that this approach can reveal causative microdeletions and/or duplications in 10–20% of patients with unexplained mental retardation and congenital malformations, and this percentage is likely to increase with the introduction of full-coverage 50- to 100-kb resolution arrays. These pilot studies already have provided insight into the quality and reproducibility aspects


Patient DNA
(e.g. DNA from a patient with Prader Willi Syndrome)
Cy3 labeling
Cy5 labeling
Sex mismatch hybridization *vs*
Cy5 labeled reference pool
Sex mismatch hybridization *vs*
Cy3 labeled reference pool
Genome-wide profile
Chromosome 15 profile
Averaged genome-wide profile
Combined chromosome 15 profile
up
down
Identify deletions/duplications
Identify target genes on genome browser
Validate array findings (FISH/MLPA)
Test parents for *de novo* occurrence

of the array CGH procedure, the need for validation of microarray findings by independent technologies such as fluorescent *in situ* hybridisation (FISH) or multiplex ligation-dependent probe amplification (MLPA; 34), as well as the way to translate these molecular findings into clinical practice.

In this light it is important to note that these pilot studies have indicated that submicroscopic copy number alterations do not always have phenotypic consequences, as in some of the cases identical alterations were found in either one of the normal parents. This notion has been substantiated by recent studies revealing the presence of large copy number variations (LCVs) in apparently normal individuals (35, 36). In addition, once it has been established that a copy number alteration has occurred *de novo* in a patient, it may be that this alteration has not been described before in the literature, posing serious problems for genetic counseling. However, in due time increasing numbers of these abnormalities will be documented, either in individual case reports (see for example 37) or in publicly available online databases, and thus further our understanding of the genetic basis of these disorders.

11.4 Gene identification by array CGH

Haploinsufficiency of specific genes is a known cause of disease, both in acquired (cancerous) and congenital disorders. Haploinsufficiency can be brought about by single-base changes or deletions of stretches of DNA. The genome-wide detection of DNA alterations by array CGH can mark genomic sites where genes associated with a particular disease may be located. Previously, Albertson *et al.* (38) used this approach to identify *ZNF217* and *CYP24* as putative oncogenes in breast cancer. Similarly, we narrowed down the critical region of deletion for the gene(s) causing congenital aural atresia, located on chromosome 18q22.3-q23 (21). More recently, we have applied this approach successfully to identify the causative gene for

Figure 11.1.

Proposed flow-scheme for application of the array CGH procedure in a clinical setting. As an example DNA from a patient with Prader Willi Syndrome, containing a known microdeletion on 15q11-q13 is labeled and hybridized onto our genome-wide tiling resolution BAC array (~32 000 clones, each printed once onto the microarray slide) in a label swap experiment versus a sex-mismatched reference pool. The normalized $\log_2$ test-over-reference ratios (y-axis) are plotted for each clone ordered per chromosome by Mb position (x-axis), from p-ter to q-ter. Chromosome X and Y serve as control for deletion and duplication detection in these sex-mismatch experiments. The results of the individual experiments indicate the presence of potential copy number alterations; however, the combination of both experiments increases the statistical significance of an observed copy number alteration. This is indicated by the solid line at –1 for the microdeletion region (see combined chromosome 15 profile), which means a 100% probability for a downstate (deletion) as computed by a Hidden Markov Model. The gene content of this region can now be studied, if needed array findings can be validated (especially useful for deletions smaller than 1 megabase) and finally parents can be tested to distinguish causative alterations from normal variation.

CHARGE syndrome (39). CHARGE syndrome (OMIM 214800) is a non-random pattern of congenital anomalies including choanal atresia and malformations of the heart, inner ear and retina. In this study a *de novo* microdeletion of 4.8 megabases was identified on 8q12 by hybridizing genomic DNA from a CHARGE patient versus a normal control on our 1-Mb-resolution BAC array. A literature search identified another CHARGE patient with an apparently balanced chromosome 8 translocation, identified by routine karyotyping (40). Array CGH analysis of this patient revealed another microdeletion partially overlapping with the one encountered in our index patient. Further analysis of 17 additional CHARGE patients on a tiling resolution chromosome 8 array did not reveal additional microdeletions. As a next step, sequence analysis of nine genes located within the minimal region of deletion overlap identified 10 *de novo* heterozygous mutations in a novel gene called *CHD7*, including seven stop-codon mutations, two missense mutations and one mutation at an intron-exon boundary. This latter study showed that array CGH may indeed serve as an effective new approach to localize disease-causing genes. This approach is of particular interest for sporadic malformation syndromes that cannot be tackled by other mapping procedures because of reproductive lethality.

11.5 Summary

Chromosome analysis has rapidly developed in the post-genome era. Novel genomic tools like array-based comparative genomic hybridization (array CGH) allow the mapping of genomic copy number alterations at the submicroscopic level, directly linking disease phenotypes to gene dosage alterations. These approaches are excellently suited for gene identification studies, for genotype-phenotype mapping as well as for molecular karyotyping in a routine clinical setting, thereby rapidly bridging the gap between DNA diagnostics and cytogenetic diagnostics. With increased resolution, these technologies not only identify disease-causing alterations but also highlight variation in the normal population.

References

1. Ried T (2004) Cytogenetics – in color and digitized. *N Engl J Med* **350**: 1597–1600.
2. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**: 818–821.
3. Lichter P (2000) New tools in molecular pathology. *J Mol Diagn* **2**: 171–173.
4. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T and Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**: 399–407.
5. Pinkel D, Segraves R, Sudar D, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**: 207–211.
6. Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA and Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**: 718–725.

7. Klein CA, Schmidt-Kittler O, Schardt JA, Pantel K, Speicher MR and Riethmuller G (1999) Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc Natl Acad Sci USA* **96**: 4494–4499.
8. Stoeckert CJ Jr Causton HC and Ball CA (2002) Microarray databases: standards and ontologies. *Nat Genet* **32(Suppl)**: 469–473.
9. Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* **32(Suppl)**: 496–501.
10. Wang NJ, Liu D, Parokonny AS and Schanen NC (2004) High-resolution molecular characterization of 15q11-q13 rearrangements by array comparative genomic hybridization (array CGH) with detection of gene dosage. *Am J Hum Genet* **75**: 267–281.
11. Vissers LE, de Vries BB, Osoegawa K, *et al.* (2003) Array-based comparative genomic hybridization for the genome wide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet* **73**: 1261–1270.
12. Cowell JK, Wang YD, Head K, Conroy J, McQuaid D and Nowak NJ (2004) Identification and characterisation of constitutional chromosome abnormalities using arrays of bacterial artificial chromosomes. *Br J Cancer* **90**: 860–865.
13. Veltman JA, Schoenmakers EF, Eussen BH, *et al.* (2002) High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet* **70**: 1269–1276.
14. Harada N, Hatchwell E, Okamoto N, *et al.* (2004) Subtelomere specific microarray based comparative genomic hybridisation: a rapid detection system for cryptic rearrangements in idiopathic mental retardation. *J Med Genet* **41**: 130–136.
15. Yu W, Ballif BC, Kashork CD, *et al.* (2003) Development of a comparative genomic hybridization microarray and demonstration of its utility with 25 well-characterized 1p36 deletions. *Hum Mol Genet* **12**: 2145–2152.
16. Locke DP, Segraves R, Nicholls RD, Schwartz S, Pinkel D, Albertson DG and Eichler EE (2004) BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *J Med Genet* **41**: 175–182.
17. Klein OD, Cotter PD, Albertson DG, Pinkel D, Tidyman WE, Moore MW and Rauen KA (2004) Prader-Willi syndrome resulting from an unbalanced translocation: characterization by array comparative genomic hybridization. *Clin Genet* **65**: 477–482.
18. Shaw CJ, Shaw CA, Yu W, Stankiewicz P, White LD, Beaudet AL and Lupski JR (2004) Comparative genomic hybridisation using a proximal 17p BAC/PAC array detects rearrangements responsible for four genomic disorders. *J Med Genet* **41**: 113–119.
19. Van Buggenhout G, Melotte C, Dutta B, *et al.* (2004) Mild Wolf-Hirschhorn syndrome: micro-array CGH analysis of atypical 4p16.3 deletions enables refinement of the genotype-phenotype map. *J Med Genet* **41**: 691–698.
20. Buckley PG, Mantripragada KK, Benetkiewicz M, *et al.* (2002) A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum Mol Genet* **11**: 3221–3229.
21. Veltman JA, Jonkers Y, Nuijten I, *et al.* (2003) Definition of a critical region on chromosome 18 for congenital aural atresia by array CGH. *Am J Hum Genet* **72**: 1578–1584.
22. Solomon NM, Ross SA, Morgan T, *et al.* (2004) Array comparative genomic hybridisation analysis of boys with X linked hypopituitarism identifies a 3.9 Mb duplicated critical region at Xq27 containing SOX3. *J Med Genet* **41**: 669–678.
23. Ekong R, Jeremiah S, Judah D, *et al.* (2004) Chromosomal anomalies on 6p25 in iris hypoplasia and Axenfeld-Rieger syndrome patients defined on a purpose-built genomic microarray. *Hum Mutat* **24**: 76–85.

24. Veltman JA, Yntema HG, Lugtenberg D, *et al.* (2004) High resolution profiling of X chromosomal aberrations by array comparative genomic hybridisation. *J Med Genet* **41**: 425–432.
25. Snijders AM, Nowak N, Segraves R, *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29**: 263–264.
26. Fiegler H, Carr P, Douglas EJ, *et al.* (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* **36**: 361–374.
27. Ishkanian AS, Malloff CA, Watson SK, *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet* **36**: 299–303.
28. Larrabee PB, Johnson KL, Pestova E, Lucas M, Wilber K, LeShane ES, Tantravahi U, Cowan JM and Bianchi DW (2004) Microarray analysis of cell-free fetal DNA in amniotic fluid: a prenatal molecular karyotype. *Am J Hum Genet* **75**: 485–491.
29. Schaeffer AJ, Chung J, Heretis K, Wong A, Ledbetter DH and Lese Martin C (2004) Comparative genomic hybridization-array analysis enhances the detection of aneuploidies and submicroscopic imbalances in spontaneous miscarriages. *Am J Hum Genet* **74**: 1168–1174.
30. Schoumans J, Anderlid BM, Blennow E, Teh BT and Nordenskjold M (2004) The performance of CGH array for the detection of cryptic constitutional chromosome imbalances. *J Med Genet* **41**: 198-202.
31. Smeets DF (2004) Historical prospective of human cytogenetics: from microscope to microarray. *Clin Biochem* **37**: 439-446.
32. Shaffer LG and Bejjani BA (2004) A cytogeneticist's perspective on genomic microarrays. *Hum Reprod Update* **10**: 221–226.
33. Shaw-Smith C, Redon R, Rickman L, *et al.* (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet* **41**: 241–248.
34. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F and Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30**: e57.
35. Sebat J, Lakshmi B, Troge J, *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.
36. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004) Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949–951.
37. Koolen DA, Vissers LE, Nillesen W, Smeets D, van Ravenswaaij CM, Sistermans EA, Veltman JA and de Vries BB (2004) A novel microdeletion, del(2)(q22.3q23.3) in a mentally retarded patient, detected by array-based comparative genomic hybridization. *Clin Genet* **65**: 429–432.
38. Albertson DG, Ylstra B, Segraves R, Collins C, Dairkee SH, Kowbel D, Kuo WL, Gray JW and Pinkel D (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat Genet* **25**: 144–146.
39. Vissers LE, van Ravenswaaij CM, Admiraal R, *et al.* (2004) Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat Genet* **36**: 955–957.
40. Hurst JA, Meinecke P and Baraitser M (1991) Balanced t(6;8)(6p8p;6q8q) and the CHARGE association. *J Med Genet* **28**: 54–55.

Protocols

CONTENTS

MATERIALS

Reagents

Clone preparation	DOP PCR primers: 5′-CCG ACT GCA GNN NNN NAT GTG G-3′, dissolved in distilled water or TE (100 μM) – store at –20°C
	dNTP mix (dATP, dCTP, dGTP, dTTP; 100 mM) – store at –20°C
	Magnesium chloride (100 mM)
	Taq 2000 DNA polymerase (5 U/μl, Stratagene) – store at –20°C
	10 × PCR buffer (supplied with Taq DNA polymerase) – store at –20°C
	BAC DNA sample
	TE (10 mM Tris-HCl, 1 mM EDTA)
	Spotting buffer: 30% dimethyl sulfoxide (DMSO)
DNA purification	QIAamp kit containing lysis buffer AL, wash buffer AW1, wash buffer AW2 and AE elution buffer (QIAgen)
	Ethanol (100%)
	DNA samples
DNA sonication	Distilled water
	DNA samples
DNA labeling	Distilled water
	Bioprime labeling kit containing 2.5 × Random Primed and Klenow polymerase (Amersham Biosciences)
	dNTP mix containing 2 mM of the nucleotides dGTP, dCTP, dATP and 1mM of dTTP, 1 mM EDTA pH 8.0 and 10 mM Tris pH 8.0
	Cy3-dUTP (Amersham Biosciences)
	Cy5-dUTP (Amersham Biosciences)

Purification of labeled DNA	Labeled DNA samples
	QIAquick kit containing PB buffer, PE buffer and EB elution buffer (QIAgen)
DNA precipitation	Purified labeled DNA
	Human COT1-DNA (1 μg/μl, Roche)
	3 M NaAc pH 5.2
	100% ethanol
Probe preparation	Precipitated DNA (test and reference)
	Distilled water
	Yeast tRNA (100 mg/ml, Invitrogen)
	20% SDS
	Mastermix containing 50% formamide, 10% dextran sulfate, 2 × SSC and 4% SDS
Slide pre-treatment	Distilled water
	20 × SSC
	BSA
	1 M Tris pH 9.0
	10% SDS
	Ethanolamine
Posthybridization procedure	Formamide
	Filter sterile 20 × SSC pH 5.0
	Nonidet P-40
	Distilled water
	0.2 M Na_2HPO_4
	0.2 M NaH_2PO_4
Gel electrophoresis	0.5 × TBE
	Agarose
	100 bp marker
Additional reagents and materials:	Multiscreen plate (Millipore)
	50-ml falcon tube
	0.5-ml reaction tube, amber
	1.5-ml reaction tube
	0.2-μm filter
	Coplin jar

250-ml beaker

1-l beaker

Forceps

Bunsen burner

QIAamp DNA purification kit (QIAgen)

QIAquick PCR purification kit (QIAgen)

CMT-GAPS coated slides (Corning)

Exicator

Equipment

Arrayer (Omnigrid 100, Genomic Solutions)

UV Stratalinker

Thermal cycler

Speedvac

Gel electrophoresis system

Power supply

Sonicator (Soniprep; Soniprep150)

Water bath (37°C, 42°C and 50°C)

Hybridization station (GeneTAC, Genomic Solutions)

Centrifuges

Scanner with Cy3 and Cy5 lasers (Affymetrix 428 scanner, Affymetrix)

Protocol 11.1: Clone preparation and array fabrication

1. Prepare a PCR premix containing the following reagents (the amount of each reagent is given for one PCR reaction. They should be scaled up appropriately according to the number of PCRs being set up).

   80.5 μl sterile distilled water

   10.0 μl 10 × PCR buffer

   3.0 μl 100 mM $MgCl_2$

   2.3 μl 100 μM DOP PCR primer

   1.2 μl 100 mM dNTP

   1.0 μl Taq 2000 DNA polymerase

   Dispense a 98-μl aliquot of the PCR premix into a microtiter plate (or microcentrifuge tube) containing 2 μl of DNA sample (approximately 50–100 ng).

2. Place the microtiter plate in a thermal cycler. Suggested cycling conditions: denaturing at 94°C for 3 min, followed by 30 cycles 94°C for 30 s, 37°C for 30 s and a linear ramp from 37°C to 72°C and 72°C for 1 min, followed by a final extension of 10 min at 72°C.

3. DOP PCR products are purified using a multiscreen plate (Millipore) and finally dissolved in 55 μl TE.

4. Check amplified DNA quality and fragment length by agarose gel electrophoresis on a 1% 0.5 × TBE agarose gel. Load 1–2 μl of the purified DOP PCR product.

5. Dry the amplified BAC clones in a microtiter plate using a speedvac.

6. Dissolve BAC clones in spotting buffer containing 30% DMSO.

7. Arrays can now be prepared for spotting onto aminosilane-coated slides according to manufacturers instructions using a split-pin system. Depending on the number of clones to be spotted, they can be spotted once or in replicate.

8. Dry the slide overnight in the arrayer.

9. Crosslink the slide according to procedures provided by the different manufacturers.

10. Store the slide in an exicator at room temperature for up to 6 months.

Protocol 11.2: Array CGH procedure

DNA PURIFICATION

1. Purify the high-molecular DNA samples with QIAamp columns using the standard protocol supplied by the manufacturer.
2. Determine DNA concentration and purity by spectophotometry and agarose gel electrophoresis.

DNA SONICATION

1. Place the sonicator tip in a tube filled with distilled water.
2. Start sonication to clean the tip.
3. Dry the tip and place it in the tube containing the DNA sample. Sonicate the DNA sample.
4. Place the DNA sample on ice for at least 10 min.
5. Check fragment length of the sonicated DNA by agarose gel electrophoresis on a 1% 0.5 × TBE agarose gel.
6. The DNA must show a smear with estimated fragment size between 400 and 3000 kb. If not, repeat from step 3.

DNA LABELING

1. Dilute 500 ng sonicated genomic DNA to a final volume of 34.4 µl with distilled water; use amber 0.5-ml reaction tubes.
2. Add 32 µl of 2.5 × Random primer (mix before use), vortex 5–10 s and spin down.
3. Place the reaction for 10 min at 100°C (boiling water).
4. Place the reaction on ice for at least 5 min to ensure complete cooling of the reaction.
5. Spin down before adding the next components:

   4.0 µl Cy-dye (Cy3 for patient DNA, Cy5 for reference DNA)

   8.0 µl dNTPmix, vortexed before adding

   1.6 µl Klenow fragment DNA polymerase
6. Vortex the reactions gently and spin down.
7. Incubate the reactions overnight in a 37°C water bath.

PURIFICATION OF LABELING

1. Purify the labeled DNA to remove all unincorporated dyes by use of QIAquick columns as described by the manufacturer.
2. Elute the labeled DNA in two rounds of 60 μl EB buffer.

DNA PRECIPITATION

1. Mix the labeled patient (test) and control (reference) DNA (total volume is now 240 μl).
2. Add 120 μg of human COT1-DNA.
3. Add 0.1 × volume 3 M NaAc, pH 5.2.
4. Add 2.5 × volume 100% ethanol. Invert the reaction tube and precipitate the probe for at least 15 min at –20°C.
5. Centrifuge the DNA 30 min at 16 100 *g* at 4°C.

PROBE PREPARATION

1. Dry the pellet for a maximum of 10 min. Do not overdry the pellets because this will hamper resuspension. Keep the tubes in the dark during this procedure to prevent fading of the Cy-dyes.
2. Resuspend the pellet in 7.9 μl of distilled water, 12.0 μl of yeast tRNA and 16.3 μl 20% SDS.
3. Dissolve the pellet in the dark for 15 min. Subject the probes to a visual inspection to ensure that the probe is fully dissolved.
4. Add 84.0 μl of Mastermix. Vortex carefully and spin down.
5. Denature the probes at 70°C in a water bath for 15 min. Spin down afterwards.
6. Prehybridize the probes in a water bath of 37°C for 30 min.
7. The probes are now ready for hybridization on the hybridization station.

Slide pretreatment

NOTE 1: This section should be performed simultaneously with the purification of the labeling.

NOTE 2: Use aminosilane-coated slide only.

1. Prepare the following buffers and preheat at the right temperature:

Ethanolamine (120 ml; preheat at 50°C):

360 ml ethanolamine

12 ml Tris pH 9.0

1.2 ml 10% SDS

108 ml water

Blockwash buffer (120 ml; preheat at 50°C):

24 ml 20 × SSC

96 ml water

1.2 ml 10% SDS

Prehybridization buffer (120 ml; preheat at 42°C):

30 ml 20 × SSC

90 ml water

1.2 g BSA

1.2 ml 10% SDS

2. Place the slide in a coplin jar or slide container and pour prewarmed ethanolamine blocking buffer onto the slide, and incubate the slide for 60 min at 50°C.
3. Wash the slide 5–7 times with distilled water.
4. Pour prewarmed Blockwash buffer onto the slide and incubate for 60 min at 50°C.
5. Wash the slide 5–7 times with distilled water.
6. Place the slide into a vertical slide holder (prevent drying of the slide).
7. Place the slide in boiling distilled water for 3 min for denaturation of DNA on the array.
8. Wash the slide two times with distilled water.
9. Place the slide back into the coplin jar or slide-container and pour prewarmed prehybridization buffer and incubate for 45 min at 42°C.
10. Wash the slide five times with distilled water.
11. Dry the slide as quickly as possible by centrifugation (5 min, 216 *g*)
12. Scan the slide to check for autofluorescence. The scan should now show a very low autofluorescence signal, if any. If not, repeat pretreatment of the slide.

Protocol 11.3: Hybridization and posthybridization procedure

1. Prepare PN buffer for the posthybridization washes to be performed tomorrow.

   PN buffer (1000 ml):

   Take 473.5 ml 0.2 M Na_2HPO_4/0.1% NP-40.

   Add 400 ml distilled water.

   Adjust pH to 8.0 with 0.2 M NaH_2PO_4/0.1% NP-40.

   Adjust to 1000 ml with distilled water. Store at room temperature.

2. Prepare the hybridization station as suggested by the manufacturer.

3. Place the pretreated slide in the hybridization station.

4. Start the hybridization program:

   Step 1. O-ring conditioning: 3 min 75°C

   Step 2. Introduce probe: 37°C hold step until probe is loaded

   Step 3. Hybridization: Overnight 37°C hold step (usually 18 h)

   Step 4. Formamide wash: Temperature increases to 45°C followed by five cycles of 10-s flow time, 60-s hold time

   Step 5. Preparation for PN buffer wash: Reducing temperature to 20°C

   Step 6. PN buffer wash: Five cycles at 20°C of 30-s flow time, 60-s hold time

   Step 7. PN buffer hold: One cycle at 20°C of 10-s flow time, 2-h hold time

   NOTE: During the hold time of step 7, you can abort the program when you are ready to remove the slide from the station.

5. Remove the slide from the hybridization station and place it in a coplin jar filled with PN buffer.

6. Incubate the slide for 10 min.

7. Dip the slide in water and place it in a 50-ml falcon tube. Dry the slide immediately by centrifugation (216 *g*, 5 min, room temperature).

8. Slide is now ready for scanning. Keep the slide in the dark until use.

DNA microarrays: analysis of chromosomes and their aberrations

12

Heike Fiegler, Susan M Gribble and Nigel P Carter

12.1 Introduction

Comparative genomic hybridization (CGH) has been widely used for the analysis of copy number changes in tumors to identify genes involved in the development and pathogenesis of cancers. In conventional CGH, metaphase chromosomes are used as the target for hybridization (1). However, the resolution with which copy number changes can be detected using this technique is limited to approximately 3–5 Mb (2). By replacing the metaphase chromosomes with mapped sequences (typically from large insert clones such as BACs, PACs and cosmids), analysis resolution becomes dependent only on the insert size and the density of the clones used to construct the array.

For this purpose, the sequencing of the human genome has provided a valuable resource of mapped and sequenced clones. Due to the increased sensitivity and resolution compared to conventional CGH, array CGH is now not only being applied to the analysis of copy number alterations in tumors, but also to the identification of genomic imbalances (microdeletions/microduplications) in patients with constitutional rearrangements and to rapidly map translocation breakpoints.

12.2 Array construction and application of genomic microarrays

Array construction using DOP PCR

Large insert clones such as cosmids, bacterial artificial chromosomes (BACs) and P1 artificial chromosomes (PACs) are used typically for the construction of genomic DNA microarrays. DNA derived from these clones was originally prepared from large-scale cultures (3, 4), which, when expanded to the number of clones that are required to construct an array with a resolution of 1 Mb (~3500 clones) or even a whole genome tiling path array (~37 000 clones) becomes a costly and time-consuming procedure. Therefore, several PCR-based methods have been developed to remove the requirement of large-scale cultures. These include ligation-mediated PCR (5, 6), rolling circle PCR using Phi29 (7), or degenerate oligonucleotide primed

PCR (DOP PCR) using an amine-modified version of the standard DOP PCR primer 6MW (8).

We have also chosen a DOP PCR-based approach for the construction of a large insert clone DNA microarray. DOP PCR uses a mixture of primers, whereby each of the primers consists of a defined 5′ sequence and six defined bases at the 3′ end flanking random hexanucleotide sequence (9, 10). Thus, DOP PCR allows a general amplification of any target DNA. However, the reaction relies on the frequency with which bases in the target sequence match the six 3′ bases of the primer. If the matches to the primer are infrequent in a particular target sequence, the DOP PCR product will be a poor representation of the target (9, 10). In order to increase the level of representation, we have designed three DOP PCR primers that would be efficient in amplifying human genomic DNA, but inefficient in the amplification of *Escherichia coli* DNA, a known contaminant of DNA preparations from clone cultures (11). Amplification of the contaminating *E. coli* DNA together with the clone insert will reduce the capacity of the spotted product to hybridize with the DNA of interest thus reducing the sensitivity of the array. We found that the use of these three DOP PCR primers in combination resulted in a significant increase in signal to background ratio, sensitivity and reproducibility. Following this strategy, we have constructed a large insert clone DNA microarray consisting of 3523 Golden Path sequencing clones spaced at approximately 1 Mb intervals across the human genome. Each of the clones was amplified in three separate PCR reactions using the three different DOP PCR primers. This was followed by a secondary PCR reaction with a 5′ amine-modified primer designed such that the 3′ end matched the 5′ end of the DOP PCR primers to enable covalent attachment of the products to specially coated glass slides (12). In order to interpret and report copy number changes correctly across the genome it is essential to map the exact position of each clone along the chromosomes. Clone information is available through various genome browsers but often not easily accessible for large clone collections. We have therefore generated a specific view of the human genome within the Ensembl genome browser (Cytoview, www.ensembl.org/homo_sapiens/cytoview) that displays the 1 Mb clone set in relation to the Golden Path sequencing clones. In particular, Cytoview facilitates the downloading of clone lists (from specific regions of interest to whole chromosomes and the whole genome) together with their corresponding map position. Additional information, such as location by fluorescent *in situ* hybridization (FISH), BAC-end sequence data, genes or expressed sequence tags (ESTs) for any region of interest can be viewed within the context of the 1 Mb clone set. Ensembl also provides an automatic update of all this information with every new assembly of the human genome (12).

Array CGH in tumor biology

Many tumors are characterized by the presence of copy number alterations ranging from gains or losses of whole chromosomes to less than a megabase of DNA. In human cancers, regions of gain potentially harbor oncogenes, while tumor suppressor genes are likely to be located within regions of copy number loss. We have used the 1 Mb array described above in several large-

scale studies, one of which involved the screening of 22 bladder-tumor-derived cell lines. The array results confirmed numerous genetic changes previously identified by conventional CGH, M-FISH, or LOH analyses. The most frequent copy number alterations included a complete or partial loss of chromosome 4q and gain of chromosome 20q. In addition to previously identified homozygous deletions on chromosomes 9p21.3 (harboring CDKN2A), 9q33.1 (harboring DBCCR1) and 10q (harboring PTEN) in some of the cell lines used in this study, we could also identify several potentially new homozygous deletions and high-level amplifications with a previously reported amplification at 6p22.3 being the most frequent. Subsequent real time PCR analysis of genes in that region revealed a novel candidate gene (*NM_017774*) with consistent over-expression in all the cell lines displaying the 6p22.3 amplicon (13).

Another large-scale study used the 1 Mb array to investigate copy number changes in 48 colorectal cancer cell lines and 37 colorectal primary carcinomas. Colorectal cancer (CRC) is the second most common malignancy in the Western World and accounts for around 20 000 deaths in the UK per year. So far, only a few genes have been identified in which somatic mutations contribute to the pathogenesis of CRC. These include *APC*, *SMAD4*, *p53*, *KRAS*, and *β-catenin*. Conventional CGH analyses of CRCs revealed consistent copy number changes such as gain of chromosomes 20, 13 and 8q and loss of chromosomes 18q and 8p. These observations were confirmed by array CGH; however, due to the increased resolution we could also identify the most frequently altered regions within these large-scale gains or losses. The most frequently gained regions were detected on chromosomes 20q (harboring potential candidate genes such as *LIVIN*, *HD54*, *EEF1A2* and *PTK6*), and 13q (harboring *FLT1* and *FLT3*), and the most frequently lost region on chromosome 18q (harboring *SMAD2* and *SMAD4*). In addition, we detected previously unreported copy number changes, such as a common region of amplification on chromosome 17q11.2-q12 (harboring *AATF* and *TBC1D3*) and a common deletion on chromosome 1q41 (harboring *TGFβ2*). Interestingly, both, colorectal cancer cell lines and primary carcinomas revealed a strikingly similar pattern of copy number alterations across the genome (14).

Array CGH for cytogenetic analyses

Array CGH is being increasingly applied to the identification and analysis of sub-microscopic deletions (microdeletions) or gains (microduplications) in patients with constitutional genomic rearrangements in order to identify genes contributing to the patient's phenotype. Array CGH-based approaches are particularly suited for the analysis of patients with learning disability and dysmorphology (15–17). In a study of 50 patients with cytogenetically normal karyotypes but with learning disability and dysmorphic features, we identified 12 patients (24% in total) harboring subtle genomic copy number changes. These copy number aberrations ranged in size from those involving only a single clone to regions as large as 14 Mb. Interestingly, none of the rearrangements coincided with previously reported cases from similar patient groups. We detected seven different microdeletions of which six were *de novo* and one deletion inherited from

a phenotypically normal parent, and five different microduplications of which one was *de novo* and four inherited. While the *de novo* rearrangements are likely to account for the phenotype of the patient, the pathogenic significance, if any, of the inherited copy number changes is unknown (15).

Array CGH is also proving valuable in the more detailed analysis of genomic regions contributing to cytogenetically defined syndromes. We have used array CGH to study a patient with a chromosome 21-derived marker chromosome who displayed some features similar to Down syndrome. Down syndrome is usually caused by trisomy of chromosome 21 and is characterized by, for example, cognitive impairment, hypotonia and specific phenotypic features such as flat faces and ridge formation on hands and feet. The Down syndrome critical region, which is thought to be responsible for the physical phenotype of the patients, has been mapped to 21q22.1-21q22.3. However, it is not clear whether other regions on chromosome 21 contribute to the complex phenotype of Down syndrome. The patient with the chromosome 21-derived marker chromosome did not show any of the characteristic dysmorphic features of Down syndrome, but presented with learning disability and cognitive defects typical of Down syndrome. The array CGH analysis revealed a partial tetrasomy of chromosome 21 that did not involve the Down syndrome critical region. We suggest that the genes located within the amplified region may contribute to aspects of learning disability and cognitive impairment, but do not play a role in the typical dysmorphic features associated with Down syndrome (18).

Array painting

Although array CGH is able to identify genomic copy number alterations, chromosome rearrangements which do not result in genomic imbalance, such as reciprocal, balanced translocations, cannot be detected by this method. We have therefore developed a technique ('array painting') that utilizes flow-sorted derivative chromosomes in combination with the array technology to analyze the constitution and the breakpoints in balanced translocations. Briefly, each derivative chromosome involved in the translocation is flow-sorted, amplified by DOP PCR, differentially labeled and hybridized to the arrays. Signal intensities above background will only be obtained from clones that contain sequences present in the flow-sorted derivative chromosomes. The ratio of the intensities determines from which derivative chromosome the hybridizing DNA sequence has been derived. Breakpoint-spanning clones are identified by intermediate ratio values when sequences present on both derivatives hybridize to the same clone (19). For example, we have analyzed the DNA of a patient with a *de novo* 46,XY,t(17;22)(q21.1;q12.2) translocation by array painting. The derivative chromosomes 17 and 22 were flow-sorted, differentially labeled and hybridized to the 1 Mb array. Only clones representing chromosomes 17 and 22 showed strong signals above background. We also obtained weak signals on chromosome 19. However, chromosome 19 is close to the derivative chromosome 17 on the flow karyotype and inevitably contaminated the derivative chromosome 17 isolation during the sort. Plotting the fluorescent intensities against the position of the clones along the

chromosomes clearly showed a transition from low to high ratios (chromosome 17) or *vice versa* (chromosome 22), but intermediate ratios that would identify breakpoint spanning clones were not detected using the 1 Mb array (*Figure 12.1*). However, by constructing a custom array consisting of tiling path clones within the previously identified 1 Mb intervals and hybridizing the same derivative DNAs, breakpoint-spanning clones were identified and could be confirmed by FISH analysis (19).

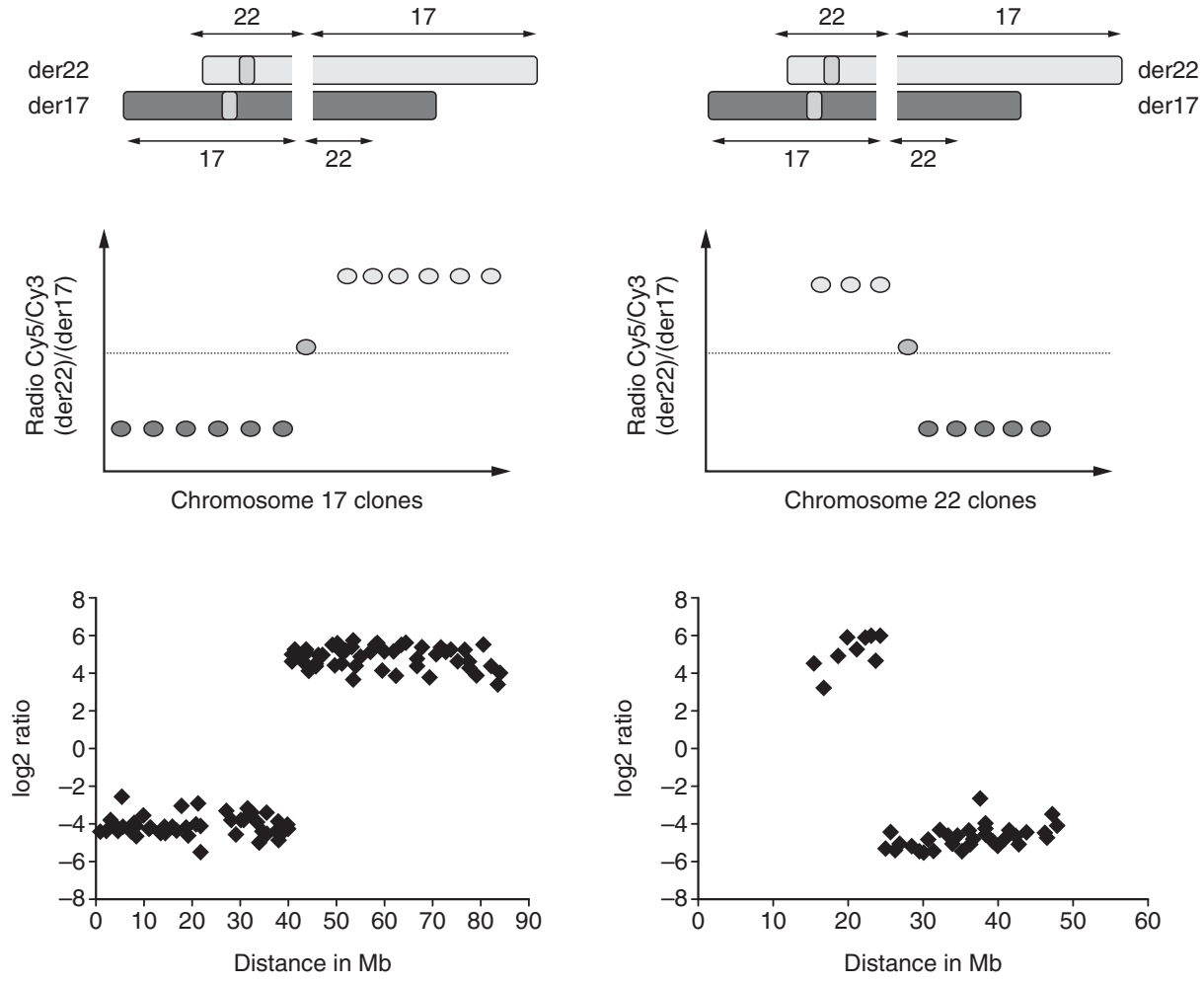


Figure 12.1.

Array painting results for chromosomes 17 and 22 in the analysis of a t(17;22) patient. The flow-sorted derivative chromosomes were differentially labeled with Cy3 (der17) and Cy5 (der22) and hybridized to the arrays. Only clones that correspond to sequences present in the derivative chromosomes will show signal intensities above background. The fluorescent ratio of the hybridizing clones will either be high or low depending to which derivative chromosome the sequence corresponds (e.g. clones mapping to the chromosome 17 sequences on the der17 will generate low ratios whilst clones mapping to the chromosome 17 sequences on the der22 will give high ratios). Breakpoint-spanning clones are identified by an intermediate ratio as both derivative chromosomes will hybridize to the same clone. Taken from *Journal of Medical Genetics*, Vol. 40, pages 664–670. Copyright (2003), with permission from the BMJ Publishing Group.

While array painting is particularly efficient in the analysis of the breakpoints of apparently balanced translocations, only the derivative chromosomes are screened. A more complete analysis of the patient's karyotype can be achieved by combining array painting of clearly rearranged chromosomes with the more general array CGH approach for the detection of genomic imbalances. In a study of 10 patients with apparently balanced chromosome translocations, we found a surprisingly high level of breakpoint complexity and genomic imbalance. Six of the patients studied showed complex rearrangements involving deletions, inversions and insertions at or near a breakpoint, the involvement of additional chromosomes in the translocation process and previously undetected microdeletions or microduplications unrelated to the translocation (20).

12.3 Conclusion

The analysis of genomic copy number aberrations and chromosomal rearrangements using large insert clone DNA microarrays is becoming an increasingly widespread application. The development of chromosome specific or whole genome-wide tiling path arrays, as well as arrays consisting of clones with inserts of even smaller size such as fosmids or PCR products, will allow the detection of increasingly subtle genomic changes to be identified by this technology. Thus, microarray analysis will facilitate the identification of new genes involved in tumor development and progression, as well as genes contributing to cytogenetically important syndromes and previously undiagnosed cases.

References

1. Kallioniemi OP, Kallioniemi A, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1993) Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. *Semin Cancer Biol* **4**: 41–46.
2. Lichter P, Joos S, Bentz M and Lampel S (2000) Comparative genomic hybridization: uses and limitations. *Semin Hematol* **37**: 348–357.
3. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T and Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**: 399–407.
4. Pinkel D, Segraves R, Sudar D, *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**: 207–211.
5. Bruder CE, Hirvela C, Tapia-Paez I, *et al.* (2001) High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. *Hum Mol Genet* **10**: 271–282.
6. Snijders AM, Nowak N, Segraves R, *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29**: 263–264.
7. Buckley PG, Mantripragada KK, Benetkiewicz M, *et al.* (2002) A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum Mol Genet* **11**: 3221–3229.
8. Hodgson G, Hager JH, Volik S, *et al.* (2001) Genome scanning with array CGH

delineates regional alterations in mouse islet carcinomas. *Nat Genet* **29**: 459–464.
9. Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA and Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**: 718–725.
10. Telenius H, Pelmear AH, Tunnacliffe A, *et al.* (1992) Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes. *Genes Chromosomes Cancer* **4**: 257–263.
11. Foreman PK and Davis RW (2000) Real-time PCR-based method for assaying the purity of bacterial artificial chromosome preparations. *Biotechniques* **29**: 410–412.
12. Fiegler H, Carr P, Douglas EJ, *et al.* (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* **36**: 361–374.
13. Hurst CD, Fiegler H, Carr P, Williams S, Carter NP and Knowles MA (2004) High-resolution analysis of genomic copy number alterations in bladder cancer by microarray-based comparative genomic hybridization. *Oncogene* **23**: 2250–2263.
14. Douglas EJ, Fiegler H, Rowan A, Halford S, Bicknell DC, Bodmer W, Tomlinson IP and Carter NP (2004) Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res* **64**: 4817–4825.
15. Shaw-Smith C, Redon R, Rickman L, *et al.* (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet* **41**: 241–248.
16. Veltman JA, Schoenmakers EF, Eussen BH, *et al.* (2002) High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet* **70**: 1269–1276.
17. Vissers LE, de Vries BB, Osoegawa K, *et al.* (2003) Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet* **73**: 1261–1270.
18. Rost I, Fiegler H, Fauth C, *et al.* (2004) Tetrasomy 21pter—>q21.2 in a male infant without typical Down's syndrome dysmorphic features but moderate mental retardation. *J Med Genet* **41**: e26.
19. Fiegler H, Gribble SM, Burford DC, *et al.* (2003) Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays. *J Med Genet* **40**: 664–670.
20. Gribble SM, Prigmore E, Burford DC, Porter KM, Carter, NP (2005) The complex nature of constitutional de novo apparently balanced translocations in patients presenting with abnormal phenotypes. *J Med Genet* **42**: 8–16.

Protocols

CONTENTS

Protocol 12.1: DOP PCR

REAGENTS

DOP PCR primer:

DOP 1: CCGACTCGAGNNNNNNCTAGAA

DOP 2: CCGACTCGAGNNNNNNTAGGAG

DOP 3: CCGACTCGAGNNNNNNTTCTAG

TAPS salt solution (final volume 96 ml), store at –20°C:

250 mM TAPS, pH 9.3	6.08 g
166 mM $(NH_4)_2SO_4$	2.20 g
25 mM $MgCl_2$	2.5 ml of 1 M stock solution

TAPS2 buffer (final volume 1 ml), UV-sterilize prior to use:

TAPS salt solution	960 µl
Bovine serum albumin (BSA), 5% stock solution	33 µl
β-mercaptoethanol	7 µl

DOP PCR (REACTION VOLUME 50 µL):

TAPS2-buffer	5 µl
DOP primer (20 µM)	5 µl
dNTPs (2.5 mM each)	4 µl
Polyoxyethelene ether W1 (1%)	2.5 µl
AmpliTaq polymerase	0.5 µl
H_2O	28 µl
Clone template DNA (1 ng/µl)	5 µl

PCR reactions are performed in PTC-225 Tetrad thermocyclers (MJ Research). After initial denaturation at 94°C for 3 min, the reaction is as follows: 10 cycles of 94°C for 1.5 min, 30°C for 2.5 min, ramp at 0.1°C s^{-1} to 72°C, 72°C for 3 min followed by 30 cycles of 94°C for 1 min, 62°C for 1.5 min, 72°C for 2 min, and a final extension step of 72°C for 8 min. The average obtained product size ranges from 0.2 to 2 kb.

Protocol 12.2: Aminolinking PCR

REAGENTS

Amino Primer:

GGAAACAGCCCGACTCGAG

|

NH_2

Aminolinking buffer:

KCl	500 mM
$MgCl_2$	25 mM
Tris pH 8.5	50 mM

The Aminolinking buffer should be made up prior to use and filter-sterilized (0.2 μm syringe filter). Storage at –20°C is not recommended.

Aminolinking PCR (reaction volume 60 μl):

Aminolinking buffer	6 μl
dNTPs (2.5 mM each)	6 μl
Aminoprimer (200 ng/μl)	3 μl
AmpliTaq polymerase	0.6 μl
H_2O	42.4 μl
Template (DOP-PCR product)	2 μl

PCR reactions are performed in PTC-225 Tetrad thermocyclers (MJ Research). After initial denaturation at 95°C for 10 min, the reaction was as follows: 35 cycles of 95°C for 1 min, 60°C for 1.5 min, 72°C for 7 min, followed by a final extension at 72°C for 10 min.

After combining the appropriate amino-linked products, the samples are prepared for arraying by adding 39 μl 4 × microarray spotting buffer (1 M sodium phosphate buffer, pH 8.5, 0.001% sarkosyl) to 120 μl of PCR products. The samples are then filtered by centrifugation at 600 *g* for 5 min through 0.2-μm filter plates.

Protocol 12.3: DOP PCR amplification of flow-sorted chromosomes

PRIMARY DOP PCR REACTION:

To each tube of flow sorted chromosomes (500 sorted chromosomes in 33 μl H_2O):

TAPS2 buffer	5 μl
DOP2 primer (20 μM)	5 μl
dNTP mix (2.5 mM each)	4 μl
Polyoxyethelene ether W1 (1%)	2.5 μl
AmpliTaq	0.5 μl

PCR reactions are performed in Biometra thermocyclers. After initial denaturation at 94°C for 10 min, the reaction is as follows: 10 cycles of 94°C for 1.5 min, 30°C for 2.5 min, ramp at 0.23°C s^{-1} to 72°C, 72°C for 3 min followed by 30 cycles of 94°C for 1 min, 62°C for 1.5 min, 72°C for 3 min, and a final extension step of 72°C for 8 min.

SECONDARY DOP PCR REACTION TO INCREASE THE CONCENTRATION OF THE PCR PRODUCTS:

Primary DOP PCR product	2 μl
TAPS2 buffer	5 μl
DOP2 primer (20 μM)	5 μl
dNTP mix (2.5 mM each)	4 μl
Polyoxyethelene ether W1 (1%)	2.5 μl
H_2O	16.5 μl
Amplitaq	0.5 μl

PCR reactions are performed in Biometra thermocyclers. After initial denaturation at 94°C for 4 min, the reaction is as follows: 35 cycles of 94°C for 1 min, 62°C for 1 min, 72°C for 1.5 min, and a final extension step of 72°C for 9 min.

Protocol 12.4: Random primed labeling of DNA for array CGH

EQUIPMENT AND REAGENTS

BioPrime Labeling Kit (Invitrogen)

10 × dNTP mix (1 mM dCTP, 2 mM dATP, 2 mM dGTP, 2 mM dTTP in TE buffer)

1 mM Cy3-dCTP (NEN Life Science)

1 mM Cy5-dCTP (NEN Life Science)

Micro-spin G50 columns (Pharmacia Amersham)

METHOD

1. 0.15 μg DNA and 60 μl 2.5 × Random Primers Solution are resuspended in water to a final volume of 130.5 μl.
2. The solution is denatured in a heat block for 10 min at 100°C, and immediately cooled on ice.
3. The following reagents are added on ice:
   - 15 μl 10 × dNTP mix
   - 1.5 μl Cy3- or Cy5-labeled dCTP
   - 3 μl Klenow fragment.
4. The reaction is incubated at 37°C overnight and stopped by adding 15 μl of stop buffer supplied in the kit.
5. Labeled nucleotides are removed from the DNA labeling reactions using microspin G50 columns according to the instructions of the suppliers.

Protocol 12.5: Array hybridization (grid size 2 × 3 cm)

EQUIPMENT AND REAGENTS

	Pre-/Hybridization buffer:	50% formamide
		10% dextran sulfate
		0.1% Tween 20
		2 × SSC
		10 mM Tris pH 7.4
	3 M NaAc pH 5.2	
	Human Cot1 DNA	
	Herring sperm DNA	
	100% ethanol, 80% ethanol	
	Yeast tRNA (100 μg μl^{-1}, dissolved in H_2O)	

HYBRIDIZATION

Tube 1:	Cy3-labeled DNA	180 μl
	Cy5-labeled DNA	180 μl
	Human Cot1 DNA	135 μl
	3M NaAc pH 5.2	55 μl
	Yeast tRNA	6 μl
	100% EtOH (cold)	1000 μl
Tube 2:	10 mg ml^{-1} herring sperm DNA	80 μl
	Human Cot1 DNA	135 μl
	3M NaAc pH 5.2	23 μl
	100% EtOH (cold)	400 μl

Samples are mixed and precipitated at –20°C overnight or for 30 min at –70°C.

Pre-hybridization

1. Precipitated DNAs are spun for 15 min at max. speed.

2. Pellets are washed with 500 µl 80% EtOH, and re-spun at max. speed for 5 min.
3. Supernatant is removed and samples are re-spun at max. speed for 1 min.
4. Supernatant is taken off with a small tip. Process is repeated until the pellet is dry.
5. DNAs are resuspended in:

   Tube 1: 35 µl Hybridization buffer

   a. Samples are denatured for 10 min at 70°C.

   b. Incubate for 60 min in a 37°C heat block (in the dark)

   Tube 2: 45 µl Hybridization buffer

   a. Samples are denatured for 10 min at 70°C.

   b. 30 µl of the denatured herring sperm/Cot1 mix are applied onto the grid which is then covered with a coverslip.

   c. The microarray slide is transferred into a humidity chamber saturated with 40% formamide/2 × SSC and incubated for 60 min.

   d. The coverslip is removed after incubation by placing the slide into a tall glass trough containing PBS.

   e. The pre-hybridization solution is washed off and the slide is dried by spinning at 150 *g* for 1 min.

Hybridization

Thirty µl of the hybridization solution is applied onto the grid which is then covered with a coverslip. The slide is placed into a slide mailer humidified with 20% formamide/2 × SSC, sealed with Parafilm and incubated at 37°C for 24–48h.

Washing

1. The coverslip is removed by placing the slide into a tall glass trough containing PBS/0.05% Tween 20 to wash off any excess hybridization solution.
2. The slide is then transferred into a new glass trough and washed in PBS/0.05% Tween 20 for 10 min at RT (shaking).
3. Slide is placed into a pre-heated 50% formamide/2 × SSC solution and incubated for 30 min at 42°C (shaking).
4. The slide is then transferred into fresh PBS/0.05% Tween 20 and washed for 10 min at RT (shaking).
5. The slide is dried by spinning at 150 *g* for 1 min and can be stored in a light-proof box until ready for scanning.

13 Mapping transcription factor binding sites using ChIP-chip – general considerations

Rebecca Martone and Michael Snyder

13.1 Introduction

The precise spatial and temporal control of gene expression is crucial for mediating cellular and developmental processes. A global understanding of transcriptional regulation will require the identification of *cis*-acting regulatory elements, the factors that bind them, and the regulatory cascades in which these factors function. With the recent determination of the genome sequence of humans and many other organisms it is now possible to identify many genes; however, the factors, DNA sequences and regulatory pathways that control their expression remain poorly understood. Below we discuss the use of chromatin immunoprecipitation and DNA microarrays to globally identify targets of transcription factors and use this information to construct regulatory networks.

A variety of indirect approaches have been used to identify targets of transcription factors. Analysis of genes with common expression patterns has led to the identification of shared motifs (e.g. 1); however the factor that binds these motifs is not usually apparent. The monitoring of gene expression patterns with and without a transcription factor of interest using expression microarrays or differential display can identify candidate targets (2). Although comprehensive, this approach identifies alterations in gene expression, which may be due to secondary or downstream signaling events. Similarly, genetic screens for identifying targets are neither direct nor comprehensive. Methods involving *in vitro* DNA binding selections do not account for the complexities of cooperative binding or cofactor regulation of transcription and therefore do not accurately mimic *in vivo* binding. Hence, a need exists for a direct method to identify all of the DNA targets of a transcription factor of interest in a single experimental system.

In the last few years a new microarray (chip) technology has allowed the use of chromatin immunoprecipitation (ChIP) (3, 4) to identify *in vivo* targets of transcription factors on a genome-wide scale; this procedure has been coined 'ChIP-chip'. ChIP-chip has gained increasing popularity as a means of identifying transcription factor targets on a global scale. The

experimental approach is built on the premise of reversibly crosslinking proteins to DNA so that the DNA attached to a protein of interest can then be isolated by purifying the protein and subsequently used to probe a microarray composed of regulatory regions containing putative transcription factor binding sites (*Plate 5*).

This approach was first invented for yeast (5, 6) and has now been successfully performed for a multitude of different transcription factors in many organisms (see *Table 13.1*). Several important considerations are critical to the ChIP-chip approach including experimental design, array selection, and data analysis. Here we will focus on experimental considerations and then report several studies employing ChIP-chip, which are meant to highlight the various microarray platforms suitable for this type of analysis.

13.2 Experimental approach

Published ChIP-chip protocols are fairly consistent, despite the organism or factor of interest (6–8). A general scheme of the approach is depicted in *Figure 13.1*. Generally, cells are grown in the conditions that are relevant to the goal of the experiment and treated with formaldehyde to crosslink protein-DNA, as well as protein-protein, interactions. Although other crosslinking agents are available, formaldehyde is attractive because heating samples at 65°C reverses the crosslinking, allowing the DNA to be used in subsequent enzymatic reactions (4). Extracts are then prepared and the chromatin is either sonicated to shear the DNA to 500–700 base pairs for spotted arrays (9), or enzymatically treated to generate much smaller fragments for higher-resolution oligo arrays (10). Following clarification of the extracts, the protein-bound DNA is selected by immunoprecipitation with either antibodies specific for the protein of interest or antibodies specific for a tag if the protein has been tagged with a moiety such as *myc* or HA. The crosslinks are then reversed and the DNA is purified. The DNA is then subjected to a labeling reaction, slides are hybridized and scanned, the image is quantified and the data analyzed.

Prior to hybridization, the purified DNA is fluorescently labeled by generally one of two methods: either direct incorporation of a modified nucleotide or chemical coupling of a fluorescent molecule after incorporation of an aminoallyl-modified nucleotide (6, 8). The reference sample is usually labeled with one fluor, that is Cy3, while the experimental or enriched DNA sample is labeled with another fluor, Cy5 for example. The probes can then be combined and hybridized to a microarray. Both the hybridization buffer and reaction temperature must be optimize, as both will impact on stringency potentially altering the final outcome of the experiment. After several washing steps of varying stringency to remove unincorporated dye and non-hybridizing DNA, slides are scanned with lasers optimized for detection at specific wavelengths corresponding to the dyes. Most scanners are equipped with dual lasers, thus allowing for rapid simultaneous data acquisition of both background and experimental probes. The results of the hybridization will allow detection of enriched segments in the IP reaction, thereby identifying transcription factor binding sites.

13.3 Experimental considerations

Several considerations are important in the experimental design. First, due to the low yield of DNA in the ChIP it is important to determine the optimal number of cells to generate robust signals on the microarray. This can also vary depending on the labeling technique, as some protocols allow for less starting material by adding an amplification step to the labeling reaction (10, 11). Others choose to avoid amplification biases by increasing the number of cells (9, 12).

Another important aspect to consider is the use of antibodies or epitope tags in the immunoprecipitation step. In organisms where inserting an epitope directly at the gene locus is readily feasible, such as the case in yeast, this approach is often ideal since a standard well-proven protocol is available. This allows researchers to directly compare tagged versus untagged samples to determine transcription factor binding sites (see below). Generally the signals generated by this approach are robust as several copies of the tag are inserted into the genome and antibodies directed toward the tag are highly specific. One drawback, however, is that the protein is modified so the binding profile is not of a native protein. Thus, it is important to ensure that the modified factor is functional using genetic and/or biochemical tests.

In most organisms it is not possible to epitope tag a protein expressed at endogenous levels. Although it is often possible to transfect tagged constructs in these systems, this runs the caveat that the levels of protein will affect site occupancy and thus the final target lists. Thus, for these systems it is most desirable to use antibodies specific for the protein of interest. Antibodies for many factors are available, although their quality and/or characterization are often poor. Thus, it is important to ensure that antibodies can specifically immunoprecipitate a factor of interest. In addition, if a *bona fide* target of the transcription factor is known, then a standard ChIP using PCR confirmation should be performed to ensure the antibodies function in a ChIP assay.

One final important consideration is the selection of a reference channel so that enrichment for transcription factor binding in the ChIP can be determined. This reference or background sample is commonly either genomic DNA or a mock immunoprecipitation (IP). However, it can also be IP DNA from a non-induced sample that is compared with the induced state. The latter is obviously attractive when studying an inducible transcription factor where binding only occurs in one state. Transcription factors that translocate to the nucleus upon activation such as NF-κB and STAT1 are suitable for this type of approach (9). Whichever reference is chosen, the main consideration is an adequate signal-to-noise ratio so that the highest quality data is generated.

13.4 Data analysis

The analysis of DNA microarray data is discussed in detail in other chapters and will be discussed only briefly here. Inherent to all microarray experiments are large, noisy datasets where the challenge is to process them

in a fashion that produces the most reliable and meaningful results. The data acquisition and analysis processes are fairly well established: first the slides are scanned to measure fluorescence intensity, signals from the features are then scored and normalized, the differential expression is determined and finally statistical methods and clustering algorithms are employed to generate meaningful datasets. The scanning and quantification steps are generally speaking very straight forward, however the normalization of the raw data and the calling of 'hits' is an area with a lot of experimental variability.

Both differential expression and ChIP-chip experiments rely on the ratio of 'reference' signals to 'experimental' or 'test' signals; however it quickly becomes apparent to the experimentalist that several parameters can potentially skew the raw data resulting in erroneous face value ratios including: differences in dye intensity, regional heterogeneity on the slide, faulty washing leaving behind dust and salt residues, and differences in spot intensity. Inconsistencies introduced in the hybridization reaction and subsequent washing steps, as well as experimental noise could alter the final outcome of the experiment if they are not accounted for (13).

Seeking to assist microarray users with generating meaningful datasets in a consistent and automated fashion, various web-accessible data processing platforms have been developed. One example is the ExpressYourself program developed by bioinformaticists at Yale University. The aim of this program is to address each aspect of the data analysis process, including: background correction, intra- and inter-slide normalization, data quality and scoring. One aspect of this program that sets it apart from others is that it allows users to select from a variety of scoring algorithms, some of which have been especially tailored to handle datasets generated by ChIP-chip experiments. Typically, researchers identify targets that have significant p values less than 10^{-4} or standard deviations from the median in multiple experiments (13).

13.5 Array selection

The ChIP-chip procedure has now been successfully carried out for many factors and a number of different organisms. Ideally the microarray used in a ChIP-chip experiment will contain the entire sequence of the genome of interest, as this provides the most comprehensive platform for monitoring all potential regulatory sequences. This has been feasible for smaller organisms such as *Saccharomyces cerevisiae* where ChIP-chip studies on individual transcription factors were first carried out (5, 14). These experiments used intergenic arrays containing PCR products representing all the non-coding sequence in the yeast genome and operated under the presumption that this sequence would contain all of the regulatory sites. Shortly after these initial studies, separate large-scale reports in *S. cerevisiae* emerged, including the mapping of yeast transcription factors involved in the cell cycle (15, 16), revealing the power of ChIP-chip in elucidating transcriptional circuitries. These whole-genome studies were feasible due to the size and relatively compact nature of the yeast genome, where the intergenic regions contain most of the regulatory regions and are relatively small.

After the success of the technique in yeast the natural progression was to scale up to mammalian systems. Ideally when setting out to do ChIP-chip, one would use arrays that tile across both the coding and non-coding regions of the genome, however until recently the lack of reliable and stable genomic sequence in higher eukaryotic organisms has been a major limiting factor. Likewise, researchers have had to grapple with several challenges associated with more complex genome sequence; these include repetitive elements, poor annotation, and large spans on intergenic and intronic sequence.

Despite these limitations, several reports of ChIP-chip carried out in higher organisms have been published (*Table 13.1*). The first reports of mammalian ChIP-chip came in 2002 (7, 11, 16, 17). Two groups independently investigated the cell-cycle-specific transcription factor E2F binding (7, 17). These initial studies overcame the challenges associated with the more complex mammalian genome mentioned above by the use of arrays containing limited regions of the genome. One group fabricated a CpG island microarray containing 7776 distinct DNA elements that were selected based on their high GC content (7). Many regulatory elements reside in CpG islands (18), and thus would presumably be a useful means to study transcription factor binding. However CpG islands do not guarantee the inclusion of all promoter elements, nor does it take into consideration binding outside of promoter sites. Despite these drawbacks, 68 target sites were identified representing genes involved in the cell cycle

Table 13.1. ChIP-chip studies to date.

Organism	Type of array	Transcription factor	Reference
Saccharomyces cerevisiae	Whole genome (intergenic)	Gal4, Ste12	(14)
		SBF, MBF	(5)
		106 general TFs	(15)
		Nine G1/S-specific TFs	(16)
		Ste12, Tec1	(21)
		203 general TFs	(22)
		Fhl1, Ifh1	(23)
	Whole genome (ORF & intergenic)	Histone H4	(24)
		TBP	(25)
Drosophila melanogaster	Whole-chromosome tiling array	Pol II, ORC	(26)
Homo sapiens	Promoter/CpG	E2F4	(7, 17)
		pRb	(27)
		HNF1α, HNF4α, HNF6 and Pol II	(28)
		ERα	(29)
		PRC2/3	(30)
	Genomic tiling	GATA1	(11)
		NF-κB (p65)	(9, 19)
		Sp1, c-Myc, p53	(10)
		CREB	(20)

as well as in DNA damage and recombination – two processes not previously known to be regulated by E2F.

The second group chose to look at E2F binding in promoter regions using an alternative approach. They constructed a microarray containing the proximal promoter regions of 1444 human genes, around 1200 of which were cell-cycle-regulated as the E2F family of transcription factors is known to function in regulating this cellular event (17). The array contained PCR products designed to amplify 700 bp upstream and 200 bp downstream of the putative transcription start site of the genes. Although successful in identifying targets of E2F, the study was less than comprehensive, as the approach was biased to cell-cycle-specific genes and limited to promoter regions identified by potentially error-prone computer algorithms. Despite their limitations, both studies mapping E2F binding represent pioneering efforts to globally map transcription factor binding sites.

The third group used a limited genomic tiling array in which all sequences of a region of interest were represented on the array. The binding profile of the transcription factor GATA1 within the β-globin locus was mapped using an array composed of 1-kb PCR fragments representing the entire 75-kb locus (11). This study was comprehensive in that no interactions with DNA go undetected due to lack of content on the array. Indeed a new binding site for the well-studied GATA–1 factor was discovered.

Building on the tiling approach researchers generated a tiling array of an entire human chromosome. The binding of the NF-κB family member p65 along human chromosome 22 was elucidated using an array that had nearly complete coverage of the non-repetitive sequence of the chromosome (9, 19). This microarray was comprised of nearly 22 000 PCR fragments representing both the coding and non-coding portions of the chromosome, making it suitable to survey transcription factor binding in an unbiased and comprehensive fashion. Human chromosome 22 is 34 MB in length, and although this represented only 1% of the genome, it was an important analysis of transcription factor binding nonetheless. It revealed that p65 bound genomic regions in addition to the expected 5′ proximal promoter sites – including intronic regions, intergenic regions, as well as near novel transcribed regions. This study was the first mapping of a transcription factor on an entire chromosome and revealed the importance of tiling arrays in identifying potential regulatory regions in the genome.

These findings were further supported by two subsequent publications, one mapping the binding sites of Sp1, Myc and p53 using Affymetrix Chromosome 21 and 22 tiling microarrays (10) and the other mapping CREB binding on the previously mentioned Chromosome 22 array (20). Both of these groups report similar chromosome-wide binding distributions for all the factors as observed for p65. Taken together, these studies suggest that the complexity of global transcription factor binding and subsequent contribution to gene regulation is perhaps underappreciated and can only be elucidated in the context of the whole-genome tiling arrays.

13.6 Conclusion

It is now possible to map transcription factor binding sites across entire genomes of organisms with small genomes and many segments of the human genome. With higher density arrays becoming available through arrays constructed using photolithography (Affymetrix and Nimblegen) it will likely be possible to carry out whole human genome analysis of transcription factor binding using ChIP-chip. The use of these arrays for the analysis of the 1500–2000 mammalian transcription factors in many cell types will provide a transcriptional circuitry for mammalian development and cell function.

References

1. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. 1998 Dec; **9(12)**: 3273–97.
2. Kirmizis A and Farnham PJ (2004) Genomic approaches that aid in the identification of transcription factor target genes. *Exp Biol Med* **229**: 705–721.
3. Solomon MJ, Larsen PL and Varshavsky A (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**: 937–947.
4. Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25**: 99–104.
5. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M and Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533–538.
6. Horak CE and Snyder M (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol* **350**: 469–483.
7. Weinmann AS, Yan PS, Oberley MJ, Huang TH, and Farnham PJ (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**: 235–244.
8. Wells J and Farnham PJ (2002) Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation. *Methods* **26**: 48–56.
9. Martone R, Euskirchen G, Bertone P, *et al*. (2003) Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci USA* **100**: 12247–12252.
10. Cawley S, Bekiranov S, Ng HH, Kapranov P, *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**: 499–509.
11. Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM and Snyder M (2002) GATA–1 binding sites mapped in the beta-globin locus by using mammalian ChIP-chip analysis. *Proc Natl Acad Sci USA* **99**: 2924–2929.
12. Weinmann AS and Farnham PJ (2002) Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* **26**: 37–47.
13. Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* **32(Suppl)**: 496–501.
14. Ren B, Robert F, Wyrick JJ, *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.
15. Lee TI, Rinaldi NJ, Robert F, *et al*. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*. 2002 Oct 25; **298(5594)**: 799–804.

16. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M and Snyder M (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev* **16**: 3017–3033.
17. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA and Dynlacht BD (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* **16**: 245–256.
18. Ioshikhes IP and Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. Nat. Genet. 26: 61–63.
19. Rinn JL, Euskirchen G, Bertone P, *et al*. (2003) The transcriptional activity of human chromosome 22. *Genes Dev* **17**: 529–540.
20. Euskirchen G, Royce TE, Bertone P, *et al*. (2004) CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* **24**: 3804–3814.
21. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*. 2003 May 2; **113(3)**: 395–404.
22. Harbison CT, Gordon DB, Lee TI, *et al*. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004 Sep 2; **431(7004)**: 99–104.
23. Wade JT, Hall DB, Struhl K. The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. *Nature*. 2004 Dec 23; **432(7020)**: 1054–8.
24. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*. 2004 Aug; **36(8)**: 900–5. Epub 2004 Jul 11.
25. Kim J, Iyer VR. Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol Cell Biol*. 2004 Sep; **24(18)**: 8104–12.
26. MacAlpine DM, Rodriguez HK, Bell SP. Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev*. 2004 Dec 15; **18(24)**: 3094–105.
27. Wells J, Yan PS, Cechvala M, Huang T, Farnham PJ. Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sites during S phase. *Oncogene*. 2003 Mar 13; **22(10)**: 1445–60.
28. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA. Control of pancreas and liver gene expression by HNF transcription factors. *Science*. 2004 Feb 27; **303(5662)**: 1378–81.
29. Jin VX, Leu YW, Liyanarachchi S, Sun H, Fan M, Nephew KP, Huang TH, Davuluri RV. Identifying estrogen receptor alpha target genes using integrated computational genomics and chromatin immunoprecipitation microarray. *Nucleic Acids Res*. 2004 Dec 17; **32(22)**: 6627–35. Print 2004.
30. Kirmizis A, Bartley SM, Kuzmichev A, Margueron R, Reinberg D, Green R, Farnham PJ. Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev*. 2004 Jul 1; **18(13)**: 1592–605.
31 Cawley S, Bekiranov S, Ng HH, Kapranov P, *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**: 499–509.

ChIP-on-chip: searching for novel transcription factor targets

14

Esteban Ballestar and Manel Esteller

14.1 Introduction

Chromatin immunoprecipitation (ChIP) is currently the most powerful technique for investigating *in vivo* interactions between a nuclear factor and its genomic target sequences (1, 2). The technique consists of immunoprecipitating chromatin with specific antibodies to isolate DNA sequences that are bound by the nuclear proteins against which the antibodies are raised. After that, immunoprecipitated DNA is typically analyzed by PCR with specific primers to investigate the presence of a candidate DNA sequence. In practice, two different aspects can be explored. One is the binding of different nuclear factors to their binding sites (3, 4) (*Figure 14.1*). The other is that, since histones are associated with DNA throughout the entire genome, it is possible to explore the association of different post-translational modifications of histones with specific genomic sequences by using antibodies that recognize these modifications (5, 6) (*Figure 14.1*). Consequently, ChIPs provide dynamic information about not only nuclear factor occupancy at their target binding sites but also specific histone modification patterns in selected DNA sequences.

Microarrays provide an excellent platform for investigating changes on a genomic scale. The first microarrays to be designed and used were cDNA microarrays, which have been routinely used to characterize variations in gene expression (7, 8). More recently, genomic microarrays have become available as the entire genome has been sequenced and the gene regulatory regions have become better known. One potential application is the use of comparative genomic hybridization (CGH) to investigate DNA copy-number imbalances in cancer at high resolution (9). The development of novel types of genomic microarray also provides an exceptional opportunity for a new application: hybridization of ChIP samples on a microarray (ChIP-on-chip). With this elegant combination of techniques it is now possible to uncover novel binding target sequences for nuclear factors or DNA sequences with specific histone-modification patterns (10, 11) on a genomic scale.

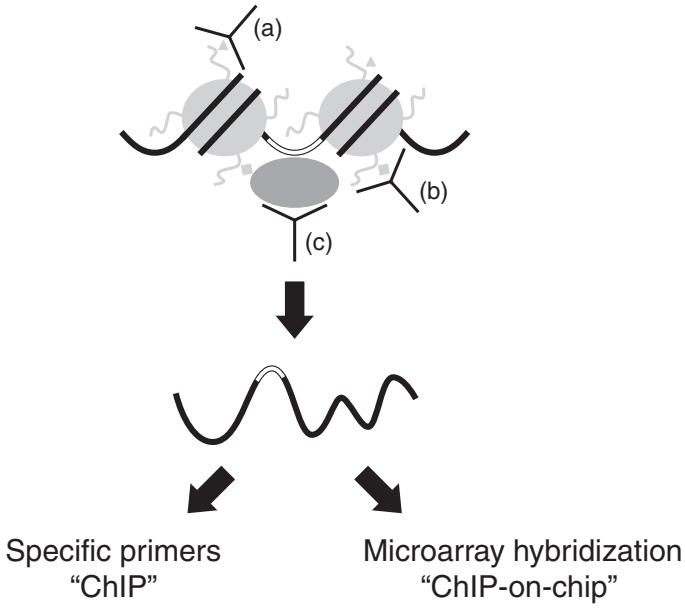


Figure 14.1.

Diagram showing the different aspects that can be investigated with ChIP assays. A nucleosome is represented as a circle, histone tails appear as protruding curly lines and DNA is represented as a black line. Antibodies (represented as three-way black lines) can be directed against different histone modifications (a, b) or against a variety of nuclear factors (c). Immunoprecipitated DNA can be analyzed either by studying candidate gene sequences, for instance by PCR amplification with specific primers, or by hybridizing appropriate microarrays (ChIP-on-chip).

14.2 Genomic microarrays

A number of genomic microarray platforms have become available in recent years. It is important to distinguish between microarrays that have a spotted selection of genomic sequences and those with a broad representation of the entire genome. Several microarrays have been designed for CGH measurements. In this case, large-insert genomic clones, such as bacterial artificial chromosomes (BACs), are used for array spots (10). Although this type of microarray can be used in the ChIP-on-chip technique, the large size of the BAC clones makes it difficult to identify the target sequence of the nuclear factor. Once a positive spot has been identified, additional studies would be necessary to map the target sequence at a higher resolution within the BAC clone.

An interesting specialized genomic microarray, designed by Tim Huang (11), consists of a library of CpG island clones. This microarray has been used in combination with a method known as differential methylation hybridization. Linker-ligated genomic DNA is digested with a methylation-sensitive restriction enzyme, amplified by PCR, and hybridized to the array. Many CpG islands become methylated in cancer and are thereby protected from methylation-sensitive restriction cleavage and so can be amplified by

PCR, producing array-hybridization signals (12, 13). Since CpG islands generally coincide with the promoter of many genes, a CpG island microarray can be useful for investigating the binding sites at the regulatory regions of CpG-island-containing genes (14). We have recently used Tim Huang's CpG-island microarray to reveal novel targets of methyl-CpG binding domain (MBD) proteins in cancer (15).

Finally, there are several promoter-based microarrays. Although this type of microarray is obviously of great interest for its potential use for studying the binding of factors to regulatory regions, it must be remembered that the entire human genome is not represented on the microarray.

14.3 Performing a successful ChIP assay

In order to guarantee the success of a ChIP-on-chip experiment it is important to optimize the conditions for the equivalent single-ChIP experiment. Many protocols describing ChIP assays (see a schematic diagram in *Figure 14.2*) have been published and are now easily accessible.

There are two major considerations when setting up an experiment: the proper fixation of DNA-protein contacts, and the fragmentation of chromatin. It is also very important to ensure that the antibody is highly specific and able to immunoprecipitate.

The most common crosslinking agent used in ChIP analysis is formaldehyde, a dipolar reagent that produces both protein-nucleic acid and protein-protein crosslinks, through the imino group of amino acids, such as Lys, Arg and His, and DNA (adenines and cytosines). A key property of

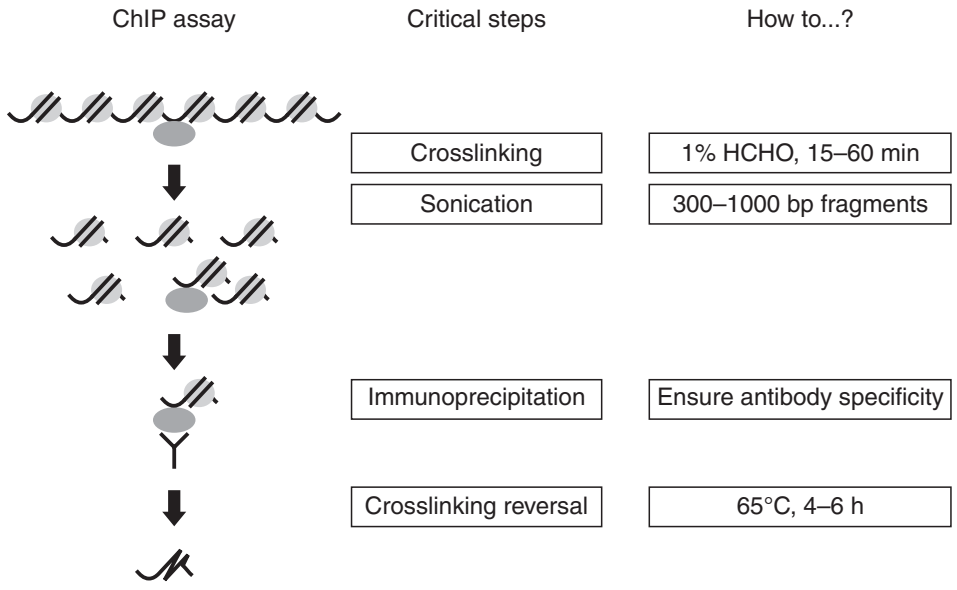


Figure 14.2.

Schematic representation of the ChIP assays. The central column shows the critical steps in the ChIP assay, as discussed in the text, and several technical tips are indicated in the right column.

the crosslinks obtained by using formaldehyde is their reversibility, which is achieved by treatment at low pH in aqueous solution or incubation at 60–70°C in the presence of SDS. Due to the small size of formaldehyde (2 Å) only proteins located within this distance of the DNA will become crosslinked. Some of the chromatin-modifying enzymes, such as histone deacetylase, do not directly bind DNA and their gene-specific regulatory functions occur through recruitment by additional DNA-binding proteins that associate regulatory sequences. Although these proteins exhibit no DNA-binding properties, it is possible to investigate their association with particular sequences by using additional protein-protein crosslinkers (16). For instance, dimethyl adipimidate (DMA) has been used to investigate the association with the yeast HDAC Rpd3 (17).

Efficient fixation of proteins to DNA is crucial for the ChIP assay. Standard conditions for formaldehyde crosslinking usually consist of a concentration of 1% and incubation times between 15 min and 1 h, depending on the proteins to be analyzed. It is important to avoid a long formaldehyde crosslinking treatment as this increases resistance to fragmentation by sonication and decreases the efficiency of the technique. Moreover, formaldehyde is a moderately denaturing agent for proteins and a high concentration or long exposure to this reagent may result in the loss of antigen epitopes. It is advisable to determine empirically the effects of formaldehyde on the protein under study. After standard fixing conditions for different exposure times, immunolocalization analysis can detect loss of fluorescence signal due to denaturation.

When choosing fixation conditions, it is important to ensure that the increased mechanical resistance of chromatin still allows fragmentation by sonication. In fact, the size of the chromatin fragments is the second critical consideration when performing ChIP assays, since these will determine both the yield of immunoprecipitated material and the degree of resolution of the technique. Chromatin fragmentation is generally achieved by sonication (although micrococcal nuclease can also be used in protocols that avoid fixation by formaldehyde) and conditions must be optimized for each sonicator prior to any immunoprecipitation experiment.

In many studies, accurate mapping can be achieved by designing primers that amplify DNA fragments of 200–300 bp. Large chromatin fragments are specifically immunoprecipitated less efficiently than small fragments. Nevertheless, the size of the fragments determines the resolution of the technique and, therefore, fragments should not greatly exceed the size of the sequence to be analyzed. If the average chromatin fragments are much larger than the sequence to be PCR-amplified or probed, one cannot be sure that the protein for which the antibody was used is actually bound to that particular region or to a neighboring region.

Finally, the quality of the antibody is extremely important in ChIP assays. It is essential to ensure, firstly, that the antibody efficiently recognizes the antigen and, secondly, that most of the immunoprecipitated material represents specific DNA sequences. Ideally, a ‘no-antibody’ control and pre-immune serum control should be included.

14.4 Obtaining material for hybridization

One important consideration in ChIP-on-chip experiments is the amount of immunoprecipitated DNA required for the hybridization. A standard ChIP DNA sample contains a variable amount of DNA of between 50 and several hundred nanograms. In a ChIP-on-chip experiment, between 1 and 2 μg of immunoprecipitated material is required for a single hybridization experiment. Two approaches have been taken to overcome this limitation and obtain the required amount. Firstly, it is possible to scale up the ChIP experiment, which generally means increasing the amounts of cells and of antibody. However, in some cases, both these quantities are limited, and an alternative approach is possible that exploits random PCR amplification of the immunoprecipitated material.

In the first case, it is best to perform multiple single-standard ChIP assays rather than amplifying the volume in a single experiment. Usually, 30 single IP experiments should yield enough material for hybridization. Samples should be treated and processed separately, and only after DNA samples have been resuspended in water should they be combined to proceed with fluorescent labeling and hybridization.

When the amount of biological material or the availability of the antibody is limited, it is possible to use an amplification step. This protocol has been described by Kuukasjarvi *et al.* (19) and modified by Huang *et al.* (20). Basically, two consecutive amplification steps are performed. The first requires the use of Thermosequenase, a degenerate primer (5′-CCG ACT CGA GNN NNN NAT GTG G-3′) and low-stringency amplification conditions (3 min at 94°C, followed by four cycles of 1 min at 94°C, 1 min at 25°C, 3 min transition at 25–74°C, 2 min extension at 74°C, and a final extension of 10 min). The second step consists of a more standard PCR amplification, standard Taq polymerase and more stringent conditions are used (3 min at 94°C, followed by 35 cycles of 1 min at 94°C, 1 min at 56°C, 2 min extension at 72°C, and a final extension of 10 min). It is important to run a confirmatory gel. There should be a smear of DNA, of length between 300 and 1000 bp, present for the antibody-treated samples. Negative controls should be added for each DOP-PCR step in order to rule out the existence of non-specific amplification of contaminant DNA.

Before labeling and hybridizing the ChIP samples, it is advisable to test a small aliquot of the samples for PCR amplification of a positive and negative control for both antibody-treated and no-antibody samples. This test depends on the availability of known *in vivo* binding targets for the protein of interest.

14.5 Labeling and hybridizing the DNA

Once the required amount of DNA (1–2 μg) has been obtained for both the antibody ChIP sample and the no-antibody control, it is labeled with the fluorescent Cy5 and Cy3 dyes. There are several commercially available DNA labeling systems for incorporating these dyes into the DNA samples. Once the labeled samples have been obtained, DNAs are cohybridized to the selected microarray. Following hybridization, the arrays are washed, scanned and analyzed like other types of microarray. Many institutions

have established core-facility units specialized in microarray hybridization and it is advisable to use their expertise during the analysis.

14.6 Validating ChIP-on-chip results

A key step when using any type of microarray is the independent validation of the results. If RT-PCR is used to validate the results of an expression-microarray analysis, in the case of ChIP-on-chip experiments, individual single-ChIP assays should be performed to confirm the target sequences identified by this technique. It would be ideal to perform ChIPs with two different antibodies raised against the same protein. Specialized validating experiments are advisable. For instance, when we performed ChIP-on-chip analysis to investigate MBD targets in breast cancer cells (15), we validated the results by using both individual ChIP assays and a specific assay. In this case, since MBDs are known to associate specifically with methylated DNA (21), we investigated the methylation status of the CpG islands that each of the anti-MBD antibodies had been able to isolate. The specific methylation profile of each of the identified targets was an independent test that served not only to validate the results from the ChIP-on-chip analysis but also to reveal novel targets of epigenetic inactivation in human breast cancer. In the same system, for instance, when studying genes for which only one allele is methylated, an appropriate validating method is the coupling of individual ChIPs with bisulfite genomic sequencing (22). For nuclear factors that have a known or inferred binding site, it would be useful to search for that particular binding site in the positive clones resulting from the ChIP-on-chip experiment. Additionally, electrophoretic mobility-shift experiments can be used to test *in vitro* the ability to bind the resulting targets (23).

14.7 Summary

ChIP-on-chip is a powerful tool that can be used to discover novel target sequences for transcription factors or to reveal DNA sequences with particular chromatin features. This potential is the result of the elegant combination of ChIP assays with microarray technology. ChIP assays allow the isolation of a genomic library of sequences that are bound by a specific factor or that contain specific histone modifications. Microarray technology makes it possible to analyze thousands of sequences in a single experiment. The applicability of this technique relies on the availability of genomic microarrays, but fortunately, both genome-wide and specialized microarrays are now available. Combination of ChIP-on-chip with other types of microarrays, such as cDNA microarrays, will surely help to lead to a functional understanding of the way by which the genome is regulated. ChIP-on-chip experiments will greatly contribute to the mapping of the epigenomic landscape.

References

1. Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation (Review). *Trends Biochem Sci* **25**: 99–104.

2. Kuo MH and Allis CD (1999) In vivo cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment. *Methods* **19**: 425–433.
3. Mencia M, Moqtaderi Z, Geisberg JV, Kuras L and Struhl K (2002) Activator-specific recruitment of TFIID and regulation of ribosomal protein genes in yeast. *Mol Cell* **9**: 823–833.
4. Crighton D, Woiwode A, Zhang C, Mandavia N, Morton JP, Warnock LJ, Milner J, White RJ and Johnson DL (2003) p53 represses RNA polymerase III transcription by targeting TBP and inhibiting promoter occupancy by TFIIIB. *EMBO J* **22**: 2810–2820.
5. Vettese-Dadey M, Grant PA, Hebbes TR, Crane-Robinson C, Allis CD and Workman JL (1996) Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. *EMBO J* **15**: 2508–2518.
6. Peinado H, Ballestar E, Esteller M and Cano A (2004) The transcription factor Snail mediates E-cadherin repression by the recruitment of the Sin3A/Histone Deacetylase 1 (HDAC1)/HDAC2 complex. *Mol Cell Biol* **24**: 306
7. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE and Davis RW (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* **94**: 2150–2155.
8. Iyer VR, Eisen MB, Ross DT, *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* **283**: 83–87.
9. Pinkel D, Segraves R, Sudar D, *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**: 207–211.
10. Ishkanian AS, Malloff CA, Watson SK, *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet* **36**: 299–303.
11. Yan PS, Efferth T, Chen HL, Lin J, Rodel F, Fuzesi L and Huang TH (2002) Use of CpG island microarrays to identify colorectal tumors with a high degree of concurrent methylation. *Methods* **27**: 162–169.
12. Huang TH, Perry MR and Laux DE (1999) Methylation profiling of CpG islands in human breast cancer cells. *Hum Mol Genet.* **8**: 459–470.
13. Paz MF, Wei S, Cigudosa JC, Rodriguez-Perales S, Peinado MA, Huang TH and Esteller M (2003) Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. *Hum Mol Genet* **12**: 2209–2219.
14. Weinmann AS, Yan PS, Oberley MJ, Huang TH and Farnham PJ (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**: 235–244.
15. Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J, Cigudosa JC, Huang TH and Esteller M (2003) Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. *EMBO J* **22**: 6335–6345.
16. Kurdistani SK and Grunstein M (2003) In vivo protein-protein and protein-DNA crosslinking for genomewide binding microarray. *Methods* **31**: 90.
17. Kurdistani SK, Robyr D, Tavazoie S and Grunstein M (2002) Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nat Genet* **31**: 248.
18. Spencer VA, Sun JM, Li L and Davie JR (2003) Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. *Methods* **31**: 67.
19. Kuukasjarvi T, Tanner M, Pennanen S, Karhu R, Visakorpi T and Isola J (1997) Optimizing DOP-PCR for universal amplification of small DNA samples in comparative genomic hybridization. *Genes Chromosomes Cancer* **18**: 94–101.
20. Huang Q, Schantz SP, Rao PH, Mo J, McCormick SA and Chaganti RS (2000)

Improving degenerate oligonucleotide primed PCR-comparative genomic hybridization for analysis of DNA copy number changes in tumors. *Genes Chromosomes Cancer* **28**: 395–403.

21. Fraga MF, Ballestar E, Montoya G, Taysavang P, Wade PA and Esteller M (2003) The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res* **31**: 1765–1774.
22. Matarazzo MR, Lembo F, Angrisano T, *et al*. (2004) *In vivo* analysis of DNA methylation patterns recognized by specific proteins: coupling chromatin immunoprecipitation and bisulfite genomic sequencing (ChIP-BA). *Biotechniques* **37**: 666–673.
23. Fraga MF, Ballestar E and Esteller M (2003) Capillary electrophoresis-based method to quantitate DNA-protein interactions. *J Chromatogr B Analyt Technol Biomed Life Sci* **789**: 431–435.
24. Snijders AM, Nowak N, Segraves R, *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29**: 263–264.
25. Carvalho B, Ouwerkerk E, Meijer GA and Ylstra B (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol* **57**: 644–646.
26. Yan PS, Chen C-M, Shi H, Rahmatpanah F, Wei SH, Caldwell CW and Huang TH-M (2001) Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res* **61**: 8375–8380
27. Kirmizis A, Bartley SM, Kuzmichev A, Margueron R, Reinberg D, Green R and Farnham PJ (2004) Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev* **18**: 1592–1605.

Protocol

CONTENTS

Protocol 14.1: Performing a successful ChIP assay

We have obtained the best results by using the following protocol, which is based on that described by *Upstate Group, Inc* and Spencer *et al.* (18):

1. Stimulate or treat cells as appropriate. Cells should be treated under conditions for which transcriptional activation of the gene of interest has been demonstrated. Use 1×10^6 cells for each antibody.
2. Crosslink histones and other nuclear factors to DNA by adding formaldehyde directly to culture medium to a final concentration of 1%. Incubate for 15–60 min (as previously determined) at room temperature. Preliminary experiments to estimate the best combination of crosslinking time and fragmentation should be performed. When planning to store crosslinked cells, glycine should be added to a final concentration of 0.125 M and incubated for 5 min. After that, 1 × phosphate-buffered saline (PBS) washes should be performed as described below.
3. Aspirate medium, removing as much of it as possible. Wash cells twice using ice-cold 1 × PBS containing protease inhibitors (there are several cocktails of protease inhibitors commercially available that cover a wide spectrum of inhibition).
4. Scrape cells and transfer to a conical tube. For suspension cells, the PBS washes need to be performed in the tube.
5. Pellet cells for 4 min at 2000 *g* at 4°C. Warm cell lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.1) to room temperature to dissolve precipitated SDS and add protease inhibitors.
6. Resuspend cell pellet in cell lysis buffer and incubate for 10 min on ice. Each 1×10^6 cells should be resuspended in 200 µl of cell lysis buffer.
7. Sonicate the cell lysate to shear DNA to lengths between 200 and 1000 bp, being sure to keep samples ice-cold. When optimizing sonication conditions, at this point, 20 µl of 5 M NaCl are added to each 500 µl and incubation at 65°C for 4 h is performed to reverse crosslinks. This incubation is followed by phenol/chloroform extraction and samples are analyzed in agarose gels to visualize shearing efficiency.

8. Once sonication conditions have been optimized, centrifuge samples following sonication for 10 min at 13 000 *g* at 4°C. The size of a sample will be the equivalent amount of 200 μl of the sonicated cell lysate.
9. Dilute the sonicated cell lysate 10-fold in ChIP dilution buffer (0.01% SDS, 1.1% Triton-X–100, 1.2 mM EDTA, 16.7 mM Tris-HCl, pH 8.1, 167 mM NaCl), adding protease inhibitors, as above. This is done by adding 1800 μl of ChIP dilution buffer to each 200 μl of sonicated cell lysate to give a final volume of 2 ml for each immunoprecipitation condition. If proceeding to PCR a portion of the diluted cell pellet suspension, a small volume can be kept to quantitate the amount of DNA present in different samples for the PCR protocol. This sample is considered to be the input/starting material, and needs to be heated at 65°C for 4 h in order to reverse crosslinks.
10. To reduce nonspecific background, it is advisable to treat the diluted cell lysate with 80 μl of protein A/G-agarose/salmon sperm DNA (50% gel slurry in TE buffer, containing 15 μg of sonicated salmon sperm DNA; Upstate Group, VA) for 30 min at 4°C with agitation.
11. Pellet agarose beads by brief centrifugation and collect the supernatant fraction.
12. Add the immunoprecipitating antibody (the amount will vary per antibody) to the 2 ml of supernatant fraction and incubate overnight at 4°C with rotation. For a negative control, perform a no-antibody immunoprecipitation and a pre-immune serum precipitation (when available).
13. Add 60 μl of protein A/G-agarose/salmon sperm DNA beads for 1 h at 4°C with rotation to collect the antibody-protein complex.
14. Pellet agarose by gentle centrifugation (2000 *g* at 4°C for 1 min). Carefully remove the supernatant and keep this fraction, which consists of unbound DNA. Wash the protein A/G agarose beads for 5 min on a rotating wheel with 1 ml of each of the following solutions:
    - Low-salt wash buffer (0.1% SDS, 1% Triton X–100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.1, 150 mM NaCl)
    - High-salt wash buffer (0.1% SDS, 1% Triton X–100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.1, 500 mM NaCl)
    - LiCl wash buffer (0.25 M LiCl, 1% NP40, 1% deoxycholate, 1 mM EDTA, 10 mM Tris-HCl, pH 8.1)
    - 1 × TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0)

Repeat.

15. After the last washing, protein-DNA complexes can be eluted from the antibody by adding 250 µl of freshly made elution buffer (1% SDS, 0.1 M $NaHCO_3$) to the pelleted protein A agarose-antibody-protein-DNA complex. Vortex briefly to mix and incubate at room temperature for 15 min with rotation. Pellet the agarose beads, and carefully transfer the supernatant fraction (eluate) to another tube and repeat elution. Combine eluates. The total volume should be approximately 500 µl.

16. Add 20 µl of 5 M NaCl to the combined eluates and reverse protein-DNA crosslinks by heating at 65°C for 4 h. The input sample as well as the unbound samples should also be treated in a similar manner by adding the equivalent volume of 5 M NaCl.

17. Add 10 µl of 0.5 M EDTA, 20 µl of 1 M Tris-HCl, pH 6.5 and 2 µl of 10 mg/ml proteinase K to the combined eluates and incubate for 1 h at 45°C.

18. After protein removal, add an equal volume of phenol:chloroform:isoamyl alcohol (24:23:1) to the sample and mix, then centrifuge at 12 000 *g* for 1 min. The upper aqueous phase is transferred to a new microcentrifuge tube and phenol:chloroform extraction is repeated. The DNA is then precipitated by adding one-tenth of the volume of sodium acetate to the aqueous phase, a carrier, such as glycogen or yeast tRNA, and 2.5 volumes of absolute ethanol, and then incubating the sample at –80°C for at least 1 h. After that, wash the DNA pellet with 70% ethanol and air dry.

19. Resuspend each pellet in 20 µl of water for PCR analysis. In standard ChIPs, the input sample is used at different dilutions to establish conditions for which specific PCR products are obtained below saturation.

Turning photons into results: principles of fluorescent microarray scanning

15

Siobhan Pickett and Damian Verdnik

15.1 Introduction

The microarray image is your window into the biology you want to measure. The clearer that window is, the easier it will be to identify, interpret, and understand the phenomena you observe. Accurate imaging, including scanning and image processing, ensures that the data reported by the instrument provide the most accurate representation of the actual fluorescent signal on the array.

15.2 Scanning parameters

Microarray scanner evolution

Microarray scanner technology evolved from scanning fluorescence microscopes that were used to image subcellular components. These microscopes combined high-powered excitation light sources, user-selectable emission filters, and a detection system to collect and store digital images of the fluorescent sample. Laser-based systems use motion control elements to scan the laser beam across the sample and photomultiplier tubes (PMTs) to collect emitted light one pixel at a time. White-light-based systems use mercury or xenon arc lamps to excite an entire field of view, and a charged-coupled device (CCD) to capture emitted light from the entire field of view. Both types of systems are used in microarray imagers today.

However, microarrays have different imaging requirements than cells and organelles. Microarray spots usually range from about 50 to 100 microns in diameter, which is extremely large compared to subcellular components. Unlike cellular imaging, the substructure of microarrays spots is of no interest (beyond optimizing spot uniformity). Microarray spots are deposited in known locations with plenty of separation between them, whereas cells and organelles are irregular in shape and location. Therefore, while cellular imaging requires ever-more powerful magnification and sub-micron pixel sizes to clearly resolve tiny structures, most

microarrays can be accurately imaged at 5- or even 10-micron pixel resolution. In addition, cellular microscopy has historically focused on the presence and location of specific elements of interest, rather than quantitation. (Note that this is changing, due in part to the development of flat-field imaging optics.) The primary purpose of microarrays is to quantify the signal from each spot. The data must be accurate and comparable over the entire 25 × 75-mm slide surface. Therefore, field uniformity is a critical parameter in microarray instrumentation.

How an image is acquired: scanner design considerations

Unlike light microscopy, which allows the viewer to look directly at the sample itself, fluorescence imaging requires a fluorescent label to be bound to the sample. When interpreting the results of microarray experiments, it's important to keep in mind that fluorescence imagers do not detect DNA, proteins, cells, or any other biological material – they only detect the fluorescent dyes that are bound to the biomolecules.

Fluorescence is the property of some molecules that absorb light of a given wavelength, and then emit light of a higher wavelength. Fluorescent dyes are characterized by their excitation and emission spectra. The excitation spectrum represents the efficiency with which the dye will absorb light over the given range of wavelengths. The emission spectrum indicates the probability that a photon of emitted light will be of a given wavelength. Thus, the peak of the excitation spectrum indicates the wavelength of light that is most efficiently absorbed by the dye, and the peak of the emission spectrum indicates the predominant wavelength of light that will be emitted as a result of the excitation. The difference between the excitation peak and the emission peak is called the Stokes' shift (1).

Not all fluorescent dyes are equally bright, even at their excitation and emission peaks. The brightness of dyes is determined by specific constants for each dye. The extinction coefficient indicates the efficiency with which the dye absorbs incident light (usually at the peak of the absorption spectrum). The quantum yield indicates the ease with which the dye molecule releases a photon of fluorescent light over the entire emission spectrum. The resulting fluorescence intensity is proportional to the product of these two constants. Therefore, equimolar amounts of two different dyes will not necessarily produce equally bright signals. Filter choices and excitation light properties (which also vary with wavelength) also contribute to differences in signal brightness.

In a fluorescence imaging system, excitation light is provided by either a halogen arc lamp or a laser. The light is delivered to the sample through a series of lenses and filters. The fluorophore on the sample emits light, which then travels through additional lenses and filters to the detector (a CCD or PMT). The analog signal from the detector is converted into a digital signal, which is then used to display an image of the sample on the computer screen.

Detailed descriptions and design considerations for white light and laser imaging systems have been discussed previously (2). This chapter focuses on critical instrument performance metrics, and some analysis methods that can be used to evaluate them.

Critical performance characteristics of a microarray fluorescence imaging system

Microarray scanners are complex instruments consisting of hundreds of individual parts, including light sources, detectors, circuit boards, moving parts, lenses, filters, wiring, and sensors. While the design and selection of each of these subunits is important, no single component defines the performance of the instrument. Ultimate instrument performance is determined by the integration of all of the subunits into a complete working system. The design of the electronic circuitry, alignment of optical elements, and behavior of moving parts all affect the data quality and long-term instrument performance. The specifications of individual components do not measure the ultimate function of combined subunits. The scanned image is the final output of the instrument; therefore the only way to characterize and compare the performance of the complete system is to compare scanned images. In addition, the appearance of the image on the screen is the result of the algorithms used to convert the analog fluorescent signal into a digital value, the color and display settings chosen, and even the monitor itself. Visual assessment of images can be misleading; therefore the signals must be quantified using appropriate background subtraction methods before valid comparisons can be drawn.

Scanner calibration

One slide does not constitute a microarray experiment. You may scan hundreds of slides over many months before reaching significant biological conclusions. Microarray scanners must perform consistently so that experimental results can be compared over time, and be validated and shared among different research groups. However, mechanical, optical, and electronic components have a finite lifetime, so instrument performance will change over time. Instrument calibration can ensure imaging reproducibility among multiple scanners of the same model over time.

Scanner calibration uses a known standard to set instrument output to predefined levels. The standard might be a precisely controlled light source or a fluorescent material. It must be stable such that it yields the same signal output after repeated long-time use. For example, GenePix® scanners are benchmarked at the factory to produce a specified signal output from a stable standard using defined scan settings. The test standards are a set of fluorescent materials that absorb and emit light consistent with each of the excitation and emission channels in the instrument. The standard is shipped with the instrument so that the user can invoke the calibration routine as often as they want to re-tune the instrument to the benchmark levels. Multiple instruments can be adjusted in the same way to produce the same benchmark signal levels.

Detection limit

A brighter image is not necessarily a better image. Absolute pixel intensity values will vary depending on different types and brands of detectors, variations in electronic signal processing, analog-to-digital conversion algo-

rithms, and other design differences. Color and display settings and even monitor settings can also influence the appearance of images scanned on different systems. Visual inspection is non-quantifiable and highly subjective, and is not a reliable method for detection limit comparisons on different instruments. Signal intensity and detection limit must be quantified using appropriate calculations and background subtraction methods (see below).

A detection limit performance specification should indicate the minimum signal that the system can quantify accurately. As a signal approaches the surrounding background level, the potential error in each measurement increases. In other words, as a spot fades into the background, so does your confidence in its existence. The signal-to-noise ratio (SNR) is the most reliable detection limit metric. The SNR calculation incorporates signal, the average background level, and the variation in background values to measure how clearly the signal can be distinguished from the background. For imaging applications, SNR is calculated as:

$$\text{SNR} = \frac{(\text{Signal} - \text{Background})}{(\text{Standard Deviation of Background})}$$

On most microarray scanners, spots may be visible below this limit; however, the accuracy of the measurement begins to diminish. As an analogy, consider looking for a 2-m-tall scarecrow in a cornfield. If all the cornstalks are 1 m tall, then the average background is uniform and lower than the signal (the scarecrow), so the scarecrow is clearly visible. If all the cornstalks are 2 m tall, then the signal is the same as the average background. Although the background noise is low, the average background is high so the scarecrow is not visible. Finally, if the cornstalks range in height equally distributed from 0.5 to 3 m tall, the average background is 1.75 m. The average background is lower than the signal, but the variation in height (i.e. the noise) will make it difficult to see the scarecrow. Thus, the signal, the average background, and the background variation must all be considered when determining detection limits in imaging applications. A commonly accepted criterion in many signal detection disciplines (including radio, electronic communications, trace chemical detection, and other fields) defines the minimum quantifiable signal at threefold greater than the background noise – that is the sample value for which SNR = 3 (3, 4).

Fluorescence imaging instruments do not detect DNA or proteins – they detect fluorescent dyes that are bound to the biomolecules. The detection limit of a fluorescence imaging instrument is measured in moles of fluorophore per square micron. The true detection limit can only be determined through careful quantitation of meticulously prepared dilutions of fluorescent dyes. In addition, the concentration of active fluorophore in a batch of dye varies according to batch preparation, age, and environmental conditions. Prior to arraying, the precise dye concentrations must be quantified on an independent platform such as a spectrofluorometer. The volume of solution that adheres to the slide during arraying must also be known. There can be no post-spotting washes or other treatment that may alter the amount of active fluorophore at each spot. The array must be used imme-

diately for detection limit determination because even the slightest decay in signal may cause the faintest spots to fade below the detection limit. These tests are time-consuming and are not practical for routine instrument comparisons.

An acceptable alternative to assess SNR differences among instruments is to compare any dilution series that covers a wide range of signal values. Several slide replicates should be scanned in alternating order on different instruments to assess and compensate for any photobleaching or slight differences among the replicates. The difference in SNR for identical spots near the background level gives a simple comparison of sensitivity across instruments, without knowledge of the absolute dye concentrations.

Field uniformity

Field uniformity is one of the most important specifications for microarray imaging. Microarray analysis entails measuring thousands of tiny signals on a relatively large field, with sufficient accuracy to allow spots to be compared among all locations on the array. A uniform imaging field ensures that the instrument is not contributing to regional variations that might bias the data and interfere with accurate comparisons.

The microarray substrate and surface matrix are the primary determinants of field uniformity. Most standard microscope slides are specified to about 40-μm flatness over the entire surface; that is, they may have hills and valleys as high as 40 μm. Optically flat slides for microarrays are also available. The slide surface variations can cause quantitative variations as the imaging plane comes in and out of focus. A scanner with a larger depth of field can better accommodate slide surface variations, ensuring accurate light collection over the entire scanned area.

Instrument components such as the slide holder, motion control mechanisms, the excitation source, and the illumination and collection optics can all affect field uniformity. In any microarray imaging system, all of these components must be precisely specified and aligned to ensure uniform illumination at all points on the sample surface.

A test standard to measure field uniformity must be more uniform than the instrument in question so that the sample itself doesn't contribute additional non-uniformity to the measurement. Such a standard doesn't yet exist in a microscope slide format. However, any fluorescent microarray can be used in a simple alternative test. You can scan the array in one orientation, rotate it 180°, and scan it again. A comparison of the signal intensities for identical spots in each scan quantifies field uniformity (*Plate 6*). Consistently lower signal in the second scan might indicate photobleaching. A third scan in either orientation can be used to quantify the photobleaching and subtract its contribution from the uniformity analysis. This rotation test is the most reliable measure of field uniformity using currently available tools. However, it is limited to variations that are asymmetrical with respect to the rotation. For example, a uniform hill or valley in the middle of the slide might go undetected.

Repeatability

A single microarray experiment is insufficient to reveal meaningful biological conclusions. Experimental replicates are critical to identify and eliminate experimental error and other variations, especially when an experiment uses many microarrays over the course of many months. Any complete microarray experiment should include array replicates, sample replicates, probe replicates (e.g. dye-swaps), and hybridization replicates. It is unwise to exclude replicates; however, you can reduce the cost of replicates and retain more data for analysis by minimizing the variation among them. If you devote some time to optimizing your protocols prior to starting a major experiment, and use care at each step in the microarray process, you will be rewarded with more reproducible results.

Instrument reproducibility is also an important parameter. Lasers and white-light sources need time to stabilize after igniting. In addition, extreme temperature and humidity fluctuations can cause variations in instrument behavior. To ensure repeatable results, users must observe the recommended warm-up times and operating conditions for their instruments. The best insurance against both short-term fluctuations and long-term signal drift is a calibration procedure such as that described above.

Signal repeatability of a microarray imaging system can be tested by repeatedly scanning a stable fluorescent standard (such as that described under "Scanner calibration" above) at identical scanner settings (*Plate 7*). Any reasonably stable fluorescent sample can also be used to test short-term scan-to-scan repeatability, although any photobleaching must be measured and subtracted.

15.3 Analysis parameters

Standardization is a hot topic in microarray data analysis. Microarrays generate so much data, which may be shared among many labs worldwide, that standardized experimental and analysis methods are becoming more important to the microarray community. When the issue of standardization is raised, conversation quickly turns to MIAME (Minimum Information About a Microarray Experiment) (see also Chapter 22) (5) and MAGE-ML (Microarray Gene Expression Markup Language) (6). MIAME is a guide to the types of information that scientists should record and report when describing microarray experiments. MAGE-ML is a file format for microarray data that contains MIAME descriptors in the file. The intended advantage of MAGE-ML over other file formats is that anyone can look at a MAGE-ML file and determine how the data was analyzed.

MIAME and MAGE-ML are successful in their intent, but they are a solution to only a very small part of a very large problem. Knowing how a microarray image was analyzed does not answer the much more important question: What is the best way to analyze a microarray image? What is the best segmentation method, the best background subtraction method, the best ratio measure, or the best normalization method? Do you even need to do all of these? The microarray community does not agree on an answer to any of these questions. This is a

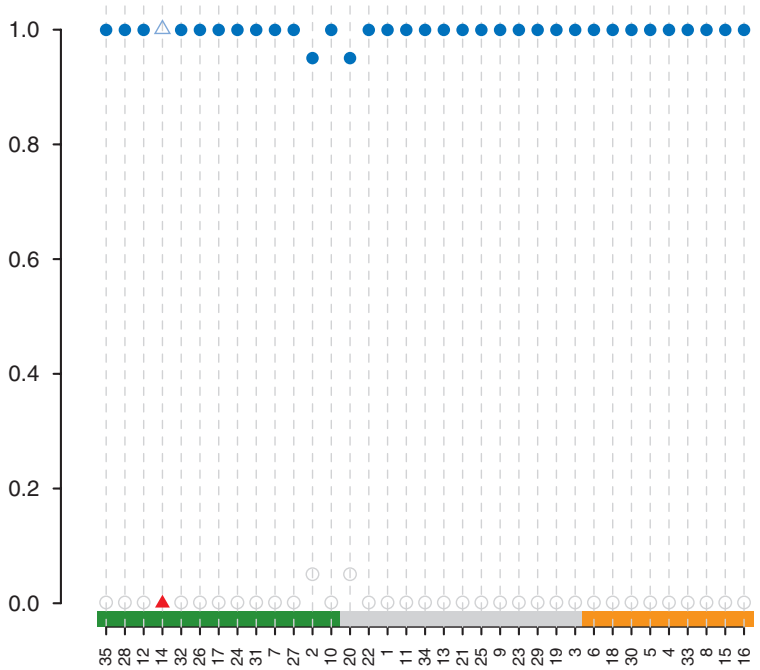


Plate 1.

Classification of renal cell carcinoma samples using prediction analysis for microarrays (PAM) (21). The frequencies of correct tumor classification (y-axis) are plotted for each tumor sample (x-axis; green: ccRCC; grey: pRCC; orange: chRCC). The sample that is incorrectly classified in all repetitions is marked with a red triangle, all others with blue circles.

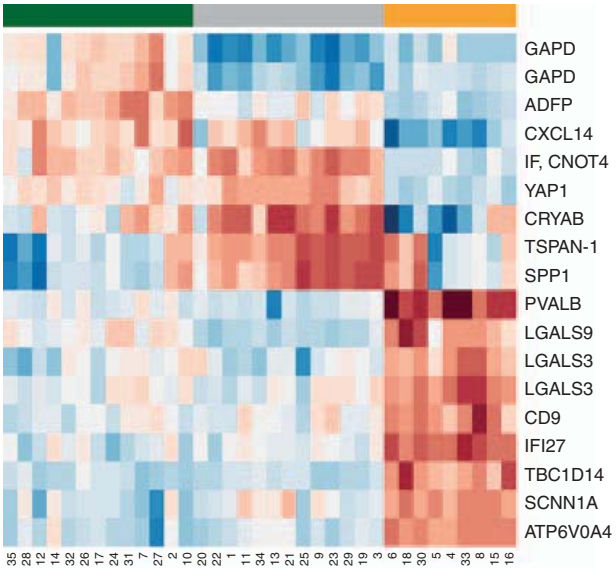


Plate 2.

Visualization of the set of 18 genes the PAM classifier is based on. The three kidney tumor types are indicated at the top (green: ccRCC; grey: pRCC; orange: chRCC). Gene designations are listed on the right hand side and the sample numbers at the bottom. Red indicates high generalized log ratios, blue denotes low log ratios.

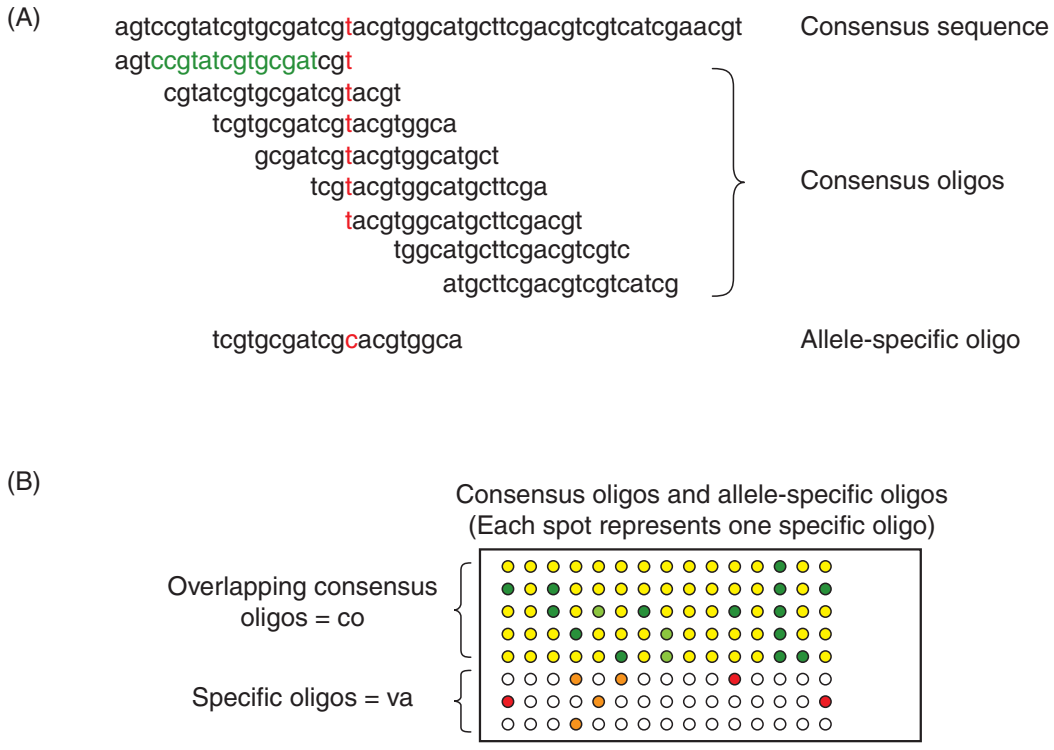


Plate 3.

(**A**) The top sequence represents the consensus sequence with the interrogating polymorphic site indicated in red. Overlapping consensus oligos are designed according to the four-nucleotide tiling system. SNPs in the flanking region (outside the green sequence of the first overlapping oligo) do not affect the hybridization pattern. At the bottom an example of an allele-specific oligo designed with the SNP positioned at the center is given. (**B**) Array printing layout. Consensus-restricted hybridization (the red labeled sample does not hybridize because it contains a SNP while the reference does hybridize) is shown in green. Balance hybridization (similar hybridization between consensus and test sample) is shown in yellow. Allele-specific hybridization (only test sample hybridizes to the allele-specific oligo while the reference sample does not) is shown in red.

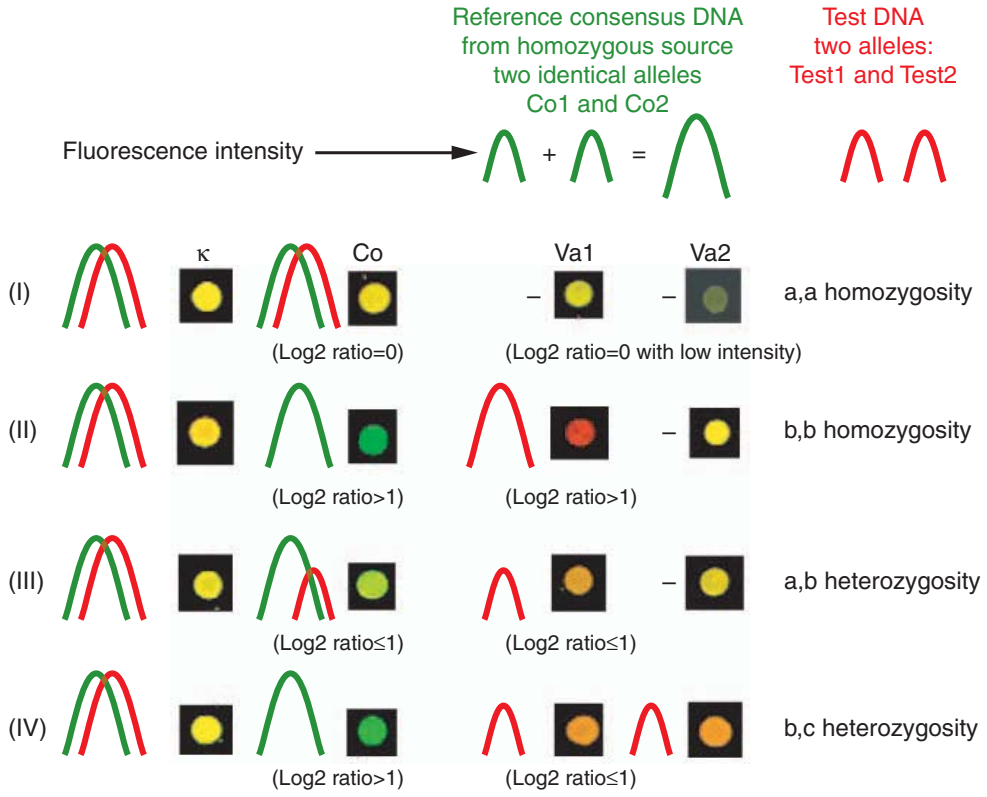


Plate 4.

Principle of SNP detection using consensus four-nucleotide tailing oligos and allele-specific oligo. Portrait of differentially labeled reference (Cy3) and test (Cy5) samples hybridized to overlapping consensus oligos (consensus oligos = Co) or variant-specific oligos (Va1 and Va2). The homozygous reference sample consists of two alleles identical to the consensus. Differentially fluorescence-labeled reference and test samples are cohybridized to an array slide spotted with the overlapping consensus oligos and variant oligos. *k* represents the most conserved region and is used to normalize the data set. Reference sample consistently hybridizes to Co and never to Va oligos. Hybridization of test sample will determine the variability in ratio of fluorescence intensity as portrayed by the digital images from a GenePix scanner. Possible combinations are: *a,a* type homozygosity (row I); *b,b* homozygosity (row II); *a,b* heterozygosity with one known allele (row III) and *b,c* heterozygosity with two known alleles (row IV).

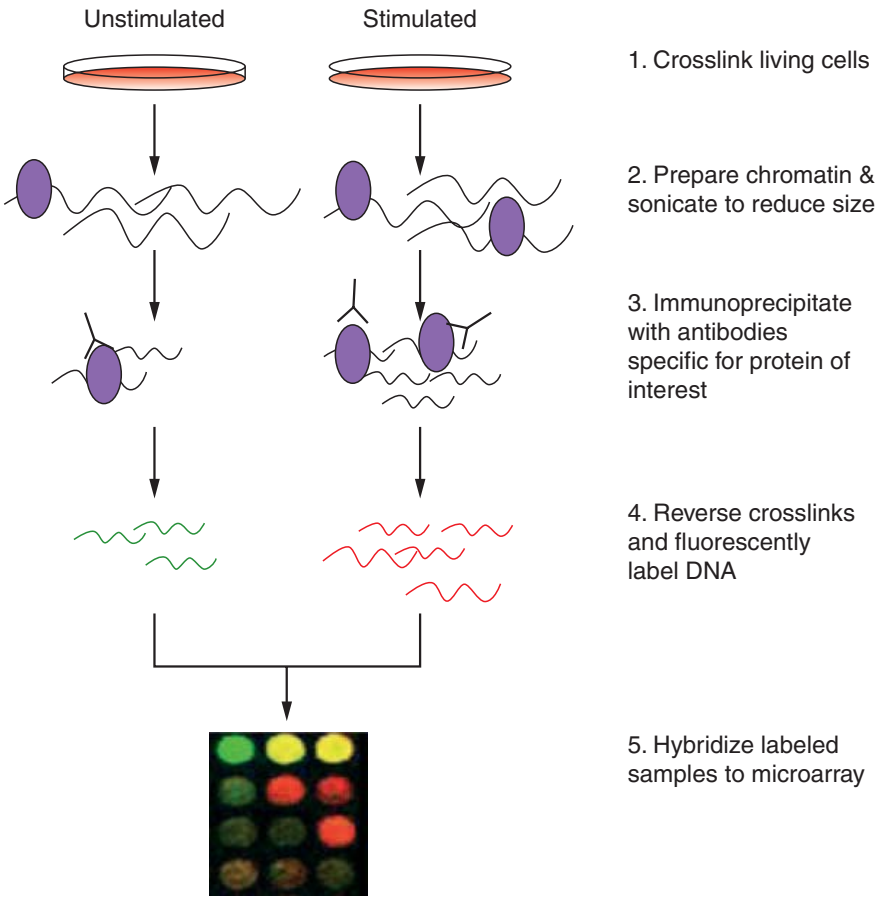
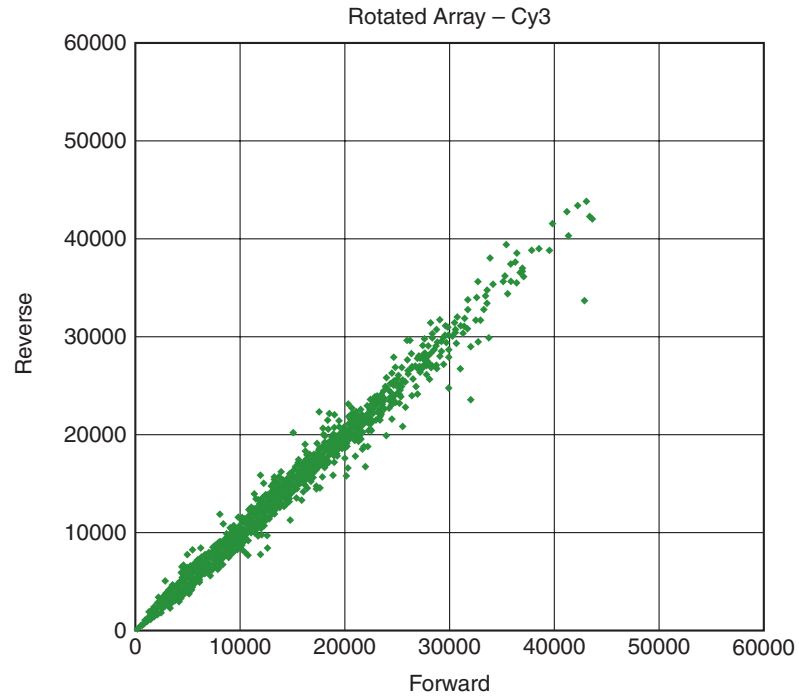


Plate 5. ChIP-chip overview.

(A)



(B)

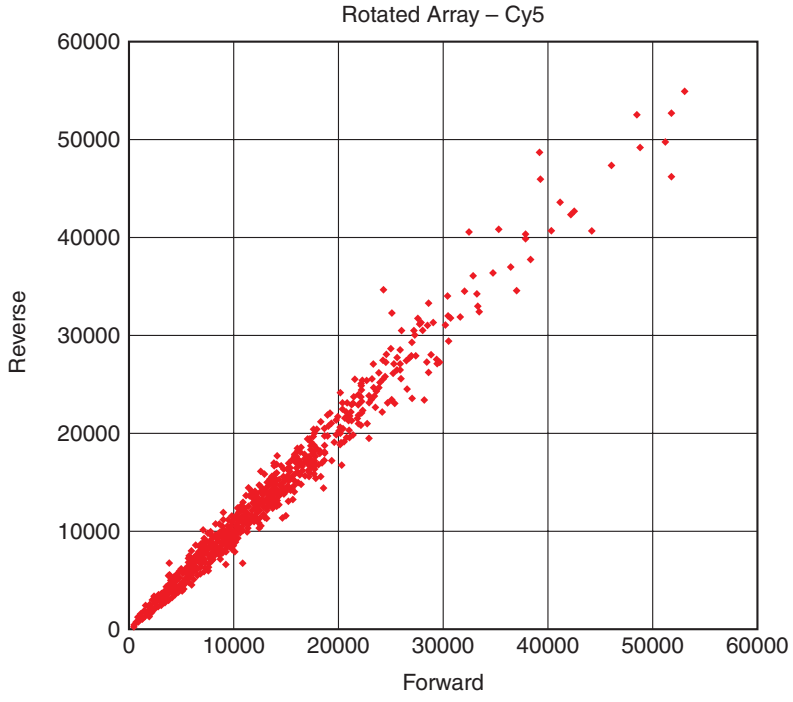


Plate 6.

Field uniformity. A hybridized gene expression microarray was scanned on a GenePix® 4000B (Molecular Devices) in the 'forward' orientation, then rotated 180° and scanned again in the 'reverse' orientation. The average difference between forward and reverse scans is 4.5% for the green channel (**A**) and 6.4% for the red channel (**B**).

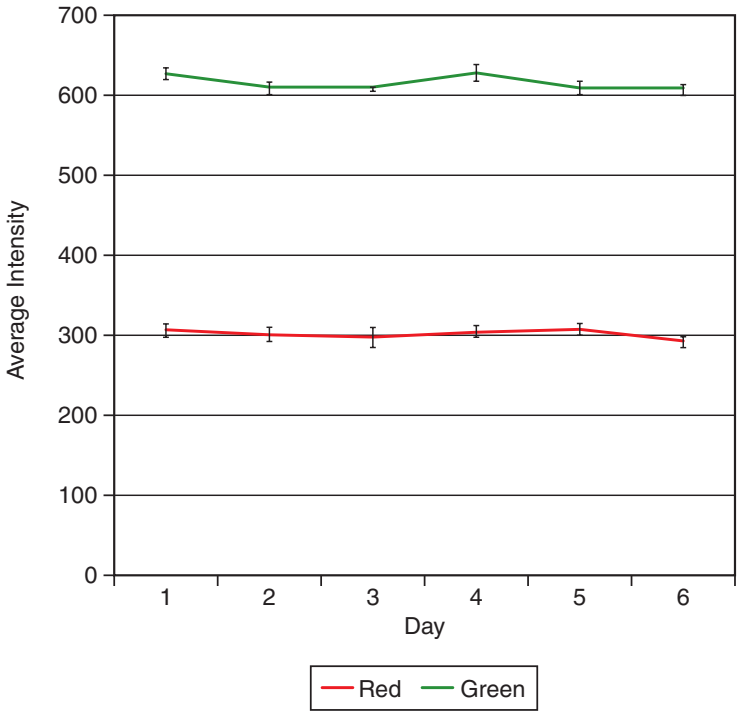


Plate 7.

Repeatability. A non-bleaching fluorescent test standard was scanned repeatedly and the average signal was quantified. Signal value variance among all scans was 1.6% in the green channel and 2.1% in the red channel. Error bars =2σ; *n* =4.

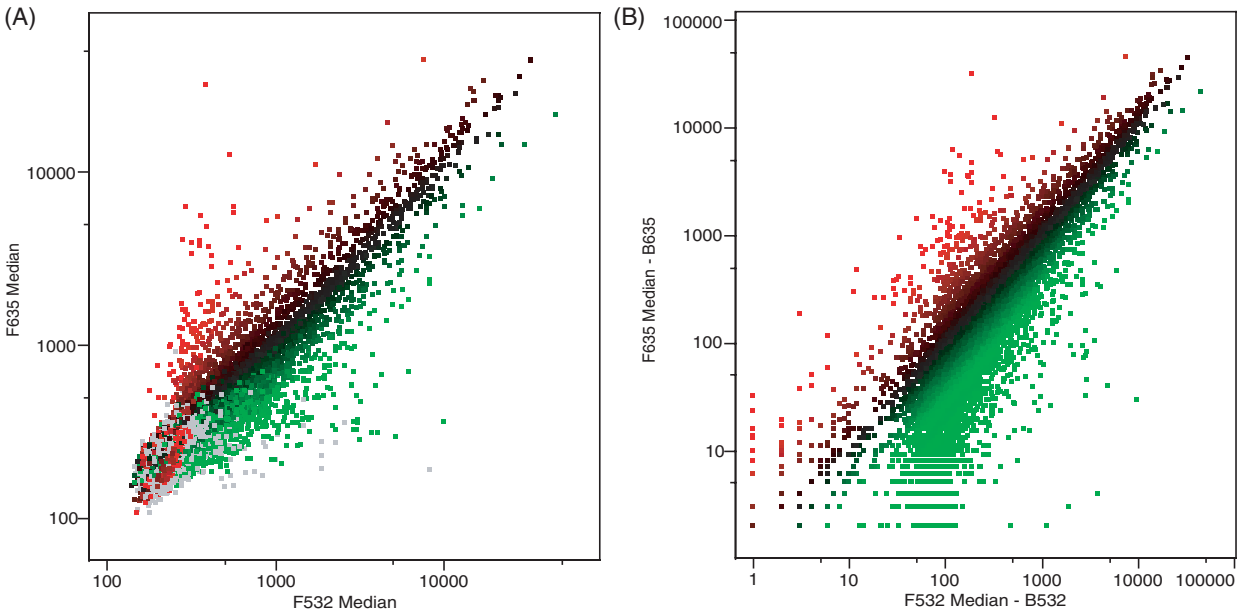


Plate 8.

Scatter plots of 635-nm versus 532-nm intensities. (**A**) No background subtracted; (**B**) local background subtracted. Data for all figures was generated using GenePix Pro and Acuity® microarray analysis software (Molecular Devices).

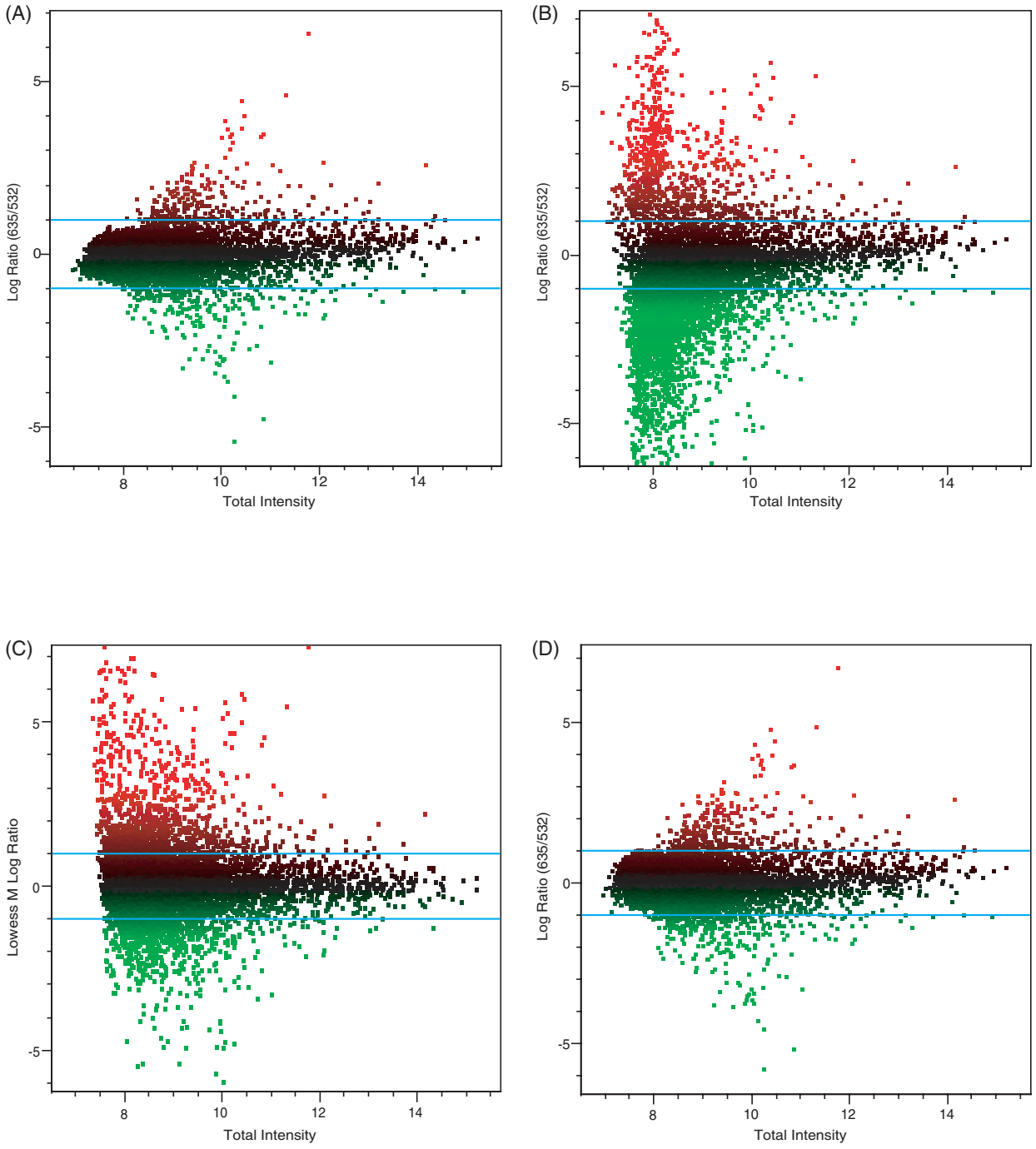


Plate 9.

MA plots (log ratio of medians versus total intensity) of the same data as in *Plate 8*. (**A**) No background subtracted; (**B**) local background subtracted; (**C**) global background subtracted; (**D**) morphological background subtracted. The horizontal lines indicate log ratios of +1 and –1.

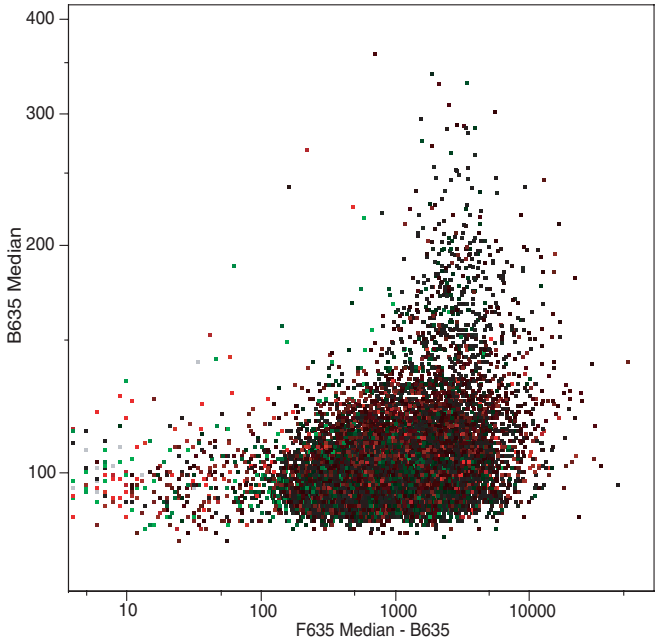


Plate 10.

Scatter plot of local-background-subtracted intensity in the red channel (x-axis) *versus* the red channel background (y-axis), showing a dependence of background on intensity. At high feature intensities, there are corresponding high background measures.

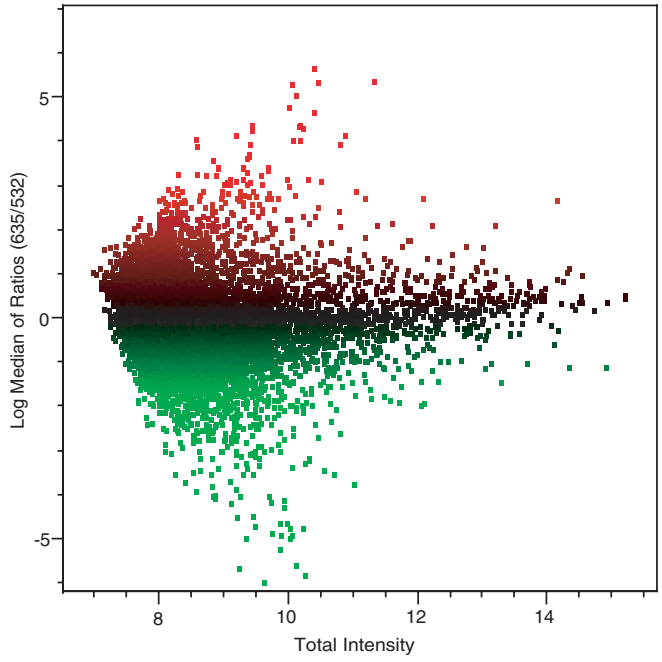


Plate 11.

MA plot of log median of ratios *versus* total intensity of the same data as in *Plate 8*.

much greater concern than agreeing on a common file format. Without data analysis standards, the existence of a common file format does not get us very far along the road to standardization.

Two outstanding microarray data analysis problems that are related to image acquisition and analysis are background subtraction and normalization. If these methods are not applied correctly, data analysis can distort the data from even the best microarray scanner.

Background subtraction

The fluorescence intensity that is measured in a feature usually includes a certain amount of stray light from various sources:

- auto-fluorescence of the slide;
- non-specific binding of labeled sample to the microarray substrate.

This stray fluorescence, known as background, needs to be accounted for in order to calculate a true measure of the fluorescence in a feature. Many methods exist for removing background from microarray images. Each method has advantages and disadvantages.

Local background subtraction methods calculate a unique background value for each feature from a region near the feature. The advantage of local methods is that they subtract only the nearby background from each feature. Local background subtraction methods are the most commonly used for microarrays because microarray features are very small relative to the entire array, and background can vary significantly across the field. However, if there are artifacts or binding variations near or within a feature, local methods may produce unrealistically high or low background values. For example, if non-specific binding is different on features compared with the space between features, then the background estimated from the space between features will not correctly estimate the background fluorescence within features.

Global background subtraction methods calculate a single value for each wavelength. The advantage of global methods is that they provide a single background estimation for the whole slide. Global methods are useful particularly when the features are so close together that local methods cannot be applied. However, if the background varies significantly across a slide, one single estimate may not accurately represent the background contribution to all features.

Negative control background subtraction methods calculate a background value from the intensity of specified negative-control features. Choose negative controls that you know do not hybridize with your sample, so that they always give the same intensity values independent of the experiment. Negative-control methods have several advantages over local and global methods. First, non-specific binding may differ where features have been printed on a slide, compared to the space between features. In such cases one can estimate non-specific fluorescent background from negative-control features, rather than from local regions between features. Second, negative-control features can be used to calculate background for nearby features, compensating for both non-specific hybridization and non-specific binding to the support matrix. However,

unlike all the other background subtraction methods, which are purely computational and can be applied to any slide, the disadvantage to using negative controls is that they must be included in the microarray slide design from the beginning. In addition, if negative controls are used as the only background method, they should be distributed widely throughout the array, and can therefore take up a lot of space on the array.

Morphological methods (7) address a problem that all of the other three methods can exhibit, namely that the background estimated for a feature can be higher than the feature intensity. This situation is rather catastrophic, as a negative background-subtracted intensity leads to a negative ratio, and hence an undefined log ratio, so that the feature is lost from all further data analysis.

In morphological methods, a copy of each single-wavelength image is created, and then each image is filtered to construct a background image for each wavelength. The two standard morphological methods are:

- Opening. A local minimum filter is applied to the whole image.
- Closing followed by Opening. Small dark regions are filled in on the background image, and then a local minimum filter is applied to the whole image.

The Opening method produces a significantly lower estimate of background than any other background-subtraction method. It guarantees that the background estimate is always lower than the feature intensity. However, it may underestimate the true background level. Closing followed by Opening produces background estimates that are slightly lower than standard local background methods for low background regions, but significantly lower for regions with bright patches in the background. However, background-subtracted intensities can still be negative using this method.

Regardless of the method chosen, the calculated background intensity is subtracted from the feature intensity before any ratios are calculated. The process of subtraction itself can have several serious effects: it can lead to negative background-subtracted intensities, it can contribute background noise to the feature measurement, and it can add dye bias, especially to low intensity features (8). The contribution of background subtraction to noise and dye bias in extracted data can be quite large, yet for an effect of such severity it is not well characterized.

Background subtraction can add noise and dye bias to a microarray. *Plate 8A* is a plot of the raw red and green intensities from an array plotted against each other on a log scale, while *Plate 8B* shows the same data with the local background subtracted. Note the bias towards the green channel at low intensities in the background-subtracted data. The log ratios show a similar effect. *Plate 9A–D* shows plots of total intensity on the x-axis and log ratio on the y-axis (MA plots) (9) for the various types of background subtraction; the horizontal lines indicate log ratios of +1 and –1 to help see the scatter in the data. Using both local (*Plate 9B*) and global (*Plate 9C*) background-subtraction methods, a large amount of scatter and dye bias is introduced at low intensities. This effect is much less marked if the background estimate is small, as is the case with morphological background subtraction (*Plate 9D*).

Another problem with some background subtraction methods is that the results can depend upon the segmentation method used to separate foreground (i.e. feature) signal from background. If the segmentation method is inappropriate for the image, foreground intensity can be counted as background. For example, very bright spots can sometimes appear larger than dim spots due to either excess material on the slide or optical flare in the instrument. If the feature indicators are smaller than the largest spots, signal from the spots will be outside the feature indicators and will be counted as background. This effect is apparent as an intensity-dependence in the background distribution (*Plate 10*).

For all these reasons, there is a growing trend in the microarray community to not use background subtraction at all, or to use methods like morphological opening that are independent of segmentation and provide relatively low estimates.

An alternative to simply ignoring background is to use a model-based method that estimates the true intensity of a spot by modeling the contribution of the background (10). Model-based methods for estimating background are also becoming popular for analyzing data from the Affymetrix platform. The Robust Multichip Analysis method of calculating expression levels from Affymetrix arrays entirely ignores the mismatch probes, which are designed to estimate background (11).

Finally, one can use a ratio calculation method that is less skewed by background subtraction. The y-axis values in *Plate 9* all represent the ratio of medians, which is the ratio of the median pixel value of a feature. An alternative ratio measure is the median of ratios, which is the median of the ratios of the individual pixels in each feature. *Plate 11* shows an MA plot of the background-subtracted median of ratios. It has less dye bias than the background-subtracted ratio of medians in *Plate 9B*, and while it shows more spread than the raw data distribution in *Plate 9A*, it is significantly closer to this distribution than to *Plate 9B*. Another advantage of the median of ratios over the ratio of medians is that it has a Gaussian distribution, while the ratio of medians does not (12).

15.4 Normalization

Normalization is the process of removing bias from a measurement. Data on a microarray may be biased for several reasons including differences in dye properties, probe labeling, and hybridization efficiencies, as well as inappropriate detector settings on the scanner (see also Chapter 17). We address only the scanner issues in this chapter.

A common misconception with microarrays is that a 1:1 ratio of signals on the array corresponds to a 1:1 ratio of gene expression. As discussed above, equimolar amounts of different dyes may not produce equally bright signals. Equal signal is guaranteed to represent equal gene expression only if the dye is the same for all probes, the probes are labeled to the same density, and the probes have equal hybridization efficiency to their respective targets. However, quantifying the contribution of each of these factors for each microarray experiment would be extremely cumbersome. The dyes and filters commonly used for microarrays have been optimized to minimize differences caused by inherent dye properties, and including dye swap

replicates is a useful way to control for dye batch, labeling and hybridization differences. With carefully prepared and properly controlled experiments, multi-channel microarrays can provide reliable estimates of actual biological ratios.

The purpose of adjusting the PMT is to maximize the range of numerical values that are available to represent the fluorescent signal from the sample. All channels should be set to the highest gain possible without causing saturated pixels. In addition, it is easier to visually evaluate microarray images if the channels are equally bright, as indicated by a ratio of approximately 1.0. Note that the human eye is very sensitive to color, and is a very reliable judge of signal intensity, so many researchers simply rely on visual inspection to balance the channels.

In many whole-genome gene expression experiments the mean of all ratio values should be close to 1.0 because for any given experimental system relatively few genes are differentially expressed, and approximately the same number are over-expressed as are under-expressed. The PMT gain can be set to balance the signal from both channels by calculating the mean ratio of all features on the array, and adjusting the feature intensities so that this mean is set to approximately 1.0.

However if the microarray contains a small or functionally specific set of genes, or in microarray experiments that examine organisms under extreme conditions, such as heat shock, starvation or stationary phase, we may expect many of the genes to be differentially expressed. In this case normalizing the data to force a global mean ratio of 1.0 may mask important differential expression. In such cases, one may prefer to use housekeeping genes or spiked-in controls (13). Spiked-in controls of known ratios across a range of expression values provide an external standard by which one can normalize all genes on a microarray, regardless of the distribution of the genes being probed. When using external controls, one must ensure that the controls are measured at the expected ratios. External controls should be calibrated against an independent technique such as quantitative PCR. Properly calibrated external controls provide a robust method of normalization.

Using so-called 'housekeeping' genes as controls has fallen out of favor because in some organisms, such as yeast, no gene exists that is unchanged under all conditions. However, if you are studying a subset of the genes in a genome, specific experimental conditions, or specific tissues it may be possible to select a set of housekeeping genes that reliably yields the same signal across all arrays in the experiment (14).

Once you have balanced the PMT settings of the scanner, you can use computational normalization methods to adjust for many other types of non-uniformity in the data, both physical and statistical:

- Spatial non-uniformity
- Print-tip non-uniformity (a special case of spatial non-uniformity)
- Intensity dependence of ratio values
- Intensity dependence of variance.

Locally-weighted scatterplot smoothing (LOWESS) normalization corrects for the first three non-uniformities (unless the spatial non-uniformity is seen within print-tip groups). A variance stabilization method such

as that proposed by Durbin *et al.* (15) can correct the last problem. Further discussion of these methods is beyond the scope of this chapter and is addressed in Chapter 17.

References

1. Haugland R (2002) *Handbook of Fluorescent Probes and Research Products, 9th Edn* Molecular Probes, Inc, Eugene, OR.
2. Pickett S (2003) Understanding and evaluating fluorescent microarray imaging instruments. *IVD Technol* **9**: 45–49.
3. The Spectroscopy Net (2003) www.thespectroscopynet.com/Educational/Detection_limit.htm.
4. West Coast Analytical Services, Santa Fe Springs, CA. www.wcas.com/tech/detlim.htm.
5. Brazma A, Hingamp P, Quackenbush J, *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat Genet* **29**: 365–371.
6. Spellman PT, Miller M, Stewart J, *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3(9)**: RESEARCH0046.
7. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systemic variation. *Nucleic Acids Res* **30**: 4 e15.
8. Huber W, Von Heydebreck A, Sultmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**: S96–S104.
9. Yang YH, Buckley MJ, Dudoit S and Speed TP (2001) Comparison of methods for image analysis on cDNA microarray data. *J Computat Graph Stat* **11**: 108–136.
10. Kooperberg C, Fazzio TG, Delrow. JJ and Tsukiyama T (2002) Improved background correction for spotted DNA microarrays. *J Computat Biol* **9**: 55–66.
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2002) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264.
12. Brody JP, Williams BA, Wold BJ and Quake SR (2002) Significance and statistical errors in the analysis of DNA microarray data. *Proc Natl Acad Sci USA* **99**: 12975–12978.
13. van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D and Holstege FC (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* **4**: 387–393.
14. Szabo A, Perou CM, Karaca M, Perreard L, Quackenbush JF and Bernard PS (2004) Statistical modeling for selecting housekeeper genes. *Genome Biol* **5**: R59.
15. Durbin BP, Hardin JS, Hawkins DM and Rocke DM (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18 (Suppl 1)**: S105–S110.

Microarray detection with laser scanning device

16

Ralph Beneke

16.1 Introduction

Microarrays are revolutionizing biology and so are the detection instruments required to read Biochips in an automated fashion, providing millions of raw data in a few minutes. Because the quality of the raw image is so critical for accurate microarray quantification, the primary goal when designing a detection system should be to maximize sensitivity, accuracy and reproducibility.

16.2 CCD or PMT?

Currently the technologies used to image microarrays are detection systems for absorbance, fluorescence or luminescence quantification. Charge-coupled device- (CCD-) based non-confocal systems image fixed area sizes with a limited number of pixels over a scalable dwell time. Sensitivity increases with dwell time and number of photons per μm^2 detected. But the longer the dwell time the higher the dark current noise. Because of non-confocality of CCD systems the intrascenic dynamic range is limited to functional 12- or 14-bit compared to 16-bit confocal laser scanning systems. The contrast from bright signal to background for the maximum resolution is 10–100 times better with microarray scanners compared to CCD imagers (*Figure 16.1*). The lower the number of fluorescence molecules the more sensitive the scanners are compared to CCD imaging systems for similar throughput. CCD chips acquire data on a limited area providing a snapshot depending on the size of the CCD chip. The data acquired by a CCD on a typical 3 × 1-inch microarray glass slide are stitched together. The background and signal correction is introduced by sophisticated software algorithms. This contributes to a lower image quality and reliability compared to images acquired by confocal scanners. The non-confocal detection of glass chips by a CCD camera results in a significant increase of background in the neighborhood of bright signals and the contribution of unspecific background from the backside of a contaminated glass chip. For transparent plastic chips and microplates the contribution of auto-fluorescence of the polymer is incredibly high for non-confocal imaging systems.

Scanners with an optimized confocality for planar glass surfaces currently provide the highest dynamic range and resolution in combination with sensitivity and speed:

(i) Laser scanning systems for gel detection (2D proteomics, 1D SAGE, etc.) have a lower optical resolution and are less confocal in order to optimize detection of larger features (mm range in all three dimensions) on larger areas (43 × 35-cm blots). Compared to microarray scanners, which are designed to scan 3 × 1-inch glass slides or even smaller chip areas to detect tens-of-thousands of 50- to 200-µm spots by 5- to 10-µm pixel digitization, typical gel scanners cannot be used for microarray detection providing sufficient data quality.

(ii) Microarray laser scanners illuminate fluorescence markers of the sample pixel by pixel by moving the optics and/or the chip very fast (10–30 Hz). The emitted fluorescence photons are gathered pixel by pixel in a photomultiplier tube (PMT) using mirrors, lenses, dicroics, filters, and pinholes. The PMT converts photons into electrons resulting in voltage-dependent analogue signal, which is translated by A/D converter into digital 16-bit raw image data. Depending on PMT gain

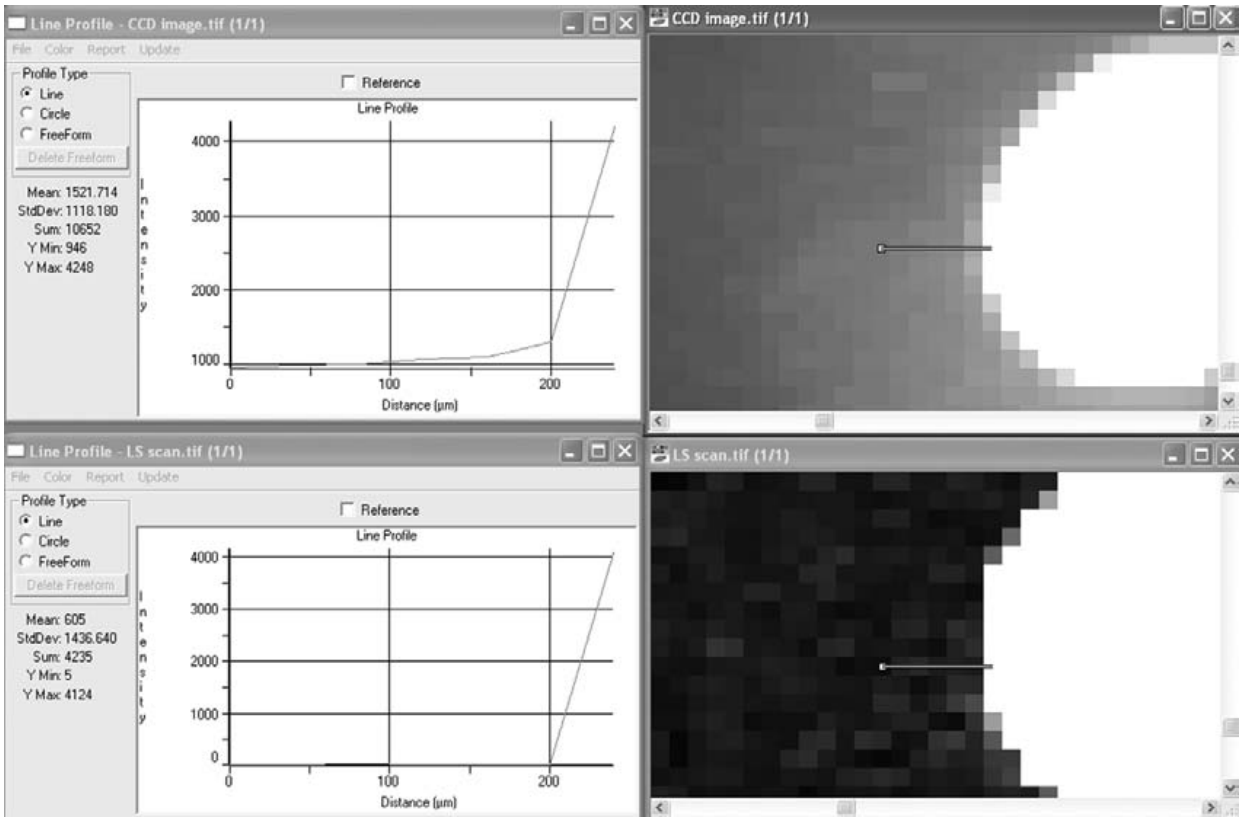


Figure 16.1.

Intrascenic Dynamic Range of images acquired by laser scanners is 10 times higher than for images acquired by CCD imagers.

(voltage-dependent) one photon generates a bunch of electrons translated into arbitrary counts of a pixel. Providing high signal-to-noise ratio on pixel, feature and image level as well as calibrating sample image data to a fluorescence standard is prerequisite to enable relative and quantitative comparable results.

16.3 Engineering of Tecan's LS series

Format flexibility

The Tecan LS series microarray scanner provides flexible scanning technology to automatically scan many different format samples such as: glass slides; evanescence resonance glass slides (NovaChip, Novartis Pharma, Basel); plastic slides (black or transparent); segmented slides and plates with tiny upper structure (e.g. HTA Greiner-Bioone); clear-bottom microplates (e.g. Nunc, Greiner-Bioone, Matrix Technologies, Schott–Nexterion Telechem); small micro-fluidic chips; membranes (FAST slides from S&S/Whatman); mini PAA gels, hydrogel protein arrays (PerkinElmer); CodeLink hydrogel slides (Amersham/GEHC); mirrored or gold-coated slides; tissue arrays; colony plates; and coverslips. Basically all formats in the range of 6 × 6-mm square to 127 × 85-mm square and up to 15 mm in height can be placed on the sample tray of the LS and automatically focused. The fluorescing sample itself can be 5 mm deeper than the top of the chip or chip-adapter. LS enables the scanning of more complex formats than just a glass slide of 3 × 1 inch size and 1 mm thickness.

Selective excitation with lasers

Laser for selective excitation

The LS series of scanners provide selective excitation by the option to choose from four different lasers. A mix or contamination of different wavelengths coming from one laser is prohibited by band-selective filters for all lasers. They excite at their predetermined fixed wavelengths only (633 nm red, 532 nm green, 594 nm orange/yellow and 488 nm blue). Currently two lasers are industry standard gas lasers: HeNe for selective 633 and 594 nm excitation. All four lasers can be placed in the same housing including the solid-state 488- and 532-nm laser. Laser power for three out of the four lasers is optimized for microarray fluorescence scanning (10 mW 532 nm, 7 mW 633 nm, 3 mW 594 nm) and is running with 100% capacity, except for the 488-nm laser, which is continuously adjustable to avoid massive bleaching (1–20 mW). Note that gas lasers by their very nature slowly lose 30% intensity over time, which results in a 20% loss of sensitivity. Usually HeNe gas lasers are not 'leaky' below this level. HeNe lasers in LS Reloaded are rated for 20 000–40 000 working hours. Both solid-state lasers of the LS Reloaded are specified for more than 8 000 working hours. After the specified lifetime is reached, 80% of the lasers are still working. After 20 minutes warm-up time the HeNe lasers are very stable during measurement and therefore no active referencing is necessary compared to several solid-state lasers. In order to avoid noise it is recommended to reduce referencing activ-

ity (e.g. of laser) during a measurement. Calculation of noise coming from instable lasers increases noise in the image. The variation of stable gas lasers in the LS series is about 0.5%. The 488-nm solid-state laser is internally power-stabilized.

Block of reflections

Depending on the optical setup and lens system several artificial optical effects have to be considered: the direct reflection of the laser beam from the sample surface and the unspecific reflections generated in the lens system. A special optical setup using a laser beam angle offset from perpendicular excitation and a system of optical apertures enable the artificial unspecific reflections in the LS Reloaded to be reduced. In addition, a pinpoint mirror and a small 'spoon' shutter block the direct reflection of the laser beam from the cone of emitted light.

Selective beam splitting: dicroics

To enable dual color simultaneous detection with LS, a set of two laser beams is synchronized and focused on the same area. The generated fluorescence spectra from different dyes are collected by a lens and mirror system. The mix of two dye spectra is subsequently discriminated by different dicroics (wavelength-selective mirrors with a discrimination line at 575 or 625 nm). To discriminate between red and green the LS uses a 625-nm dicroic and to discriminate blue from yellow and red a 575-nm dicroic is used (*Figure 16.2*).

Emission filter sets

Fluorescence dyes are adapted to be excited by lasers to cover a variety of applications. In order to cover a large range of different dyes excited in the visible range, LS Reloaded offers the option to choose more than the red (633 nm) and green (532 nm) lasers. The yellow (594 nm) and especially the blue (488 nm) laser are necessary to excite additional dyes to perform multicolor assays (e.g. specific labeling of four nucleotides in SNP/sequencing reactions). For each single dye and for dye combinations there are optimized fluorescence-emission-filter settings with specific characteristics defined as blocking efficacy, maximum pass and bandwidth (FWHM: 50% transmission efficiency) depending on the spectral characteristics of the dye used. In Tecan's LS Reloaded filter carriers most of the available fluorescence filters from more than a hundred different standard and special filters on stock can be applied. Laser and filter combination is strictly controlled by software to protect PMTs against damage. Therefore emission filter combinations, which do not block the laser lines at the same time, cannot be used.

The standard proposals are: 690/40 nm, 575/50 nm, 635/35 and 535/25 nm, or 625/25 nm and 520/10 nm, and many others.

Scanning principle

Scanner mechanism

The 20-Hz voice coil resonance scanner oscillates over the short x-axis of the glass slide and the data are acquired up to 22-mm amplitude. The shorter the amplitude the longer can be the dwell time for data acquisition without compromising speed but gaining the advantage of getting more photons to improve the signal-to-noise ratio (SNR) , which can be selected in acquisition software. The transport table holding the sample (e.g. slide adapter) is moved slowly in the y-direction in steps of 4 to 40 μm depending on the resolution selected.

Focal distance and numeric aperture

The focus or working distance is 6.5 mm. In other words, you can 'look' 6.5 mm 'into' the substrate – if required. Thus formats other than flat planar glass slides can be inserted and focused. Even with this enormous focal distance the numerical aperture (NA) of the special huge lens (0.6) is still high in order to cover a reasonable field of the emitted light providing the high sensitivity required for microarray detection.

Independent of this the maximum height of the 'carrier' on the table should not exceed 15 mm for sample loading.

Detection principle

Selectable confocality

The confocal mode of the LS series of laser scanners provides better background suppression and also better discrimination between bright and dark signals (contrast, intrascenic dynamic range) in the sample. The LS Series Laser Scanner offers three user-selectable pinhole sizes, which give the user three grades of more or less confocal measurements. While handling samples of considerable optical depth (e.g. flow-through chips, gels) or warped surfaces, the instrument can be easily switched from confocal to non-confocal mode.

Adjustable sensitivity

First generation scanners had much slower scanning speed, meaning the time over which each spot/pixel was excited was long enough to make it a concern. The Tecan scanner is much faster, meaning that photo-bleaching has for all intents and purposes become a 'non-issue'. In-house bench studies suggest 1% photo-bleaching per scan depending on default laser

Table 16.1

Small pinhole (0.3mm)	70 μm = ± 35 μm around the focus
Medium pinhole (1.0 mm)	300 μm = ± 150 μm around the focus
Large pinhole (3.0 mm)	800 μm = ± 400 μm around the focus

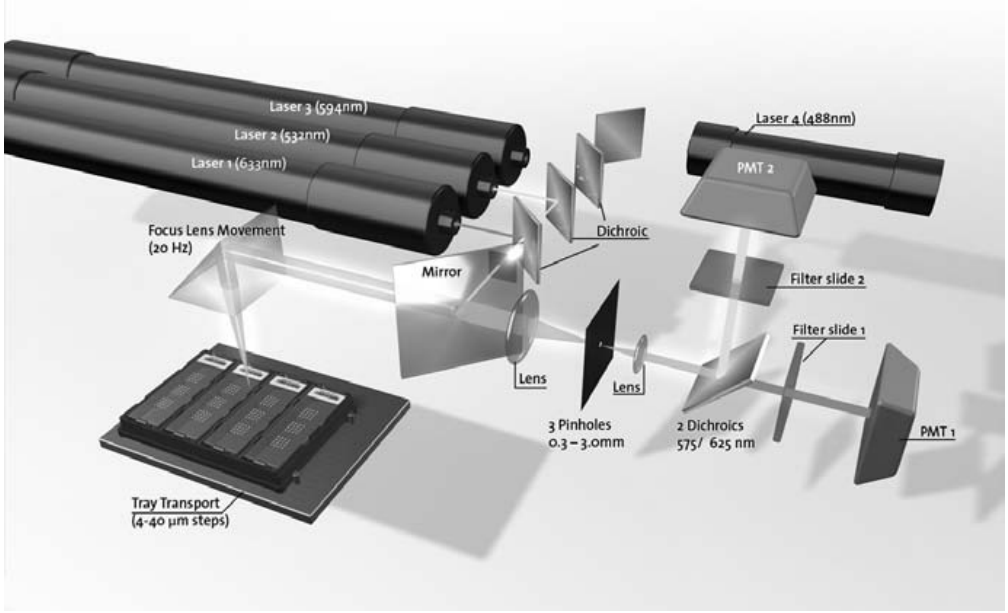


Figure 16.2.

Detection principle and optic scheme of LS for confocal or non-confocal scan.

power. Bleaching and stability in general strongly depends on oxygen, surface chemistry, temperature, humidity and light exposure. Only the latter can be influenced by the instrument.

Photon statistics dictate that the best signal/noise ratios are obtained by having maximum photon output (to a certain degree, of course), which is only achieved by having maximum laser power. Reducing the laser power therefore always reduces the achievable signal/noise ratio, whereas reducing the PMT gain for a wide range does not affect signal/noise. Therefore, the best way to adjust dynamic range is to maximize photon output and to adjust the gain of the detector accordingly. PMTs and electronics of the LS are working linear over the 16-bit dynamics. The software allows adjustment of the PMTs in LS systems over five logs to amplify the photon signal. Commonly array detectors offer two logs (1–100%) adjustable range.

In addition to using maximal laser power in order to increase photon collection, image data acquisition with LS offers several options for improvement:

(i) Use 'optimize integration time' (increase dwell time to maximum) and reduce scan width and resolution accordingly.
(ii) Average scan from two up to eight times.
(iii) Scan the sample through a glass matrix to increase photon collection by a factor of two.
(iv) Increase the pinhole size if required to achieve high uniformity on bulky and warped sample.
(v) Scan evanescence resonance glass slides from Novartis (NovaChip) and adjust the angle of incidence of laser beam according to the laser wavelength used (0° for 633 nm and 20° for 532 nm). The LS Reloaded can

achieve more than 20-fold higher sensitivity for single and dual color scans.

Adjustable laser beam angle: evanescence resonance scanning

The high sensitivity achievable with Evanescence Resonator Chip technology offers the advantage of detecting signal of low abundance genes from limited amounts of sample (e.g. biopsy assays). The continuous adjustable angle of incidence of the laser beam in LS scanners enables high sensitivity dual-color scanning of slides supporting evanescence resonance technology.

Selectable resolution

Resolution in an optical scanning system is always the result of a combination of several factors. Most important parameters are: pixel size, laser beam size, confocality and the scanning scheme. Only if we consider the combination of all parameters do we get the full picture (in terms of resolution). Let's start with a fundamental law of information theory that says that you always need to sample a signal with twice the bandwidth or frequency you actually want to see or hear. In a first order approach for an optical scanner system this means that the pixel size should be approximately half of the laser beam diameter. For a laser beam diameter of 12–16 μm the ideal pixel size therefore would be 6–8 μm. Most of the microarray features have a 100 μm diameter. On the other hand this also means that reducing the size of the laser beam without reducing the pixel size does not give the full benefit in terms of resolution. Therefore a pixel size of 5 μm and a laser beam size with the same dimensions might be slightly better, but is no way near twice as good as 5 μm pixel size with a laser beam of 10 μm diameter. However, there are two drawbacks of a very tight laser focus at the sample. First there is a higher risk of bleaching because the intensity at the small focus is much higher. The intensity providing higher sensitivity, and therefore the risk of bleaching, scale with the square of the beam diameter.

The second problem is under-sampling. If a laser beam diameter is rated as 10 μm according to standard definition this means that within a range of ± 5 μm the intensity of the beam is already down $1/e^2$. At the borders of such a beam only 14% of the maximum intensity is left. If a fluorescent molecule happens to sit away from the center of the beam it will be excited by a much lower intensity. In other words many of the labeled molecules will not be very efficiently excited and contribute much less to the overall signal than they could. Obviously this is a very sub-optimal (less sensitive!) way to make use of the rare sample molecules found in a spot at the end of the whole array process. The LS is designed to scan microarray spots from 60 to 200 μm in order to generate a minimum 100 pixels per feature on small spots and without the issue of under-sampling on larger spots.

Scattered light scan: slide quality control

Microarray spots, which have no label, can be detected for quality reason without using an indicator dye, if they contain salt crystals (SSC in spot-

ting buffer). Any light will be scattered at these spot crystals and can be detected by using the delivered absorbance filter of the LS for scanning. An automated image analysis for local and global background, spot density, spot morphology, and spot position, can be applied. A report of 'passed or failed' spots and slides by using appropriate (customized) statistics means that low quality slides can be rejected from hybridization and false negative and positive spots can be discarded from the analysis and re-evaluated.

Correction of crosstalk from dyes

High-throughput sequencing, single nucleotide polymorphism (SNP) detection, gene expression analysis and many other applications often demand multiplexed labeling on one array chip. The option of parallel dual-channel detection allows two dyes to be measured at the same time for maximum imaging throughput. A single dye calibration step is sufficient to give the required specificity for each measurement. Unwanted signals from dyes in either channel are easily accounted for by rapid recalculations using accompanying inbuilt software. Sequential scanning of three or more dyes or with different gain settings to achieve extended dynamic range can be performed by running batch mode scripts.

The LS series of instruments operate up to four lasers and 28 filters. Each LS Series Laser Scanner has one to four lasers (633-nm HeNe). On the detection side, more than 100 fluorescence emission filters are available. A collection of 28 emission filters can be used for each instrument at a time.

Signal-to-background and signal-to-noise

Most microarray processing steps suffer from introducing high unspecific background by contamination or sub-optimal stringency of incubation and washing. The detector itself produces a dark current signal as a more or less flat (dark current noise) baseline background. This background can be subtracted by default from the image data by a given offset implemented by the system. The offset of a detection system cannot be simply measured by a user. Spreading the entire signal range over 16 bit, some holes in the histogram indicate the value of subtracting the dark current count. More substantially the background and the noise of the image can be an indicator of the image (and detection) quality. The 'quick and easy' evaluation is to take a reference slide with a bright (not saturated) signal and the background from a surrounding area (10 pixel distance to signal) and calculate the CVs and the signal-to-background ratio. For measuring the detection limit Tecan recommends to achieve a factor of 2 from SNR calculation ((raw intensity – background)/standard deviation background)). SNR can vary and are highly dependent upon the type of sample and what dyes are used. Although Tecan has not established specific SNR numbers for Cy3/Cy5 'standard' arrays, note that noise from the sample (vs noise from the instrument) limits SNR in 98% of cases. As a rule of thumb one could say that at maximum signal the LS Reloaded can achieve SNR of 5 logs. The electronic dark current count noise of LS Reloaded – independent of a sample – is typically 3–5 counts. Comparing two scanner systems it is recommended to use a set of two replicate slides on both systems and make

sure that bleaching and degradation of the dyes is a non-issue for the test. Calibration or validation of fluorescence scanners is not trivial because most of the current test tools are not stable.

Dynamic range

Dynamic range of a detection system is established here as the ratio between the highest signal and electronic noise measuring a full spot containing 100 pixels. Electronic noise floor = square root of number of pixels × 5 counts per pixel. Maximum signal = number of pixels × 65 000 counts per pixel. Maximum signal results = 6.5 million counts/50 counts which yields five orders of magnitude difference.

The intrascenic dynamic range is important to quantify bright signal against low signal or background. An adjacent area to a very bright spot should not be influenced by the high signal value (about 60 000 counts mean) of the bright spot area. The image should have a background of close to zero in the spot pitch distance (about 10 pixels) of a bright spot. Almost all CCD systems have problems in achieving four logs here because they have broad area illumination and imaging capacity. They cannot excite and detect pixel-by-pixel and lack a confocal scanning mode (*Figure 16.1*).

Automatic sample load, focal adjustment and scan

Batch mode capability

A slide holder with the same outside dimensions as a standard (SBS) microplate allows automatic sample loading with standard handling devices such as Tecan's Connect autoloader (*Figure 16.3*). Even without automatic loading, up to four standard microscope slides can be processed with no further user intervention in one run. In addition, almost any user-defined method can be applied to up to 200 slides or 50 microplates in a single run using autoloading system Connect™ from Tecan and ArrayPro® Analyser from Media Cybernetics allowing fully automated batch processing of image acquisition and analysis.

In addition, each individual scan area (on a slide or microplate) can be re-scanned in the same batch run by applying different scan settings as a series of gains, lasers and filters defined as a simple scanlist.txt file in order to extend the dynamic range and to use up to four dyes on one slide. With its automatic focus and gain control and the option to implement Tecan's Connect as an autoloader system the LS series instruments have batch mode capability and optimized dynamic range of each individual scan area. Array-Pro supports the fully automated analysis of multicolor images (e.g. four-color sequencing and SNP analysis) in an unlimited number of experiments.

Balance of automation and interactivity options

Each format and sample is different. Positioning the sample on a mechanical adapter or table can vary over several tens or hundreds of micrometers in every device. This is absolutely critical if the scan is done confocally.

Figure 16.3.

Autoloading of slide adapters and microplates on LS scanners.

Because of small mechanical tolerances LS systems use an autofocus system, which is activated just before the actual scan takes place in order to adjust the sample automatically in the focus (max peak of CV on slide area <5%). The autofocus and sample tray corrects for z-positioning and for two angles of tilt before a planar scan starts. In addition, automatic gain control enables walk-away time and avoids saturation effects on single slides out of a batch. Online display during all scans and additional interactive dialogues for setting scan positions in X/Y and Z-prescan images makes parameter definition convenient and intuitive.

Several other options for manual interference and adjustments enables a high customization and optimization level of the application, which are filed as user templates. The user template library, in combination with the automatic focusing, assure reproducibility and reliability.

Quality control and preventive maintenance

Operation quality control

Tecan offers a reusable LS performance check tool (LaserCheck), which enables a convenient quality control routine of LS Series Laser Scanners in the customer's lab independent from using standard spotted slides which suffer from bleaching and degradation. The procedure is automated and wizard-guided. The report gives pass/fail feedback about optical alignment, electronics, sensitivity and mechanics. Tecan's service and production departments are using the same tool for detailed parameter readout (*Figure 16.4*). Tests performed are:

Figure 16.4.

LaserCheck: calibrated laser scanner operation quality check tool for unlimited re-use.

- Gain calibration
- Barcode reader
- Electronic noise
- Oscillator amplitude and data synchronization
- Laser intensity
- Sensitivity
- Autofocus
- Alignment
- Filter blocking

The laser-light-resistant fluorescing surfaces (vaporized organic dye) used for sensitivity and autofocus tests are calibrated in the Tecan factory. Each plate comes with a calibration certificate and the calibration values (serial number, date, time, operator, calibration data, expiration date) in a calibration file on the LaserCheck CD. The calibration file and the certificate are valid for 1 year. The LaserCheck plate should be recalibrated once a year in Tecan's factory for further unlimited use in the lab. The overall system performance check with more than 100 different parameters for a four-laser system takes a routine of 60 to 45 min. The raw data are stored as Excel spread sheets and 16-bit tiff images. The results are referenced to the system specifications. The module tests are reported as passed/failed and filed electronically and as hard copy without further need of manual data analysis.

Normalization strategies for microarray data analysis

17

Christine Steinhoff and Martin Vingron

17.1 Introduction

The basic requirement for a gene expression microarray experiment is that the measurement of intensities of each spot can in some way be interpreted to reflect the corresponding number of mRNA molecules in the sample under consideration. However, it is well known that raw intensity measurements are highly influenced by a number of different external factors, for example effects due to pins (1–3), PCR plates, sample preparation, array coating, spotting, labeling efficiencies, nonlinearity of dye-labeling, scanning, and so on (for an overview see 3, 4). Thus, it is obvious that raw intensity measurements of microarray spots typically do not reflect respective mRNA levels.

In order to achieve a biologically meaningful interpretation of the experiment, these influences have to be statistically described. It is imperative to bring samples which are compared in the course of analysis not only to a common scale (scaling methods), but particularly to remove effects which are not meant to be part of the biological interpretation as thoroughly as possible. The process of normalization should lead to the correction of those effects that are due to variations in the experimental procedure. Furthermore, the dynamic range of the data as well as the distribution of intensities might be different when comparing several arrays within one series of experiments. Recapitulating the goal of applying any normalization method is to adjust for many influences other than those due to the biological differences in the RNA samples.

Over the last few years, a number of so-called normalization methods have been published to overcome the problem of various effects and to end up with a dataset that allows for further statistical analysis (for reviews see 3–5).

The basic question is: which mathematical description can explain the underlying data best, or what kind of data description is properly specifying the biological nature of a microarray experiment? While it is impossible to rule out all influencing factors and to exactly describe the underlying biology, it is nevertheless crucial to find out whether setting up a data description of higher complexity is more appropriate in biological terms.

We will review frequently used normalization methods and demonstrate their application to biological datasets.

Overall, the normalization methods which have been published during the last few years can be divided into procedures that are based on the assumption that the majority of genes detected by the array change in expression or remain unchanged in the experiment. In this article, we are focusing on the second group, namely experiments in which the majority of genes remain unchanged. Normalization methods of the second group can be divided into (i) scaling or standardization methods and (ii) normalization methods using a normalizing transformation of the data (see *Figure 17.1*). Note, that scaling methods can in fact only correct for globally multiplicative effects by appropriate scaling of the data. Nevertheless, they are often called normalization methods as well.

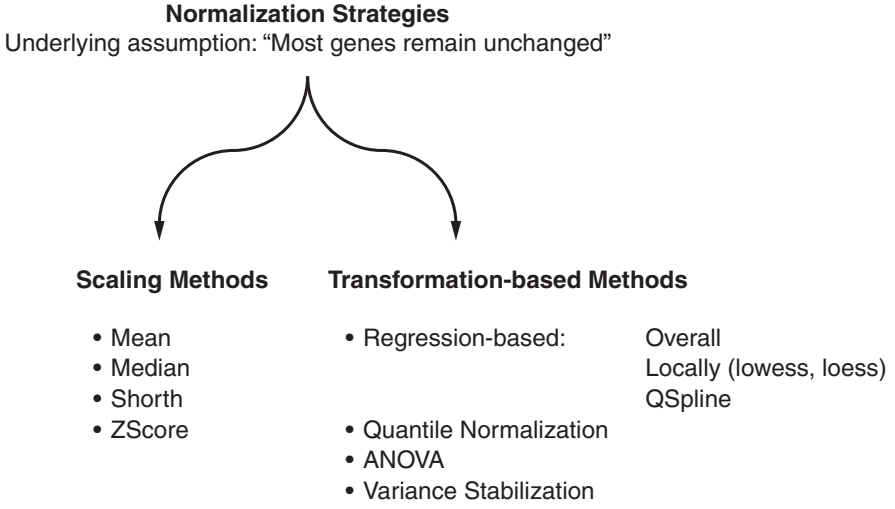


Figure 17.1.

Overview of normalization strategies used for microarray data analysis.

In this article we outline mathematical procedures to describe and remove various kinds of effects in microarray data. Some of these variations are systematic, for example pin effects, and can be estimated using the measured data in many cases. Others are random effects, and appropriate error models for these will be discussed. In the following sections we first introduce the experimental data we are using. Then, examples of scaling methods are explained and we discuss the problem that these methods can only correct for globally multiplicative errors. Subsequently, we describe some of the most frequently applied normalization methods which are based on data transformation. We demonstrate the application of the presented normalization methods using two published biological microarray datasets. Note that we describe the normalization methods for cDNA array technology. However, they can also be applied with only minor changes to Affymetrix datasets.

17.2 Experimental data

We used the BDNF dataset published by Gurok *et al.* (6). In that study, neural progenitor cells were prepared from postnatal day 7 CD1 mice (Charles River, Wilmington, Mass.) and their differentiation process was examined. For comparison of normalization methods, we only used the experimental data of undifferentiated cells and the first day of differentiation with BDNF. This data consists of one dye-swap experiment, so two arrays were used, and signal intensities of two laser channels were measured per array. The data is available as supplementary information to the publication (http://www.molgen.mpg.de/~dna_microarrays-neural_differentiation/neural.html). This data serves as an example of a high number of genes showing a very low expression.

As a second example we used a gene expression dataset from the swirl zebrafish cDNA microarray experiment which is available in the Bioconductor (22) Marray-package. The dataset consists of two dye-swap experiments of which we used the first array (slide 81). It serves as an example of nonlinearity of microarray data.

17.3 Normalization methods

Microarray data is frequently displayed in logarithmic scale. From the graphical representation of microarray intensities and ratio-display it is immediately clear that the log-display shows a more convenient image. When plotting a histogram of raw microarray intensities one normally gets a shape similar to a geometric distribution. Plotting the logarithm of the data points results in a shape which is similar to a normal distribution. Another example is the display of ratios when comparing two RNA samples. Plotting the intensities of each sample against each other results in a display where most data points are clustered in the lower left hand corner of the plot. Instead, the graphical display of the log-product ($\log(\text{sample}_1 \times \text{sample}_2)$) plotted against the log-ratio ($\log(\text{sample}_1/\text{sample}_2)$) is more informative (see *Figure 17.1*) and the log-ratio can be interpreted as a measure for differential gene expression.

Thus, using logarithmic scale evens out skewed distributions of the data and gives a more realistic picture for outliers when displaying the log-product versus the log-ratio of two samples. Furthermore, by applying the logarithm of the intensities multiplicative effects become additive.

As already mentioned, in logarithmic scale the intensities are rather equally distributed across their dynamic range while this is not the case for the untransformed display of the data. In fact, this is a big advantage for the visualization but not necessarily for the analysis. Problems arise with the many low intensity values or negative values which are frequently evident after performing background subtraction. Thus, for low intensities we get a very strongly scattered plot and for zero or negative values we can even get non-defined data points.

Typically, log-ratios are normally distributed and at least for high intensities the variance is independent of the intensities, which is another advantage of log transformation for data analysis. However, this does not hold for low intensities. There, the variance is dependent on the intensity

and variance of log-ratios and decreases with increasing log-product values. Thus, when visualizing log-ratios, one has to take into account that the variance is not constant along the whole dynamic range. To overcome this problem, variance-stabilizing normalization has been introduced (8, 9), which is reviewed at the end of this section.

17.4 Scaling methods

All scaling methods are based on the determination of a common scaling factor which is subsequently used to rescale the whole dataset, such that a global rescaling is achieved.

Let x_i (i =1,…,n) denote the logarithm of background-subtracted raw intensities of array spots. When applying any scaling method, one subtracts the determined common factor from the measured data x_i for all i =1,…,n. Scaling methods fix a range of intensities such that experiments are comparable over a common range. This is true for the following scaling methods:

Overall mean

Each set of signal intensities in a hybridization experiment is normalized by the mean such that the mean logarithmic signal ratio of each set is zero. In case of a cohybridization experiment where two differently fluorescence-labeled target samples are used, a set of signal intensities is the set of measured data resulting from one of the two dyes. That means, one subtracts the mean of all logarithmic background-subtracted data points from x_i for all i =1,…,n.

Overall median

Each set of signal intensities is normalized by the median such that the median logarithmic signal ratio of the two channels is zero analogous to the scaling for the overall mean.

Shorth of the data

The shorth of a univariate distribution is defined as the shortest interval containing half of the values. In the unimodal case the mode, which is the most frequently occurring value, is a robust estimator. Here, the shortest interval containing half of the genes is determined and, as an estimator for the shorth, the median of the genes in that interval is calculated and used as scaling factor.

Normalization using ZScores

Sets of signal intensities are centered on zero with overall variance of one. The underlying assumption is that overall the logarithmic data follows a normal distribution $N(\mu,\sigma)$ which can be transformed into $N(0,1)$ by $\log(\hat{x}_i) = (\log x_i - E(\log x_i))/\sigma(\log x_i)$. Normally, as estimators for E, the mean, and for σ, the standard deviation is used (for example, see (10)).

Scaling methods can only correct for systematic effects that are globally multiplicative. They were used in the very beginning of technology development (11, 12). Also, early analysis packages used scaling methods (AtlasImage 1.101, ClonTech).

17.5 Transformation methods

In this section we review some of the most frequently used transformation-based normalization methods such as regression methods (global-, local- and qspline-based), ANOVA, variance stabilization and quantile normalization. Overall tendencies in the data might be corrected by choosing an appropriate regression model. Several regression-based models have been used recently.

The basic idea of introducing some error model is to describe the relation of the measured signal intensity with regard to the true abundance of RNA molecules. Assuming that the true intensity level x_{kg} of the kth sample and gth gene is disturbed by some random multiplicative (b_{kg}) and additive (a_{kg}) factors, the actual measurement of the gth gene in the kth sample y_{kg} can be described as follows: $y_{kg} = a_{kg} + b_{kg}\, x_{kg}$. Proposing models and approaches that determine and optimally describe the factors a_{kg} and b_{kg} in stochastic terms has been the focus of many publications over the last few years.

Introducing an error model which describes the nature of intensity measurements including systematic and random effects, and which estimates true gene expression according to the error model, should improve the analysis. While normalization without introducing an error model might be able to correct for systematic effects that frequently appear in the data, noise effects that stochastically show up can be captured by an appropriate error model.

One of the first approaches to determine a multiplicative term was proposed in 1997 by Chen *et al.* (13). An integrative description of multiplicative and additive factors in an extensive error model was introduced by Rocke and Durbin (14) and led to the normalization model of variance stabilization (8, 9). A good overview of the development of recent error models for describing microarray datasets is given in Huber *et al.* (4).

Regression methods

Regression methods correct a dataset of signal intensities by either overall (15, 16) or locally (17) estimating an optimal polynomial function (either linear or of higher degree) that explains the local or entire dataset's tendency.

1. Linear regression
   Here, a function $f(x) = ax + b$ is fitted to the *log-log* plot of two sets of signal intensities resulting from two differently fluorescence-labeled target samples by the least-squares method. The dataset is normalized according to $f(x,)$ that means $(f(x)-b)/a$.
2. Polynomial regression of degree >1

A function f is fitted to the dataset (*log-log* plot of both sets of signal intensities) as in the linear case, but f is a polynomial function of degree >1:

$$f(x) = \sum_{i=1,\ldots,m} a_i x^i + b, \text{ where } m > 1.$$

3. Local regression via loess/lowess (locally weighted scatter plot smooth)
For each z in a sliding window, a linear (lowess) or quadratic polynomial (loess) weighted regression function is estimated locally. Here, a descending M estimator is used with the Tukey's biweight function. For normalization of gene expression microarray datasets as proposed in Yang *et al.* 2001 and 2002 (2, 18) a lowess curve is fitted to the A versus M scatterplot, where A denotes the *log* product intensity of two channels r and g ($A = \log \sqrt{rg}$) and M the *log* ratio ($M = \log r/g$). Loess/lowess-normalization is widely used and has been widely adapted to many applications, specific problems and trends such as print tip effects (1, 3, etc.).
4. Local regression via Locfit
The underlying model is $Y_i = \mu(x_i) + e_i$, where $\mu(x_i)$ is assumed to be smooth and is estimated by fitting a polynomial model within a sliding window. For each point x consider a locally weighted least-square criterion $\sum_{i=1,\ldots,m} w(\frac{x_i - x}{h})(y_i - (a_0 + a_1(x_i - x)))^2$ where $w(v)$ $= (1- |v|^3)^3$ for $|v| > 1$ *and* $w(v) = 0$ otherwise; h denotes the band width. As in the case of loess/lowess normalization, the locally estimated curve is fitted to the A versus M scatterplot.
5. QSpline normalization
Workman *et al.* (19) proposed a normalization method where intensity pairs of two arrays are interpolated according to a cubic spline function. Here, smoothing B-splines are fitted to the quantiles from raw array signals of both channels. Then, the splines are used as signal-dependent normalization functions. This method is implemented in the affy-package in Bioconductor (http://www.bioconductor.org/; library: affy, function: normalize.qspline).

Analysis of variances (ANOVA)

Applying ANOVA for microarray data analysis was first proposed by Kerr *et al.* (20). Assuming a dye-swap experiment setting, the underlying statistical model is: $\log(y_{ijkl}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg}$ where μ is the overall mean, A_i the overall array effect, D_j the overall dye effect, V_k the overall variety effect and G_g is the overall gene effect across the other factors. The $(AG)_{ig}$ term describes the potential effects which are specifically due to the variation in the amount of spotted cDNA on the array. $(VG)_{kg}$ describes signal intensities explained by the considered variety (probe vs control), which is the main factor of interest in detecting biological differences. The error term ε_{ijkg} is assumed to be normally distributed around zero. Parameters are calculated by maximum likelihood estimation. The

authors provide codes for Matlab and R on their homepage: http://www.jax.org/sta_/churchill/labsite/software/.

Variance stabilization

Normalization by variance stabilization (8) comprises data calibration, the quantification of differential expression, and the quantification of measurement error. In particular, this normalization method leads to the correction of the variance-versus-mean dependence that can typically be observed when examining the variance-to-mean plots of background-corrected microarray intensity data. For the transformation *h,* the parametric form $h(x)=arsinh(a+bx)$ is derived from a model of the variance-versus-mean dependence for microarray intensity data. The difference statistic Δh has approximately constant variance for the whole intensity range of the array. Note that for high intensities, *h* coincides with the logarithmic transformation. For low intensities the *arsinh*-transformation is continuous in contrast to logarithmic transformations. This is because there is no singularity around zero as in the case of logarithmic transformation. The parameters of *h* together with those of the calibration between experiments are estimated with a robust variant of maximum-likelihood estimation. The variance stabilizing model was introduced by Huber *et al.* (8) and Durbin *et al.* (9) and is implemented in the vsn-package in Bioconductor (http://www.bioconductor.org/; library: vsn).

Quantile normalization

In order to get the same overall distribution of intensities, the array-intensity values of *n* arrays are normalized by projecting each quantile of intensities to lie along the unit diagonal. In *n* dimensions all *n* data vectors should have the same distribution such that plotting the normalized quantiles in *n* dimensions leads to the unit vector $(1/\sqrt{n},\ldots,1/\sqrt{n})$. To end up with the same distribution for all arrays, one takes the mean quantile and substitutes it with the value of the data point in the original dataset. Quantile normalization was proposed by Bolstad *et al.* (21) and is implemented in the affy-package in Bioconductor (http://www.bioconductor.org/; library: affy, function: normalize.quantiles and normalize.quantiles.robust).

17.6 Application of normalization methods

We applied all normalization methods to the described microarray datasets by using MATLAB (version 6.0.0.88, release 12, MathWorks), R (7) and Bioconductor (22). QSpline, quantile, loess and variance-stabilizing normalization were performed using the default setting.

BDNF dataset

Variance stabilization was performed for each dye swap (two experiments) separately. Apart from ANOVA, all normalizations were used separately for each experiment. After normalization, the sample repetitions were averaged. Thus, we ended up with two datasets corresponding to the two RNA

samples (undifferentiated and differentiated cells) which have to be compared. As graphical display we used the representation of log product versus log ratio. This is shown in *Figure 17.2*. The dye-swap experimental data (6) shows a high percentage of low-expressed genes after background correction. Variance-stabilizing normalization leads to a very good correction of the small-intensity values while other methods cannot correct for this effect and still show – in logarithmic scale – a highly scattered plot in the low-intensities range. Due to the fact that for ZScore scaling the values are corrected for mean and divided by the standard deviation, this scattering effect is strengthened.

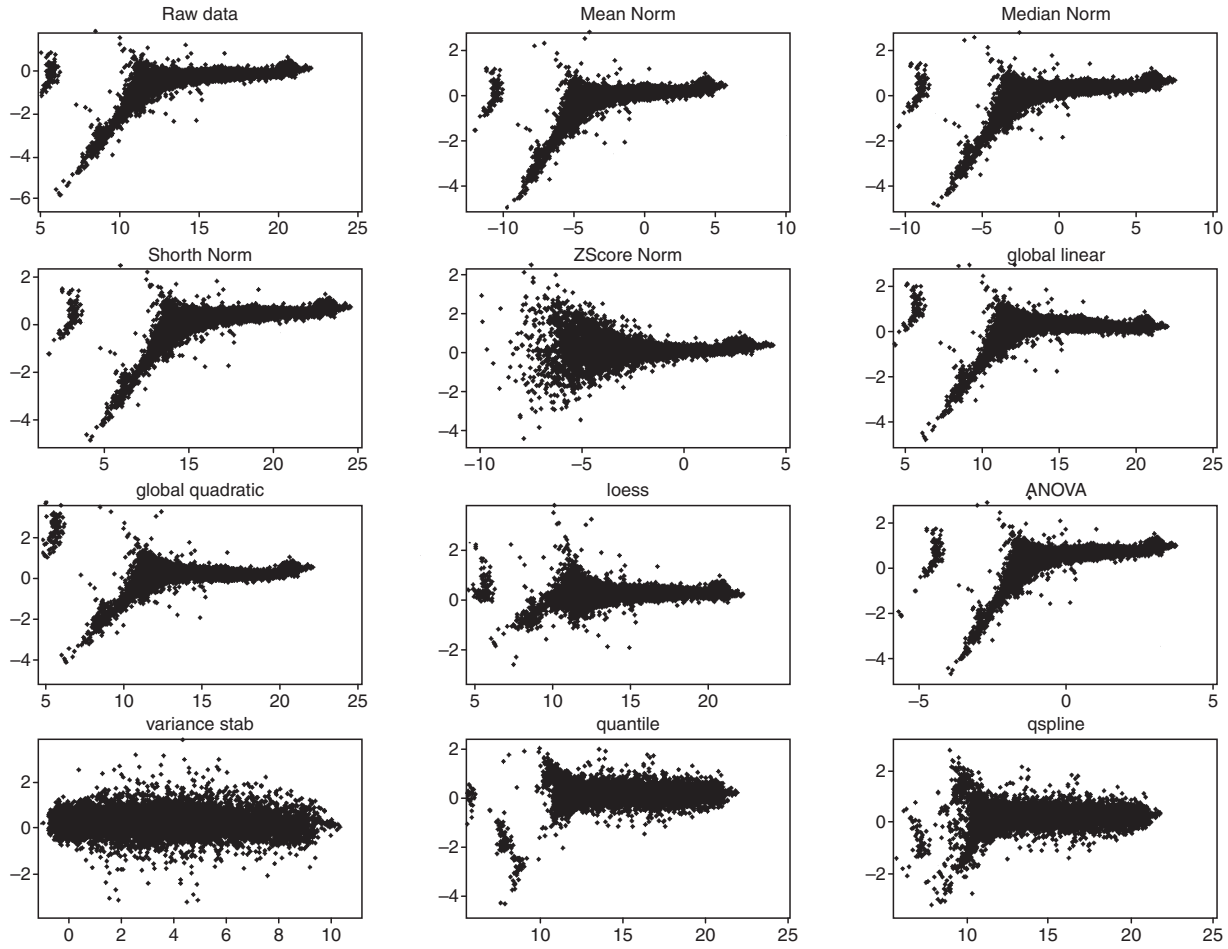


Figure 17.2.

Normalization of a dye-swap experiment. Each subplot displays the scatterplot of log-product (x-axis) versus the log ratio (y-axis) of the mean of one dye-swap repetition. The subplot in the upper left displays the mean values of background-corrected raw intensities without any normalization. The other subplots show scatterplots after application of global mean scaling, global median scaling, shorth scaling, Zscore normalization, global linear regression, global quadratic regression, loess regression, ANOVA normalization, variance stabilization, quantile normalization and qspline normalization.

Swirl dataset

Since the dye-swap dataset shows no nonlinearity, we chose a second dataset with a typical nonlinearity in the logarithmic scale. We used the first experiment of the swirl dataset that consists of one pair of RNA samples (wildtype and swirl). Background-subtracted intensity values were normalized by applying global mean scaling, global linear regression, loess regression, quantile normalization and qspline normalization.

For this dataset, local regression should out-perform global linear regression methods. The result is shown in *Figure 17.3*. Already from the visual inspection it is obvious that the application of local regression is more appropriate than applying a global normalization method. Quantile normalization and qspline normalization also seem to be superior to global methods in terms of correction of nonlinearities.

17.7 Summary

It is obvious that as scaling methods can only correct for globally multiplicative effects these methods appear insufficient to normalize raw microarray datasets in a way that the normalized dataset fulfills the requirement to approximately reflect the corresponding number of mRNA molecules in the sample under consideration. Random effects cannot be captured at all. However, a number of transformation methods have been proposed in the last few years which seem to be more appropriate for the analysis of microarray data. While ANOVA normalization assumes a specific set-up of the experiment which is not always given, local regression methods as well as variance stabilization seem to be appropriate for many experimental settings. Variance stabilization in particular is superior to all other methods when the dataset contains a large proportion of low-intensity values. Local regression-based methods can especially deal with nonlinearities.

Appropriate normalization methods should be able to identify and correct for systematic and random effects in the data. Though one can detect such effects, it is impossible to correct for them in each single intensity measurement within one single experiment. Effects that are due to one specific plate, pin, enzymatic reaction, and so on, can be detected within data preprocessing and a separate normalization is possible. As for the example proposed by Smyth and Speed (3), separate normalization for a specific outlier pin improves the normalization. Thus, in some cases a separate normalization for each pin might be advisable. Problems arise if the same kind of effects is present for plates and other technical issues. It is impossible to adjust for all at the same time since the data subsets get too small. If one wants to use composite normalization one has to decide which kind of technical issue is the most likely influencing factor. Overall, local regression-based methods, such as loess and variance stabilization, emerge as the most appropriate ones. Especially, in the case of strongly scattered values in the low-intensity range, variance stabilization out-performs all other methods.

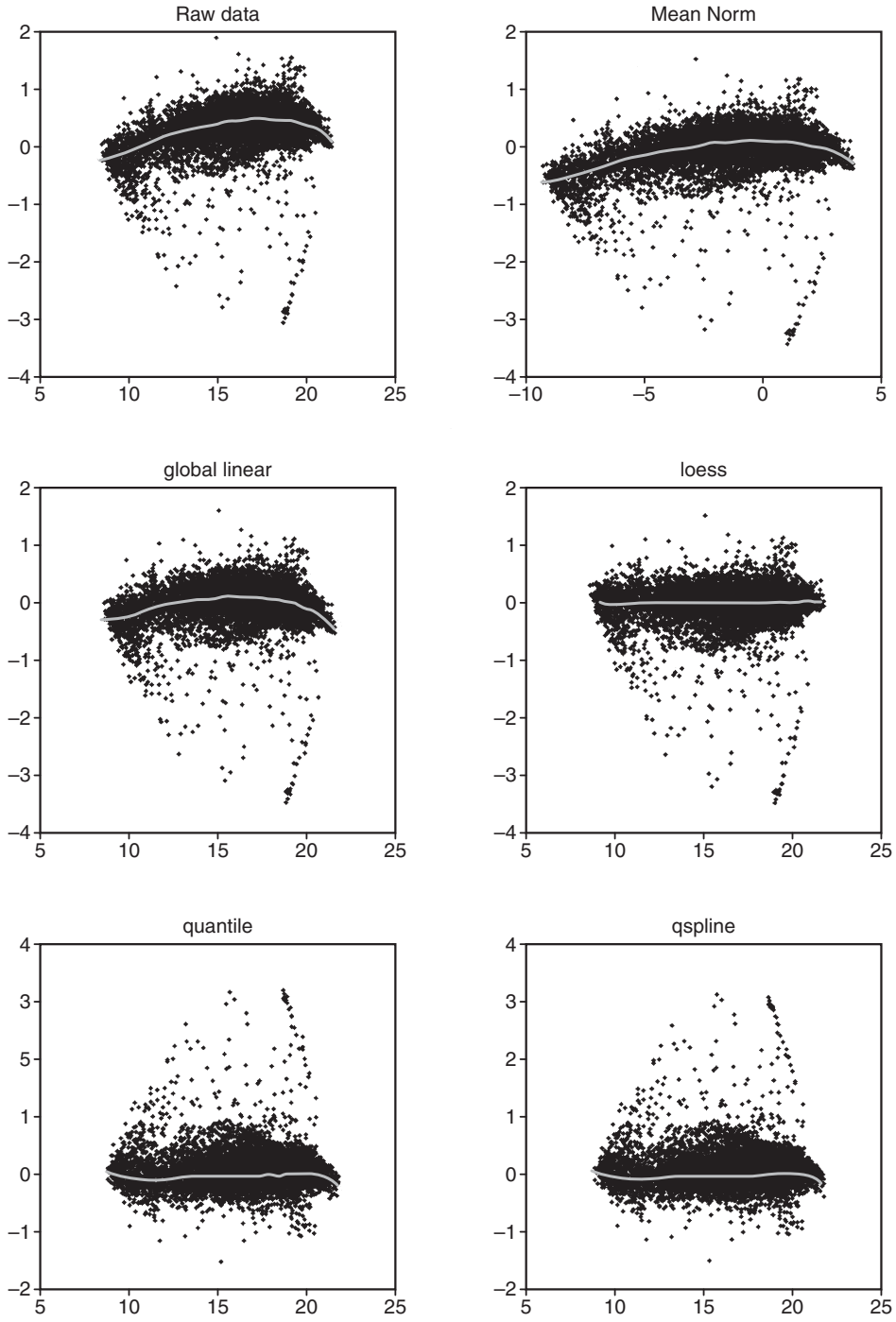


Figure 17.3.

Normalization of one experiment of the swirl dataset. Each subplot displays the scatterplot of log-product (x-axis) versus the log ratio (y-axis) of red and green intensity values. The subplot in the upper left displays the background-corrected raw intensities without any normalization. The other subplots show scatterplots after application of global mean scaling, global linear regression, loess regression, quantile normalization and qspline normalization. The gray line shows the local regression line (applying loess function for each 1% quartile) for each resulting raw or normalized dataset.

References

1. Dudoit S, Yang YH, Callow MJ and Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin* **12**: 111–140.
2. Yang YH, Dudoit S, Luu P and Speed TP (2001) Normalization for cDNA microarray data. *Proc SPIE* **4266**: 1–12.
3. Smyth GK and Speed TP (2003) Normalization for cDNA microarray data. *Methods* **31**: 265–273.
4. Huber W, Von Heydebreck A and Vingron M (2003) Analysis of microarray gene expression data. In: *Handbook of Statistical Genetics* (eds DJ Balding, M Bishop and C. Cannings). John Wiley & Sons, Chichester.
5. Kroll TC and Wolfl S (2002) Ranking: a closer look on globalisation methods for normalisation of expression arrays. *Nucleic Acids Res* **30(11)**: e50.
6. Gurok U, Steinhoff C, Lipkowitz B, Ropers HH, Scharff C and Nuber UA (2004) Gene expression changes in the course of neural progenitor cell differentiation. *J Neurosci* **24(26)**: 5982–6002.
7. Ihaka R and Gentleman R (1996) R: a language for data analysis and graphics. *J Computat Graph Stat* **5(3)**: 299–314.
8. Huber W, Von Heydebreck A, Sultmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18(Suppl1)**: S96–S104.
9. Durbin BP, Hardin JS, Hawkins DM and Rocke DM (2002) A variance stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18(Suppl1)**: S105–S110.
10. Cheadle C, Vawter MP, Freed WJ and Becker KG (2003) Analysis of microarray data using Z score transformation. *J Mol Diagnost* **5(2)**: 73–81
11. Shena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270(5235)**: 467–470.
12. DeRisi JL, Iyer VR and Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278(5338)**: 680–686.
13. Chen Y, Dougherty ER and Bittner ML (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt* **2**: 364–374.
14. Rocke DM and Durbin BP (2001) A model for measurement error for gene expression arrays. *J Computat Biol* **8(6)**: 557–569.
15. Golub TR, Slonim DK, Tamayo P, *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531–537.
16. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfeld CD, de la Chapelle A and Krahe R (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci USA* **98**: 1124–1129.
17. Tseng GC, Oh MK, Rohlin L, Liao JC and Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29(12)**: 2549–2557
18. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30(4)**: e15.
19. Workman C, Jensen L J, Jarmer H, *et al.* (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3(9)**: research0048.
20. Kerr MK, Martin M and Churchill GA (2000) Analysis of variance for gene expression microarray data. *J Computat Biol* **7**: 819–837.

21. Bolstad BM, Irizarry RA, Astrand M and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19(2)**: 185–193.
22. Gentleman RC, Carey VJ, Bates DM, *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Microarray data analysis: Differential gene expression

18

Stefanie Scheid and Rainer Spang

18.1 Introduction

Producing a useful list of differentially expressed genes from microarray data sounds like an easy task. However, our experience is different. Being caught in various traps ourselves, we will warn of the most dangerous pitfalls. Some 'roads to the list' are relatively safe to travel on, and we will point them out. The trip usually splits into two parts, and can be stopped after the first. At the beginning is explorative analysis which leads to candidate genes. Having arrived at this destination, you can transit to statistical analysis. This leads to quality measures, and allows you to distinguish between a reliable result and an unreliable one.

You can choose almost every generic statistical software for analysis. We recommend using the open source language R (1) together with the bioinformatics R-packages collected by the Bioconductor project (2). R is freely available at http://www.r-project.org, and Bioconductor at http://www.bioconductor.org. Throughout the chapter we will give you precise information for your 'trip to the gene list using R'.

18.2 Getting started

Repeat experiments

We start with a short field trip, that does not reach the outland of statistics yet. If you only have two microarrays and you want to compare them, you proceed as follows: Normalize the data, rank genes according to fold changes, pick from the top of the list as many genes as you like, and acknowledge in your paper that this is an explorative experiment. We believe that no claims on the statistical significance of your findings should be made.

Coming to repeated experiments, our recommendation is: the more the better. Of course, we will not do the benchwork for you and we will not pay for the microarrays. However, we want to raise two important points. First, repetitions are not only for the sake of *p*-values, they also improve the quality of your list. Without repetitions your list is much worse than it could be. Second, there is almost no statistical analysis

without repetitions of experiments. If you want to go beyond explorative research, you need repetitions.

Packing your bag

Collect your expression data in a matrix where columns correspond to samples and rows to genes. The first row contains sample labels and the first column contains gene labels. Separate all entries (labels and expression values) by tabs. Common data management tools like Microsoft© Excel and SPSS© offer to save data in tab-delimited formats. Say, the matrix is stored in a text file called data.tab. In R, read it into a matrix *X* by typing: X <– read.delim("data.tab"). In a second file, store additional information like condition labels for each sample (which sample is a wild-type, which sample is a mutant?). Similarly, read the reference values into a vector *y*.

Switch to additive scale

Expression levels can be signal intensities from oligonucleotide chips or 'red-green' ratios from cDNA chips. You will be used to compare expression levels by fold changes. This amounts to operating on a multiplicative scale. For statistical analysis, the multiplicative scale is not convenient and you should transform the data to the additive scale by taking for example logarithms. On the additive scale, differences between values correspond to ratios on the original scale. For example, the difference between two log expression values relates to the fold change on the original scale. Several normalization methods already return the data on the additive scale, for example the variance stabilization method by Huber *et al*. (3). If your data is not on an additive scale, type X <– log (X). If your data has negative values, refer to a text on expression data normalization, for example Huber *et al*. (4).

Be aware of your experimental design

Before choosing a scoring method that ranks your genes, consider the underlying experimental design. Ask yourself the following questions: first, how many conditions do I examine? If you compare samples of a wild-type to a mutant group, you need a *two-condition* score. If you compare samples of more than two conditions, you are in a *multi-condition* setting. The next question is: Are the samples *paired* or *unpaired?* A 'wild-type versus mutant' comparison is unpaired: the samples are not related to each other. A 'before and after treatment' comparison on the same mouse is paired: the gene expression of the *same* animal was measured before and after treatment, that is each sample in one condition has its counterpart in the second condition. Like the two channels in a 'red-green' experiment, these coupled data must not be separated from each other. *Figure 18.1* helps in deciding which design, and thus which scoring method, is appropriate.

Pick a score that fits your design

Within each experimental setting, you still have some selection of scores to pick from. Results might be quite different. What are the problems? To

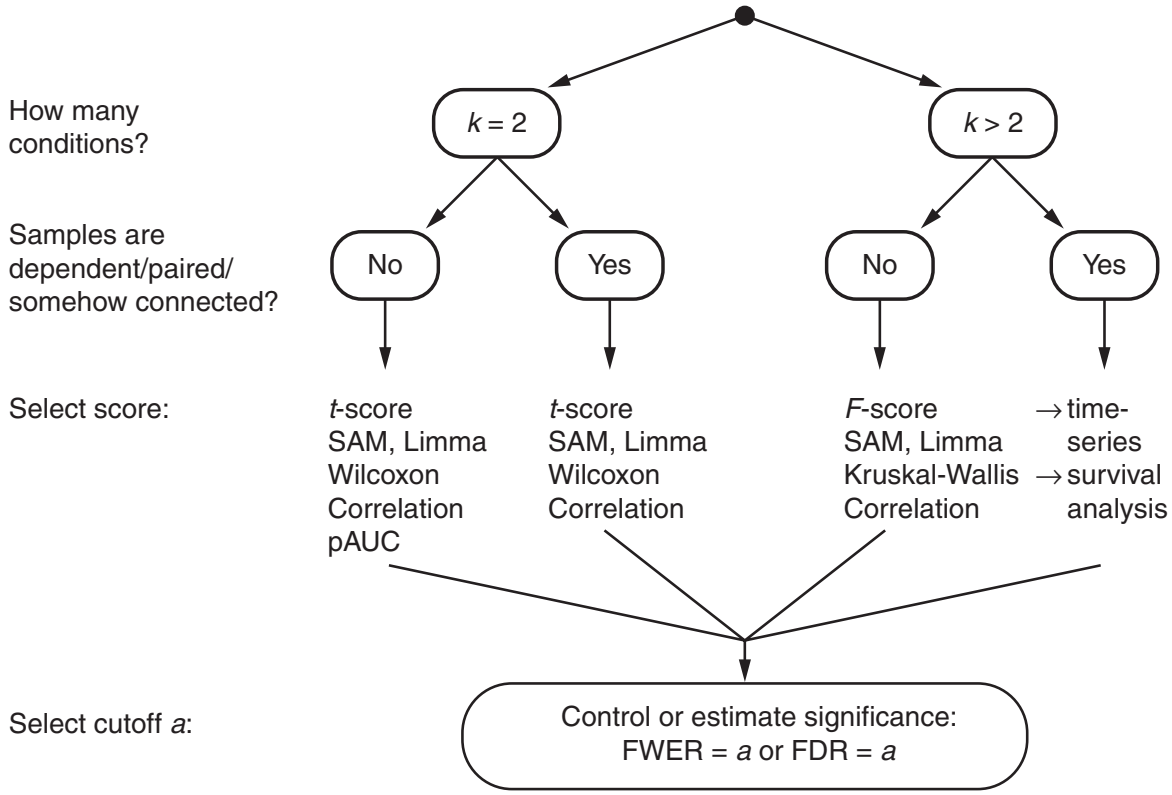


Figure 18.1.

The roadmap to the gene list.

start with, 'differentially expressed' is not a well-defined term. *Figure 18.2* displays gene expression values of two genes A and B in two conditions (light and dark gray). The y-axis shows expression levels on additive scale, and two horizontal lines indicate the mean expression in each condition. The distance between the two lines is larger for gene A than it is for gene B. Is gene A the better candidate? Gene B is rather constant in both groups, while gene A seems to be quite variable. Should we take variance into account? We can do so, using *t*-scores. The *t*-score is higher for gene B. Now, B is the better candidate.

You reach a different result with a third argument. For gene B, the expression values in the two conditions share only a small overlap. For gene A, this is not the case. Note that two dark gray values exceed the light gray mean. Do you feel that this is an important observation? You can formalize it using the pAUC-score. It is higher for gene B, consistent with the *t*-score but inconsistent with the fold change. By reflecting on *Figure 18.2*, we came up with three different perspectives on differential gene expression, and they led to conflicting results. Applied to a microarray experiment, the three scores result in different rankings of genes. Eventually, the rankings can be so diverse that you will be left with bad feelings when it comes to biological conclusions. The problem is what is a good ranking? We are not able to answer this question. In Section 18.3 we will discuss several scores and point out pitfalls. The final choice is upon

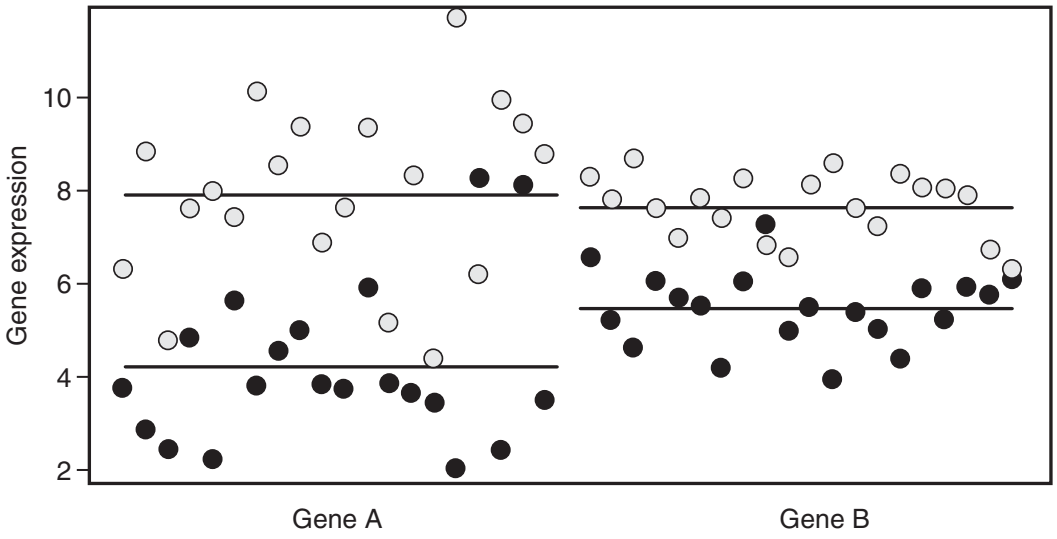


Figure 18.2.

Expression values of two genes A and B in two conditions. Light and dark circles refer to conditions. Horizontal lines mark the average expression in each condition.

you and depends on your definition of differential gene expression. We will not go into details, for deeper and more comprehensive reviews see Pan (5) and Troyanskaya *et al.* (6).

18.3 Explorative analysis

We will describe several scoring options to rank the genes on the array. Some of them will not work for you, since they do not match your experimental design. But we are confident that some methods in our selection will suit your purpose.

The design has two conditions and samples are unpaired

For example, you observe expression values in 10 wild-type and 10 mutant organisms. For each gene, separately compute the average expression in each condition and subtract the two averages. Differences on additive scale are ratios on original scale. Thus, the difference in means relates to the mean fold change on original scale and is therefore called *log ratio.* Log ratio scores are the simplest method to rank genes. In R, load for example package twilight by calling library(twilight) and type:

```
score1 <– twilight.pval(X,y,method="fc",paired=FALSE)
```

where "fc" stands for fold change.

The difference in means can be misleading if the gene-wise variances differ much. Consider again the example in *Figure 18.2*. Taking variability into account leads to the classical *t*-test score. Type:

```
score2 <– twilight.pval(X,y,method="t",paired=FALSE)
```

Consider fudge factors and be robust against outliers

Often, the *t*-score gives top ranks to genes with very small mean intensity differences if the expression values are almost constant in each condition. In terms of the applicability of the *t*-test this is no problem. As a biologist, however, you might not like it. To block these genes from top ranks, you can artificially enlarge low variances. First attempts to correct for low variances were made by Efron *et al*. (7) and Tusher *et al*. (8) by introducing a *fudge factor* that is added to each gene's variance estimate. Typically, the fudge factor is chosen from the set of variances of all genes. Like Efron *et al*. (7) you can choose this value manually. Choosing a very large fudge factor approximates differences of means, a zero factor results in the classical *t*-score, and probably an intermediate value will work best for you. An automatic choice is implemented in the software SAM (Significance Analysis of Microarrays), see Tusher *et al*. (8). The implementation in package twilight selects the median of all standard deviations as fudge factor. Type:

```
score3 <– twilight.pval(X,y,method="z",paired=FALSE)
```

Genes with small variances are one type of gene with high *t*-scores but little biological relevance; the other type are genes with outlying values. The Wilcoxon rank-sum score works on ranks instead of numerical values, and is less sensitive to outliers. The price for robustness is loss of information that was contained in the numerical values. In R, you can define a new function wilc that incorporates the base function wilcox.test. Note that we changed the Wilcoxon score such that it is distributed around 0. Assume that your condition labels are coded as 1 for the first condition and 0 for the second condition such that the label vector *y* is binary. Function wilc is then applied to each gene.

```
wilc <– function(expression,labels){
result <– wilcox.test(expression~labels,paired=FALSE)$statistic
return( result – sum(labels)*sum(1–labels)/2 )
}
score4 <– apply(X,1,wilc,labels=1–y)
```

A score that protects against both outliers and constant genes was introduced by Smyth (9) as *moderated t*-score. The idea is 'to borrow information across genes' with the intention to raise small variances and to shrink larger variances to an overall variance. The approach is implemented in the R package limma. The computation of limma scores in R needs three steps. First, fit a linear model for two conditions to each gene. Second, define the *contrast* you are interested in. That is the difference between the two condition averages. Third, compute moderated *t*-scores with function ebayes.

```
library(limma)
a <– lmFit(X,design=cbind(y,1–y))
b <– contrasts.fit(a,contrasts=matrix(c(1,–1),ncol=1))
score5 <– ebayes(b)$t
```

Alternatively: Score by separation

The expression values of gene A in *Figure 18.2* share a large numerical overlap whereas the overlap in the case of gene B is numerically smaller. Can we rank the genes based on how well the two conditions can be separated from each other without allowing too much overlap? The concept of separation is used by Pepe *et al.* (10) who suggest using pAUC-scores. A high pAUC-score indicates that the expression values in one condition are well, albeit not necessarily perfectly, separated from the values in the other condition. For small pAUC-values there is essentially no separation. Note that the concept of separability is different from comparing averages.

Currently, the pAUC-score is not implemented in R. You might want to read more about pAUC-values and their interpretation in Pepe *et al.* (10). To do the computations described in this publication you can use the following R function. It works for up-regulation only. For exploring down-regulation, reverse the class labels in your binary vector y to $1–y$.

```
pauc <– function(x,A,B){
  u <– sort(unique(A),decreasing=TRUE)
  t <– numeric(length(u))
  r <– numeric(length(u))
  for (i in 1:length(u)){
    t[i] <– sum(A[B==0]>u[i])/sum(1–B)
    r[i] <– sum(A[B==1]>u[i])/sum(B)
  }
  roc <– numeric(length(x))
  for (i in 1:length(x)){
    z <– which(t<=x[i])
    z <– z[length(z)]; roc[i] <– r[z]
  }
  return(roc)
}
```

You need to choose a false-positive rate, say 10%. Finally, calculate pAUC-scores by applying function integral to each gene:

```
integral <– function(a,b){integrate(pauc,0,0.1,A=a,B=b)$value}
score6.up     <– apply(X,1,integral,y)     # up-regulation
score6.down   <– apply(X,1,integral,1–y)   # down-regulation
```

For the combination of up- and down-regulation, compute the two variants and take the maximum of the two scores for each gene.

Example

Here is an application that illustrates different roads to different gene lists starting from the same data. The data set is by Golub *et al.* (11). Samples from 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML) are hybridized to Affymetrix© HU6800 microarrays coding for 7129 transcripts. The R package golubEsets contains a slightly transformed version of this data set in the variable golubMerge. We normalize with vsn (3) and compute absolute scores as described above.

```
library(golubEsets); library(vsn)
golubNorm <– vsn(exprs(golubMerge))
X <– exprs(golubNorm)
y <– as.numeric(golubMerge$ALL.AML)–1
```

Note that twilight.pval returns scores in the first column of a matrix called result. Genes are ordered by empirical *p*-values. To return the original order type:

```
genes <– rownames(X)
ttest <– twilight.pval(X,y,method="t",paired=FALSE)
score <– ttest$result[genes,1]
```

Table 18.1 displays the ranks of genes ranked highest by *t*-score and selected ranks. The first four columns contain ranks of *t*-like scores. Going from left to right, the scoring methods put more weight on the difference in means and less weight on a gene's variance (recall that the log ratio score is simply the difference in means). Ranks of genes with small differences and small variances increase going from *t*-test to log ratio, for example genes *CD33* and *MLP*. Ranks of genes with larger differences but large variances decrease, for example genes *FCER1G* and *SPI1*. The comparison between *t* and Wilcoxon ranks highlights which scores are confounded with outlying expression values, for example gene *DF*. The last column contains ranks of combined pAUC-scores which in this example lead to quite similar results.

The design has two conditions and samples are paired

If your data is paired you cannot travel on the roads described above. In many cases there are parallel tracks for you to use. Say you have an experiment with 10 patients before and after a treatment. In this setting you do not want to average the expression levels before and after treatment sepa-

Table 18.1. Example on data set of Golub *et al.* (11); ranks of selected genes resulting from different scoring methods.

Gene	*t*-score	Limma	Fudge	Log ratio	Wilcoxon	pAUC
MGSTl	1	1	3	21	5	27
DF	2	2	1	1	22	4
CD33	3	3	8	87	1	3
CST3	4	4	2	2	4	1
TCF3	5	5	11	58	3	5
MLP	6	7	22	118	8	28
CSTA	7	6	5	18	11	10
CTSD	8	8	27	144	7	12
SPTANl	9	9	19	62	12	17
CCND3	10	11	17	51	10	6
PSMA6	20	18	24	63	21	30
CD 63	30	30	46	120	29	158
FCER1G	40	38	23	29	49	164
SPI1	50	48	20	10	46	64
LTC4S	60	63	150	359	105	45

rately, instead you want to compute the difference in expression for each patient and the average over differences. This is a *mean difference* instead of a difference in means. In R, change the optional argument paired=FALSE to paired=TRUE to compute *t*-like or Wilcoxon scores. Note that the Wilcoxon function has to be changed such that the scores are distributed around 0.

```
wilc <– function(expression,labels){
   result <– wilcox.test(expression~labels,paired=TRUE)$statistic
   return( result – sum(labels)*(sum(labels)+1)/4 )
}
```

The design has more than two conditions and samples are independent

If you have more than two conditions, you can compare them pairwise which raises serious statistical difficulties. Instead, use the multi-condition equivalent of *t*-scores: *F*-scores. In R, use mt.teststat(. , test="f") in package multtest. The Wilcoxon equivalent are Kruskal-Wallis scores (kruskal. test). Note that multi-condition scores are sensitive to every gene whose expression is different in any one of the conditions. If the average expression in one condition deviates strongly from all others, the score is high.

A special case in multi-condition settings are dependent samples. Consider a set of 10 patients, each measured at five distinct time-points. This is a typical design in time-series or survival analysis. There are scores specialized for both, but this is beyond the scope of this chapter.

Find genes correlated to a reference gene

Like most biologists, you will have a favorite gene, and you want to find all those genes with expression values that correlate with the values of your pet gene. You might think that clustering all expression data and then looking up your gene in a red-green colored diagram is the best way to do this. We think it is not. In spite of the pitfalls already mentioned, we believe that a scoring approach is the safer way. Clustering is a mine field, and you only need to enter it if you do not want to focus on a pet gene but aim for a global view on the correlation structure in the data.

Consider the Golub *et al.* (11) example above. How can you identify genes with a high correlation to gene *MGST1?* Set the expression values of gene *MGSTl* as *reference vector* and apply either a correlation score based on numerical values (Pearson) or on ranks (Spearman) to each gene. With reference vector refvec, type:

```
score7 <– twilight.pval(x,refvec,method="pearson"
# or method="spearman"
```

Interpret your ranking

In principle, the presented scores have one feature in common: they are distributed around 0. A high positive value indicates up-regulation or corre-

lation, a high negative value indicates down-regulation or anti-correlation. To rank for differential gene expression in general, take absolute scores which is abs(score) in R.

If you want, you can stop here. You have a ranking of all genes on the array and you can pick genes from the top of the list as you like. This is exploratory data analysis, and there is nothing wrong with it. You might say: 'I can handle 50 genes. Hence, I take the top 50 genes.' This is a practical argument, and we do not see any problems with it, too. However, you might feel some discomfort once you realize that there are always 50 genes on top of a ranked list. Even in cases with no differentially expressed genes at all. For relief, you need to follow the track to statistical analysis.

18.4 Statistical analysis

A first visual inspection

Like in the previous section we start with something very simple. In this case a visual approach, which is part of the method SAM (Significance Analysis of Microarrays) by Tusher *et al.* (8). The Microsoft© Excel plug-in software is available at http://www-stat.stanford.edu/~tibs/SAM/. SAM computes scores for differential gene expression as given in the last section *(observed* scores). In addition, it simulates randomness by shuffling the patient labels and computes *expected* scores. Roughly speaking, these scores would occur if all genes in the experiment were non-induced. Plotting expected versus observed scores displays how much your data deviates from random noise. *Figure 18.3* shows SAM-plots for three situations: data sets with low, medium and high contents of differentially expressed genes. The diagonal line denotes the perfect agreement between your data and random data. The more the line deviates from the diagonal, the more evidence for differential expression you have. Up-regulated genes result in points above the line and down-regulated genes result in points below the line. The amount of points deviating from the diagonal gives you a first hint of the level of differential gene expression.

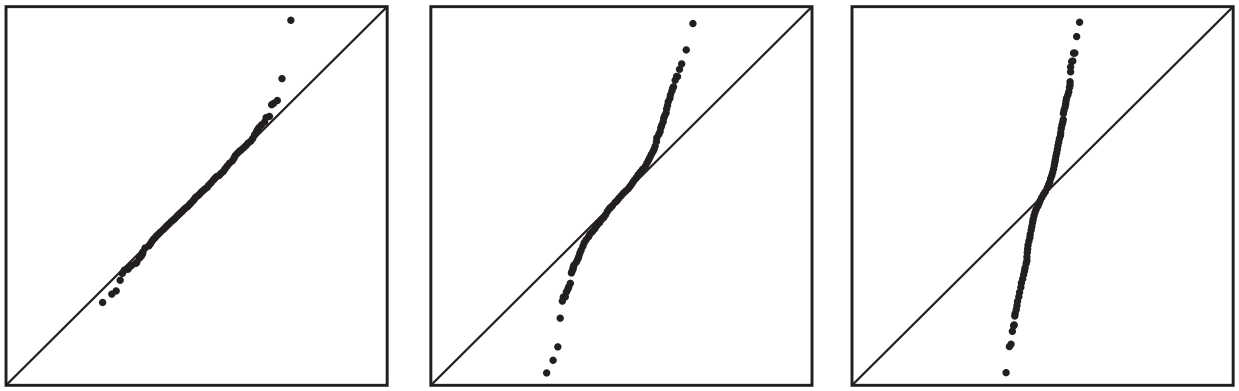


Figure 18.3.

SAM-plots for three simulated data sets with expected scores on the x-axes and observed scores on the y-axes. Diagonal lines denote equality between observed and expected scores. From left to right: Low, medium and high content of induced genes.

In R, use the command score <– twilight.pval(. . .) to compute observed and random scores. The values are stored in matrix score$result as observed and expected. For convenience, use plot(score,"scores") to get a SAM-like plot.

P-values

Now we proceed deeper into statistical analysis, describing classical main roads first. You want to filter out genes that are not differentially expressed. The most widely used filters are *p*-values. *P*-value filters consider false positives (genes that are not differentially expressed but passed the filter) as dirt, and they follow firm standards on how much pollution is tolerable. They will comply to these standards and prevent pollutants from entering your list. Be aware that the filter will not hesitate to absorb truly induced genes, too. This cannot be controlled directly. Fortunately, it is you who sets the standards, and thus you can indirectly calibrate the resulting list of genes. How do these standards express themselves in statistical analysis? In our experience, the most widely spread association with *p*-values is: 'They need to be below 0.05.' That is a standard of cleanness. What happens if we cut off all genes with $p \leq 0.05$? For a non-induced gene, the chance to survive this treatment is 5%. With 20 000 genes on the array and 19 500 of them non-induced, this leads to around 975 (= 19 500 × 0.05) false positives in the list. Is your standard of hygiene higher than that? In this case you can adjust the filter. Note that the simple computation above depends on the number of non-induced genes on the chip. If the chip was much smaller, say 500 genes and 250 of them non-induced, you only have to expect about a dozen false positives (250 × 0.05 = 12.5). That might be tolerable. In general, larger chips need stronger filters to achieve the same standards for clean gene lists.

As genes are not independent of each other but connected through pathways and coregulation, we recommend using *empirical p-values* that are derived by permuting the condition label vector. The empirical *p*-value is the percentage of random scores that exceed the original score. To compute empirical *p*-values from, say, 10 000 permutations, type:

```
score <– twilight.pval(X,y,…, B=10 000)
pvalue <– score$result$pvalue
```

Control contamination

By reducing the cutoff, for example to $p \leq 0.01$, you can decrease the number of false positives even for the large chip. However, you can do better using more efficient and more complicated methods of multiple testing. 'Doing better' means use a tighter filter complying to the same standard. A classical but very rigid standard is the *family-wise error rate* (FWER). Essentially, it does not tolerate false positives in your list at all. While it cannot completely ensure a perfectly clean list, it can do so at least with a high probability. Hence if you set the FWER to 0.05, there is only a 5% chance that a single false-positive gene could sneak through the filter into the list. For a good introduction into the FWER see Dudoit *et al.* (12). In R,

function mt.rawp2adjp in package multtest offers several procedures to build FWER filters. For the classical Bonferroni-Holm procedure, type:

```
library(multtest)
FWER <– mt.rawp2adjp(pvalue,proc="Holm")
```

Be less restrictive

FWER based standards are about the highest available. But it might happen that not a single gene passes the filter. While this is a 'spotless list', it is probably not what you have hoped for. If you are willing to tolerate some false positives in your list, say 5%, you can do better using the *false discovery rate* (FDR) introduced by Benjamini and Hochberg (13). In R type:

```
FDR <– mt.rawp2adjp(pvalue,proc="BH")
```

Don't let the statistician leave the biggest challenges with you, instead challenge him

You have a short list, and start screening for a gene which you know is up-regulated. You have confirmed this in single gene assays several times, and it can also be found in the literature. However, the gene is not in the list. Did the microarray experiment disprove previous results? Or did you learn better not to believe in microarrays? We think that the most likely problem is the *p*-value filter. It was too harsh. You start looking for your pet gene in the complete ranked list and you find it up-regulated but somewhat below the cutoff line. Is it allowed to enlarge the list and include all genes down to this gene? Yes, but it comes at a price: the list will be contaminated by more false positives. How badly contaminated is it? This question lies not on the main tracks of statistics, and it is controversial as to whether it is a good one. The main track is that you first define the standard, then the statistician produces the list of genes, and you are left with the challenge to interpret it. Extending the list, as suggested above, assigns jobs differently. Now, you define the list and the statistician is left with the challenge to estimate its degree of contamination. Recently, this challenge was accepted by parts of the statistical community. And not surprisingly, the first software tools like SAM became very popular.

Back to the question: how badly contaminated is the extended list? Or, statistically, what is the expected proportion of false positives in it? Storey (14) gave a first answer by introducing the *q*-value. The *q*-value of a gene is, roughly, an estimated FDR of the list that includes all genes up to this gene. The main difference between the Benjamini-Hochberg ideas and the Storey ideas is that of controlling the FDR versus estimating it. Or in easier words, it is the difference between whether you need to define a tolerable FDR and the computer is producing the list or vice versa.

You want to talk about single genes and not lists of genes

Now you have a list of genes and you know that about 10% of them are false positives. You want to know which ones. This is of course not possible. On the other hand, the gene on top of the list is less likely to be a

false positive than the one on the cutoff line. Efron *et al*. (7) introduced the local FDR (LFDR) which is the probability that a gene is a false positive. Note, that the *p*-value is a different probability. For computing the LFDR, apply function twilight which is based on the estimation procedure described in Scheid and Spang (15).

```
score <– twilight.pval(…)
LFDR <– twilight(score)
```

Plot the LFDR over the range of *p*-values by calling plot(LFDR, "fdr"). Following the LFDR from low *p*-values to high *p*-values, you get an impression of the level of differential gene expression in your experiment, and whether there is a twilight zone where clear differential expression fades into clear non-differential expression.

18.5 Final remarks

Following the roads of statistical analysis to the end, you get an informative and statistically valid first description of your microarray experiment, and can go on to other topics like classification and prediction. We guided you that far by recommending methods and software tools that we are most experienced with. We traveled mostly on main roads of microarray analysis. You might want to try different methods. We recommend that you explore the Bioconductor collection yourself, and find new paths of analysis.

References

1. Ihaka R and Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* **5**: 299–314.
2. Gentleman R, Carey V, Bates D, *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5(10)**: R80.
3. Huber W, von Heydebreck A, Sültmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18(Suppl. 1)**: S96–S104.
4. Huber W, von Heydebreck A and Vingron M (2003) Analysis of microarray gene expression data. In: *Handbook of Statistical Genetics,* 2nd Edn. (eds M Bishop *et al.)*. John Wiley & Sons, Chichester.
5. Pan W (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18(4)**: 546–554.
6. Troyanskaya OG, Garber ME, Brown PO, Botstein D and Altman RB (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18(11)**: 1454–1461.
7. Efron B, Tibshirani R, Storey JD and Tusher VG (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* **96(456)**: 1151–1160.
8. Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98(9)**: 5116–5121.
9. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Gen Mol Biol* **3(1)**: Article 3.

10. Pepe MS, Longton G, Anderson GL and Schummer M (2003) Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**: 133–142.
11. Golub TR, Slonim DK, Tamayo P, *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531–537.
12. Dudoit S, Shaffer JP and Boldrick JC (2002) Multiple hypothesis testing in microarray experiments. *U.C. Berkeley Division of Biostatistics Working Paper Series* **110**.
13. Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**: 289–300.
14. Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat* **31(6)**: 2013–2035.
15. Scheid S and Spang R (2004) A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE Transactions on Computational Biology and Bioinformatics* **1(3)**: 98–108.

Clustering and classification methods for gene expression data analysis

19

Elizabeth Garrett-Mayer and Giovanni Parmigiani

19.1 Introduction

Efficient use of the large data sets generated by gene expression microarray experiments requires computerized data analysis approaches (1, 2). In this chapter we briefly describe and illustrate two broad families of commonly used data analysis methods: class discovery and class prediction methods. Class discovery, also referred to as clustering or unsupervised learning, has the goal of partitioning a set of objects (either the genes or the samples) into groups that are relatively similar, in the sense that objects in the same group are more alike than objects in different groups (3, 4). A typical application is to generate hypotheses about novel disease subtypes (5, 6). Class prediction, also referred to as classification or supervised learning, has the goal of determining whether an object (usually a sample, but sometimes a gene) belongs to a certain class (7, 8). A typical application is classification of patients into existing disease subtypes or prognostic classes (9, 10) using gene expression information.

In our discussion, 'sample' refers generically to any type of biological material that is processed and hybridized to a chip. For example, in a study of breast cancers, the samples could represent RNA isolated from breast cancer tissues biopsied from a group of women. 'Gene' is used loosely to refer to the features on the arrays, such as sequences from genes or ESTs, single oligonucleotides in Agilent arrays, oligonucleotide sets in Affymetrix arrays and so forth. 'Object' refers to the entity being clustered, and can be either a gene or a sample, as the same algorithms can often be applied symmetrically to both. 'Attribute' is any feature of the object being clustered. If we cluster samples, genes are typically attributes, and vice versa. 'Phenotype' refers to any clinical or biological characteristic of a sample or the person or organism from which the sample is derived, such as disease subtype, age, gender, or time to disease progression.

To demonstrate the clustering methods in this chapter, we use a gene expression microarray dataset published by Hedenfalk and colleagues (11) and including samples from 22 breast cancers, of which seven are from

patients with known BRCA 1 mutations, eight from patients with known BRCA2 mutations, and seven are sporadic. Complementary DNA (cDNA) labeled with Cy3 or Cy5 was obtained from each tumor sample and hybridized to two channel cDNA arrays which included spots for 3226 genes and ESTs. The reference sample was cell line MCF-10, a nontumorigenic breast cell line. Data from this study is available at http://www.nhgri.nih.gov/DIR/Microarray.

Statistical computing environments typically offer a rich set of alternatives for clustering and classification. In particular the free and open source computing environment R (12) and the associated Bioconductor (13) project cover most standard tools, a wide variety of developmental tools and offer the flexibility for implementing custom solutions. A range of free and open source tools can be accessed via the website www. arraybook.org. The site http://ihome.cuhk.edu.hk/~b400559/arraysoft.html maintains a catalog of both free and commercial microarray data analysis software.

19.2 Clustering

Clustering techniques can be used in microarray analysis to (i) facilitate visual display and interpretation of experimental results, and (ii) suggest the presence of subgroups of objects (genes or samples) that behave similarly. The input of a cluster analysis are the gene expression values of the samples in an experiment, with no additional phenotype information. Depending on the approach, the output can be a list of subgroups, or a visualization that simplifies manually establishing subgroups. In some applications, unsupervised methods are used even though phenotype information is available. The goal is often to see how the clusters of samples that arise from an unsupervised approach compare to the known phenotypes.

Distance and similarity

To determine which objects cluster together, we must have a way of measuring how similar, or dissimilar any two of them are. Most clustering approaches will allow as input a matrix whose entries measure similarity, or dissimilarity, between each pair of objects. Choosing this measure is one of the most critical, yet often underappreciated, aspects of a cluster analysis. Different measures reflect different goals, and thus can have a strong influence on the resulting clusters. Here we discuss in detail three: the correlation coefficient, which will bring together objects whose patterns of change are similar; the Euclidean distance, which will bring together objects whose absolute expressions are similar, and the uncentered correlation, which achieves a compromise between the previous two.

The Pearson correlation coefficient measures the strength of a linear association between the expression levels of objects. In the case of genes *j* and *k,* it is defined by

$$\rho_{jk} = \frac{\sum_{s=1}^{S}(x_{sj} - \bar{x}_j)(x_{sk} - \bar{x}_k)}{\sqrt{\sum_{s=1}^{S}(x_{sj} - \bar{x}_j)^2 \sum_{s=1}^{S}(x_{sk} - \bar{x}_k)^2}} \tag{19.1}$$

where x_{sj} is the gene expression for gene j in sample s and $\bar{x}_j$ is the average gene expression of gene j across all samples. A symmetric definition applies to the correlation between samples. The correlation takes values ranging from –1 (perfect negative correlation) to 1 (perfect positive correlation). A correlation of 0 means that there is no linear relationship between the two genes. For analyses that require positive similarity matrices, it is common to use the absolute value of the correlation with the rationale that high negative and high positive correlations both may imply an underlying common mechanism. The correlation coefficient is unitless, but is sensitive to nonlinear transformation of the data, such as the logarithm. For non-linear relationships, the correlation coefficient may not adequately describe similarity. Another drawback is that it may be sensitive to noise.

The Euclidean distance measures geometric distance between two objects. In the case of genes j and k, it is defined by

$$d_{jk} = \sqrt{\sum_{s=1}^{S} (x_{sj} - x_{sk})^2}. \tag{19.2}$$

A symmetric definition applies to the correlation between samples. It takes values from 0 to ∞ and it retains the units of the input gene expression measurements. It grows with the number of samples included in the dataset.

The uncentered correlation (14) is similar to the Pearson correlation but is evaluated without centering:

$$e_{jk} = \frac{\sum_{s=1}^{S} x_{sj} x_{sk}}{\sqrt{\sum_{s=1}^{S} x_{sj}^2 \sum_{s=1}^{S} x_{sk}^2}}. \tag{19.3}$$

As the Pearson correlation, this is unitless, but is sensitive to absolute magnitudes as the Euclidean distance. As a result it will be less likely to be influenced by genes whose variation is mostly noise.

For a summary of other distance and similarity metrics, see (15).

Hierarchical clustering

Hierarchical clustering is used to partition objects into a series of nested clusters (5, 6), by contrast with approaches that find a single partition (16). To illustrate, a hierarchical clustering analysis of both genes and samples in the Hedenfalk data is shown in *Figure 19.1*, along with a gray scale image of gene expression levels. The similarity used is the uncentered correlation. The hierarchy of clusters of samples is displayed using a tree-like structure called a dendrogram. Dendrograms join objects, or clusters of objects, to form increasingly large clusters. The height at which two clusters are joined represents how similar they are, with low heights representing high similarity. Samples in *Figure 19.1* are labeled by their type (BRCA1, BRCA2, or sporadic), though these types are not used in constructing the dendrogram.

There are two kinds of hierarchical clustering approaches: agglomerative and divisive. The agglomerative approach begins by assuming that each object belongs to its own separate cluster. At the first step, the two most

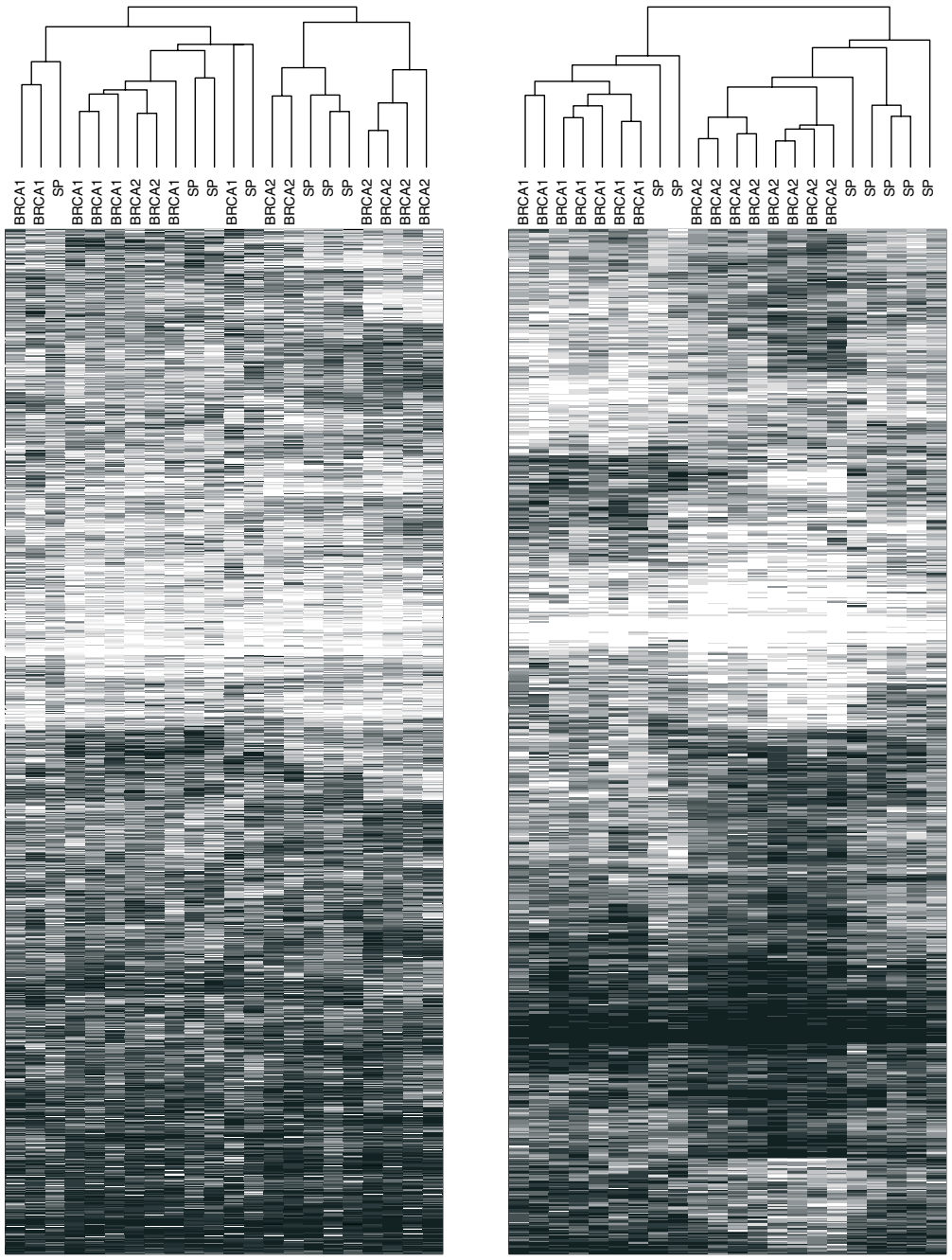


Figure 19.1.

Hierarchical cluster analysis of the Hedenfalk breast cancer data. The gray scale image represents gene expression levels, with levels lower than the reference represented by white to light gray and levels higher than the reference represented by medium gray to black. The left panel includes all samples and genes. The right panel includes all samples and the top 25% genes most strongly associated with the presence of BRCA1 and BRCA2 mutations. The dendrograms for genes have been omitted.

similar objects are combined to form a new cluster. Then the next most similar clusters or objects are combined and so forth. This is a bottom-up approach in the sense that the clustering starts at the bottom of the dendrogram of *Figure 19.1* and works its way up until all objects belong to one cluster. As part of the agglomerative approach, it is necessary to specify a linkage method, that is a way of defining similarity of clusters based on similarities of cluster members. Some of the commonly used linkage methods are single, average, and complete in which clusters are linked based on the similarity of the closest members, the average similarity, and the similarity of the furthest members.

The divisive approach works from the top of the dendrogram, where all objects belong to one cluster. At the first step, it finds the best division of the objects so that there is the highest similarity among objects within clusters and the most dissimilarity between clusters. This process continues, where the best cluster partition is chosen at each step until all objects are in their own clusters. Details of hierarchical clustering can be found in (4).

An important consideration when applying or interpreting hierarchical clustering results is that there is not a unique dendrogram for a given hierarchical clustering result. For each split in a dendrogram, it is arbitrary which branch is drawn to the right or left, and users need to specify criteria for this choice. As such, many dendrograms can be drawn for a given hierarchical clustering result and closeness of objects should be judged based on the height at which they are joined, rather than their ordering in the dendrogram.

Preselection of genes can significantly affect clustering of samples and vice versa. Selecting genes that show at least a certain amount of variation across samples is useful to reduce the sensitivity of clustering results to noise variation. Selecting genes whose variation is associated with a phenotype of interest is also common, though when that is done the correspondence of clusters to phenotype cannot be invoked as validation of the clustering results, as the correspondence will be inflated by the preselection. To illustrate, compare the left panel of *Figure 19.1*, which includes all genes in the experiment, to the right panel, where only the top 25% of genes associated with the BRCA types are included. The dendrogram on the left has short branch links and cascading patterns, both of which weaken the case for the existence of clusters. None of the main partitions has any relation to the BRCA type. On the right, the branch links at the top are longer and there is some evidence of two major clusters, which separate well the BRCA1 from the BRCA2 cases. While in general a correspondence between clusters found by unsupervised analyses and sample phenotypes can be taken as independent supporting evidence of the existence of clusters of biological significance, in this case this argument would be circular, because the sample phenotypes were used in selecting the genes for clustering.

K-means clustering and self-organizing maps

K-means clustering (17) partitions objects into groups that have little variability within clusters and large variability across clusters. The user is required to specify the number k of clusters *a priori*. Estimation is iterative, starting with a random allocation of objects to clusters, re-allocating to

minimize distance to the estimated 'centroids' of the clusters, and stopping when no improvements can be made. The centroid is the point whose attributes take the mean expression level of the objects in the clusters. K-medoids clustering is similar, except that the center of the clusters is defined by 'medoids', similar to centroids, but based on medians (4). Specification of *k* can be difficult, though there are ways of gaining insight into the appropriate number of clusters, such as using principal components analysis. A closely related approach is that of self-organizing maps (7, 15, 18), now common in gene expression data (16).

Principal Components Analysis and Multi-Dimensional Scaling

Principal Components Analysis (PCA) (19–21) and Multidimensional Scaling (MDS) are techniques whose goal is to reduce the dimensionality of data to facilitate visualization and additional analysis. They are often used as a preliminary step to the clustering of large data sets and are commonly applied to gene expression data (22–28).

PCA creates summary attributes, or 'components', that are weighted averages of the original attributes, are uncorrelated to each other, and are such that most of the variability in the data is concentrated in few components. During the estimation process, as many components as there are attributes are calculated. Users select a small number, chosen to retain a sufficient fraction of the variability. These are often plotted to visually search for clusters. A strength of PCA is that redundant information is represented in a single component, while a drawback is that the components may lack clear biological interpretations.

The first three PC's for the Hedenfalk data are shown in *Figure 19.2*. Here, instead of having to visualize thousands of genes per sample, we use three weighted averages of genes. Together, they describe 38% of the variability in the data. The samples appear to cluster in subgroups. When phenotype information is available, one can check putative subgroups against the phenotype information, or gauge how the variability in expression relates to the variability in phenotypes. For example, in the top-left panel, the sporadic samples tend to have high values for component 2 and relatively low values for component 1. BRCA2 samples are distributed differently with most having either very low values for component 2 or high values for both components 1 and 2. The four areas in the plot created by the two intersecting lines are discriminating between different BRCA types. The results for components 1 versus 3 and 2 versus 3 also show some clustering, though these are not as clearly related to BRCA types.

MDS starts from a distance matrix between objects and finds the locations of these objects in a low dimensional space that best preserves the original distances. For example, given objects in three dimensions, MDS may find the two-dimensional map of these objects that is most faithful to the original three-dimensional distances. The result is similar to the PCA result: we have summary variables, the coordinates of the map, that describe a large fraction of the variability in the gene expression measures, and that can be visually inspected to identify clusters. Two examples of MDS as applied to gene expression data can be found in Khan *et al*. and Bittner *et al*. (29, 30).

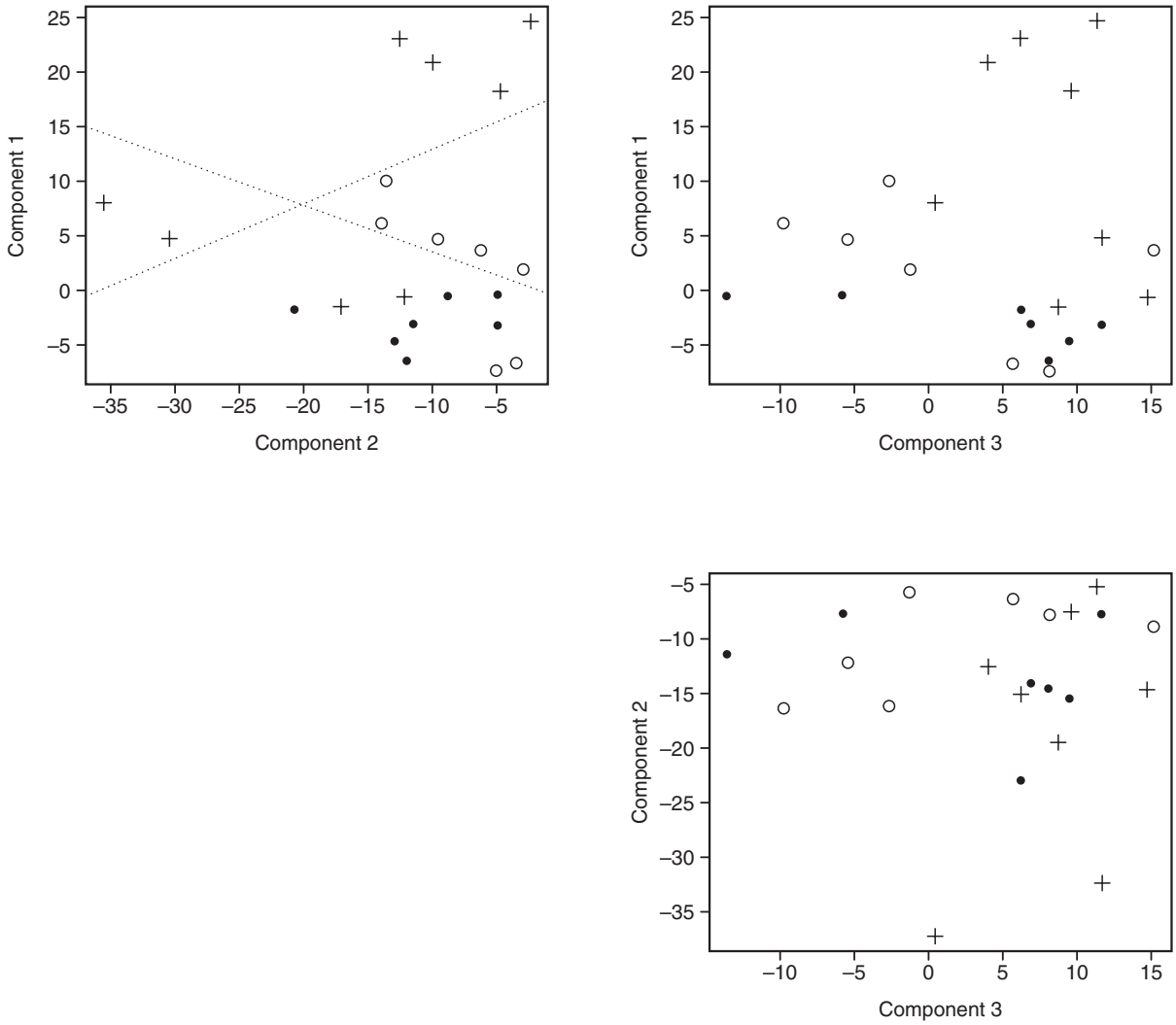


Figure 19.2.

The first three principal components of the Hedenfalk breast cancer data. Open circles indicate sporadic samples, closed circles indicate BRCA1 samples, and plus symbols indicate BRCA2 samples. Dotted lines in the plot of component 1 versus component 2 distinguish the three types.

Limitations of cluster analysis

Clustering techniques for high dimensional data are exploratory. Their strength is in providing rough maps and suggesting directions for further study. Substantial additional work is necessary to provide context and meaning to groups found by automated algorithms. This includes cross-referencing of existing knowledge about genes and samples as well as additional biological validation.

Clustering results are sensitive to a variety of user-specified inputs. The clustering of a large and complex set of objects can, like arranging books in a collection, be planned in different ways depending on the goals. From

this perspective, good clustering tools are responsive to users' choices, not insensitive to them, and sensitivity to input is a necessity of cluster analysis rather than a weakness. This also means, however, that use of a clustering algorithm without knowledge of its workings, the meaning of inputs, and their relationship to the biological questions of interest is likely to yield misleading results.

Clustering results are generally sensitive to small variations in the samples and the genes chosen and to outlying observations. This means that a number of the data-analytic decisions made during normalization, filtering, data transformations, and so forth will have an effect on results. When conclusions drawn from clustering go beyond simple data visualization, it is important to provide accurate assessments of the uncertainty associated with the clusters found. Uncertainty from sampling and outliers can be addressed within model-based approaches (31) or alternatively using resampling techniques (32–34). The consequences of choosing among plausible alternative transformations, normalizations, and filtering should be addressed by sensitivity analysis, that is by repeating the analysis and reporting conclusions that are consistent across analyses.

19.3 Classification

Classification techniques can be used in microarray analysis to predict sample phenotypes based on gene expression patterns. While novel and microarray-specific classification tools are constantly being developed, the existing body of pattern recognition and prediction algorithms provide effective tools (35). Dudoit and colleagues (36) offer a practical comparison of methods for the classification of tumors using gene expression data. Relevant tools from the statistical modeling tradition include: discriminant analysis (37), including linear, logistic, and more flexible discrimination techniques; tree-based algorithms, such as classification and regression trees (CART) by Breiman *et al*. (38) and variants; generalized additive models (39); and neural networks (7, 40, 41). Appropriate versions of these methods can be used for both classification and prediction of quantitative responses such as continuous measures of disease aggressiveness. Some of these methods are briefly reviewed here.

Dimension reduction

Because of the large number of genes that can be used as potential predictors, it is useful to preselect a subset of genes, or composite variables, likely to be predictive and then investigate in depth the relationship between these and the phenotype of interest. For example, genes with nearly constant expression across all samples can be eliminated. Additional screening can be based on measures of marginal association, such as the ratio of within-group variation to between-group variation, or the measure used in Slonim *et al*. (42), though these can miss important genes that act in concert with others but have no strong marginal effects.

Parsimonious representations of the data may be identified when there is knowledge of important pathways that can be used to manually construct

new and more highly explanatory variables. When such knowledge is not available we need to apply discovery techniques such as those described earlier; for example, the centroids of clusters or the variables identified by PCA can be used as predictors. Composite variables that are easily measurable and interpretable in terms of the original gene expression are generally preferable. Automatic approaches for preclustering variables before classification are also useful (43).

Evaluation of classifiers

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings (44, 45).

The most satisfactory approaches to validation require the use of data other than those used to develop the classifier. When only a single study is available, this can often be achieved by setting aside samples for validation purposes, as illustrated by Dudoit *et al*. (36). Statistical validation of probabilistic models (46) should focus on both refinement, that is, the ability of the classifier to discriminate between classes, and calibration, that is, the correspondence between the fraction predicted and the fraction observed in the validation sample.

An alternative to setting aside samples for validation is the so-called cross-validation. For example, *K*-fold cross-validation consists of splitting the data in *K* subsets, and training the classifier *K* times, setting aside each subset in turn for validation. The average classification rates in the *K* analyses is then an unbiased estimate of the correct classification rate (47).

A potentially serious mistake is to evaluate classifiers on the same data that were used for training. When the number of predictors is very large, a relatively large number of predictors will appear to be highly correlated with the phenotype of interest as a result of the random variation present in the data. These spurious predictors have no biological foundation and do not generally reproduce outside of the sample studied. As a result, evaluation of classifiers on training data tends to give overly optimistic assessments of validity. In plausible settings, classifiers can appear to have a near perfect classification ability in the training set without having any biological relation with phenotype (48). All aspects of learning a classifier need to be properly cross-validated to avoid inflated estimates of performance.

Prediction Analysis of Microarrays (PAM)

A straightforward approach to classification is the nearest centroid classifier. This computes, for each class, a centroid given by the average expression levels of the samples in the class, and then assigns new samples to the class whose centroid is nearest. This approach is similar to *k*-means clustering except clusters are now replaced by known classes. With a large number of genes this algorithm can be sensitive to noise. A recent enhancement uses shrinkage: for each gene, differences between class

centroids are set to zero if they are deemed likely to be due to chance. This approach is implemented in the Prediction Analysis of Microarray, or PAM (49), software. Shrinkage is controlled by a threshold below which differences are considered noise. Genes that show no difference above the noise level are removed. A threshold can be chosen by cross-validation, as shown in *Figure 19.3* for the Hendefalk data. High thresholds, on the right, include few genes, and lead to classifiers that are prone to errors. As the threshold is decreased more genes are included and estimated classification errors decrease, until they reach a bottom and start climbing again as a result of noise genes – a phenomenon known as overfitting.

Top-scoring pairs

Another simple and very effective tool is the top-scoring pair(s), or TSP, classifier (50). In a two-class classification, this looks for pairs of genes such that gene 1 is greater than gene 2 in class A and smaller in class B. This

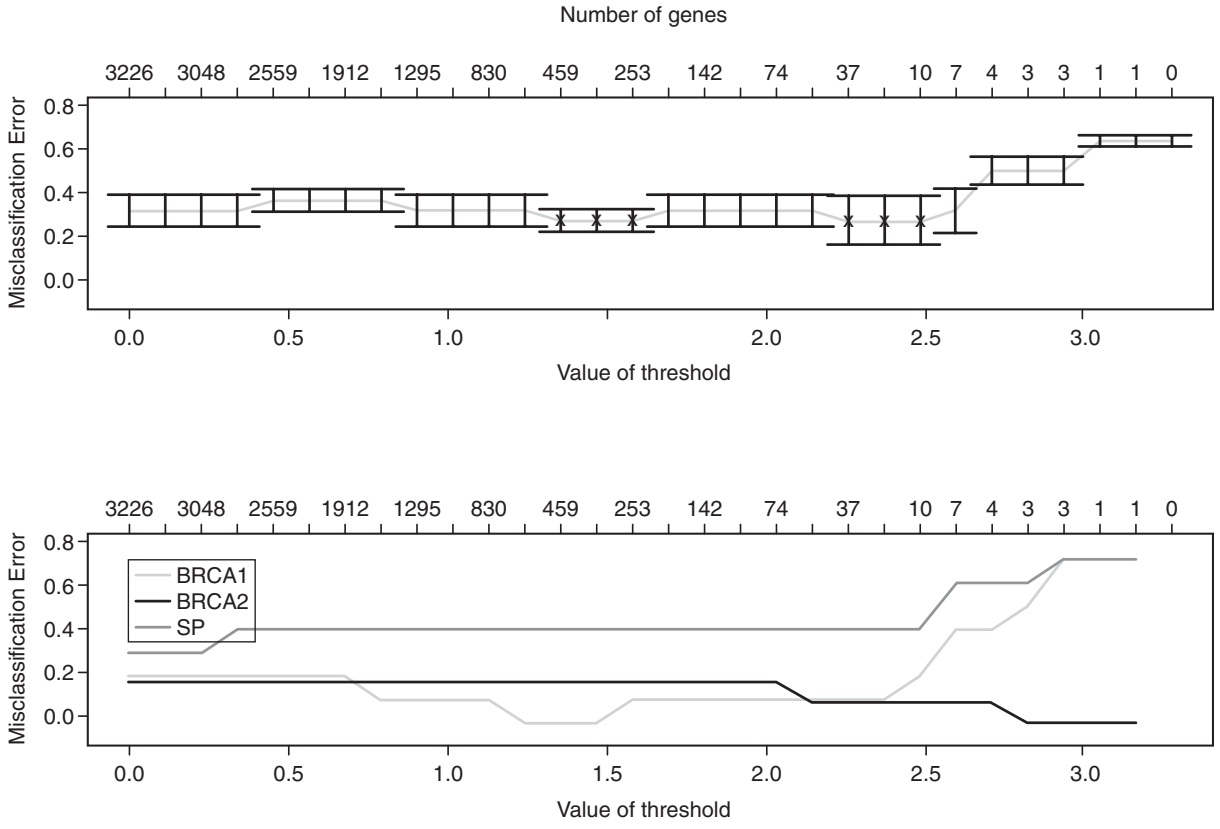


Figure 19.3.

Misclassification error of PAM classifiers on the Hedenfalk breast cancer data. The top panel shows the overall classification error as a function of the threshold used to set centroid differences to zero. Classifiers on the right have a higher threshold and a more parsimonious use of genes. As the number of genes increases, the error rate decreases until about 10 genes when the effects of overfitting offset the additional predictive ability of adding genes, and the error rates starts increasing. The bottom panel shows the classification error by class.

handles effectively issues of normalization as the pair provides an internal control and is likely to give generalizable results. TSP classifiers are transparent and interpretable and provide specific hypotheses for follow-up studies. In cancer data the TSP classifier achieves prediction rates that are as high as those of alternative approaches which use considerably more genes and complex procedures (50).

Nearest-neighbor classifiers

Nearest-neighbors classifiers (51), assign samples to classes by matching the gene expression profile to that of samples whose class is known. A simple implementation is to choose a rule for finding the *k* nearest neighbors and then deciding the classification by majority vote. Nearest-neighbor classifiers are robust, simple to interpret and implement, and do not require, although they may benefit from, preliminary dimension reduction. Nearest-neighbor algorithms are also used in several statistical software packages for imputation of missing data.

Support vector machines

Support vector machines (SVMs) (52) seek cuts of the data that separate classes effectively, that is by large gaps. Technically, SVMs operate by finding a hypersurface in the space of gene expression profiles, that will split the groups so that there is the largest distance between the hypersurface and the nearest of the points in the groups. More flexible implementations allow for imperfect separation of groups. See Burges (53) and Christianini and Shawe-Taylor (54) for details of SVMs and generalizations, while Lee and Lee (55) and Brown *et al*. (56) give examples of analysis of gene expression data using SVMs.

Discriminant analysis

Discriminant analysis (57) and its derivatives are approaches for optimally partitioning a space of expression profiles into subsets that are highly predictive of the phenotype of interest, for example by maximizing the ratio of between-classes variance to within-class variance. Ripley (7) and Everitt (21) give details, while Hastie *et al*. (58) discusses flexible extensions of discriminant analysis (FDA) and Li and Yang (59) provides a discussion of discriminant analysis in the context of gene expression array data.

Classification trees

Classification trees recursively partition the space of expression profiles into subsets that are highly predictive of the phenotype of interest (38). They are robust, easy-to-use, and can automatically sift large data sets, identifying important patterns and relationships. No prescreening of the genes is required. The resulting predictive models can be displayed using intuitive graphical representations. An example in which classification trees have been applied to gene expression data can be found in Zhang and Yu (60).

Regression-based approaches

Linear models, generalized linear models, generalized additive models and the associated variable selection strategies provide standard tools for selecting useful subsets of genes and developing probabilistic classifiers. A limitation of these techniques is that they cannot generally handle more genes than there are samples. This can be circumvented using forward selection approaches that progressively add genes to the classifier. Recent, more accurate approaches are based on the so-called stochastic search methods (61), that generate a sample of plausible subsets of explanatory variables. The selected subsets are then subjected to additional scrutiny to determine the most appropriate classification algorithm. A combination of stochastic search with principal component analysis and other orthogonalization techniques has proven effective in high-dimensional problems (62, 63), and has recently been employed in microarray data analysis (23).

Probabilistic model-based classification

Model-based classification is based on the specification of a probability distribution that describes the variability of the expression values. Typically, this is a mixture model, in which mixture components represent known classes (64). Model-based approaches are computation-intensive and can be sensitive to assumptions made about the probability model, but can provide a solid formal framework for the evaluation of many sources of uncertainty, and for assessing the probability of a sample belonging to a class.

19.4 Summary

A wide range of alternative approaches for clustering and classification of gene expression data are available. While differences in efficiency do exist, none of the well-established approaches is uniformly superior to others. Choosing an approach requires consideration of the goals of the analysis, the background knowledge, and the specific experimental constraints. The quality of an algorithm is important, but is not in itself a guarantee of the quality of a specific data analysis. Uncertainty, sensitivity analysis and, in the case of classifiers, external validation or cross-validation should be used to support the legitimacy of results of microarray data analyses.

Acknowledgment

Work of Parmigiani partly supported by NSF grant NSF034211 and NCI grant 5P30 CA06973-39.

References

1. Speed TP (ed.) (2003) *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall, London.

2. Parmigiani G, Garrett ES, Irizarry RA and Zeger SL (2003) The analysis of gene expression data: an overview of methods and software. In: *The Analysis of Gene Expression Data: Methods and Software,* Parmigiani G, Garrett ES, Irizarry RA, Zeger SL, (eds), Springer, New York, pp. 1–45.
3. Hartigan JA (1975) *Clustering Algorithms*. Wiley, New York.
4. Kaufmann L and Rousseeuw PJ (1990) *Finding Groups in Data: An introduction to Cluster Analysis*. Wiley, New York.
5. Perou CM, Jeffrey SS, van de Rijn M, *et al*. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA* **96(16)**: 9212–9217.
6. Bullinger L, Dohner K, Bair E, *et al*. (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med* **350**: 1605–1616.
7. Ripley BD (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
8. Hastie T, Tibshirani R and Friedman J (2003) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.
9. Golub TR, Slonim DK, Tamayo P, *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531–537.
10. Ross DT, Scherf U, Eisen M B, *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**: 227–235.
11. Hedenfalk I, Duggan D, Chen Y, *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med* **344(8)**: 539–548.
12. Ihaka R and Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* **5**: 299–314.
13. Gentleman R (2003) BioConductor: open source software for bioinformatics. http://www.bioconductor.org.
14. Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Nat Acad Sci USA* **95**: 14863–14868.
15. Gordon AD (1999) *Classification*. Chapman and Hall/CRC, New York.
16. Tamayo P, Slonim D, Mesirov J, *et al*. (1999) Interpreting gene expression with self organizing maps: methods and application to hematopoietic differentiation. *Proc Nat Acad Sci USA* **96**: 2907–2912.
17. Hartigan JA and Wong MA (1979) A k-means clustering algorithm. *Appl Stat* **28**: 100–108.
18. Kohonen T (1989) *Self-Organization and Associative Memory*. Springer-Verlag, Berlin.
19. Kachigan SK (1991) *Multivariate Statistical Analysis: A Conceptual Introduction*. Radius Press, New York.
20. Dunteman GH (1989) *Principal Components Analysis, Vol. 69*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07–064. Sage, Newbury Park, CA.
21. Everitt B (2001) *Applied Multivariate Data Analysis*. Edward Arnold, London.
22. Yeung KY and Ruzzo WL (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* **17**: 763–774.
23. West M, Blanchette C, Dressman H, *et al*. (2001) Predicting the clinical status of human breast cancer using gene expression profiles. *Proc Nat Acad Sci USA* **98**: 11462–11467.
24. Knudsen S (2002) A *Biologist's Guide to Analysis of DNA Microarray Data*. John Wiley and Sons, New York.
25. Quackenbush J (2001) Computational analysis of microarray data. *Nature Rev Genet* **2**: 418–427.

26. Raychaudhuri S, Stuart JM and Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, (eds), *Fifth Pacific Symposium on Biocomputing*, pp. 455–466.
27. Granucci F, Vizzardelli C, Pavelka N, *et al*. (2001) Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature Immunol* **2**: 882–888 .
28. Alter O, Brown PO and Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Nat Acad Sci USA* **97(18)**: 10101–10106.
29. Khan J, Simon R, Bittner M, *et al*. (1998) Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* **58**: 5009–5013.
30. Bittner M, Meltzer P, Chen Y, *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**: 536–540.
31. Yeung K, Fraley C, Murua A, Raftery A and Ruzzo W (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**: 977–987.
32. Kerr MK and Churchill GA (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Nat Acad Sci USA* **98**: 8961–8965.
33. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC and Simon R (2001) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Technical report 2. BRB, NCI, Bethesda, MD.
34. Bhattacharjee A, Richards WG, Staunton J, *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub classes. *Proc Nat Acad Sci USA* **98**: 13790–13795.
35. National Research Council: Panel on Discriminant Analysis Classification and Clustering. (1988) *Discriminant Analysis and Clustering*. National Academy Press, Washington, DC.
36. Dudoit S, Fridlyand J and Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* **97**: 77–87.
37. Gnanadesikan R (1977) *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.
38. Breiman L, Friedman JH, Olshen RA and Stone CJ (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.
39. Hastie T and Tibshirani R (1990) *Generalized Additive Models*. Chapman and Hall, London.
40. Neal RM (1996) *Bayesian Learning for Neural Networks*. Springer-Verlag, New York.
41. Rios Insua D and Mueller P (1998) Feedforward neural networks for nonparametric regression. In: *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, New York, pp. 181–194.
42. Slonim DK, Tamayo P, Mesirov P, Golub TR and Lander ES (1999) Class prediction and discovery using gene expression data. Discussion paper. Whitehead/M.I.T. Center for Genome Research, Cambridge, MA.
43. Dettling M and Bühlmann P (2002) Supervised clustering of genes. *Genome Biol* **3**: 0069.1–0069.15.
44. Michie D, Spiegelhalter DJ and Taylor CC (eds) (1994) *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.
45. Simon R, Radmacher MD, Dobbin K and McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* **95**: 14–18.

46. DeGroot MH and Fienberg SE (1983) The comparison and evaluation of forecasters. *The Statistician* **32**: 12–22.
47. Toussaint GT (1974) Bibliography on estimation of misclassification. *IEEE Trans Inform Theory* **IT-20**: 472–479.
48. Radmacher MD, McShane LM and Simon R (2001) A paradigm for class prediction using gene expression profiles. Technical report 1. BRB, NCI, Bethesda, MD.
49. Tibshirani R, Hastie T, Narasimhan B and Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Nat Acad Sci USA* **99**: 6567–6572.
50. Geman D, d'Avignon C, Naiman DQ and Winslow RL (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Applic Genet Mol Biol* **3**: Article 19.
51. Cover TM and Hart PE (1967) Nearest neighbor pattern classification. *IEEE Trans Inform Theory* **IT-13**: 21–27.
52. Vapnik V (1998) *Statistical Learning Theory*. Wiley, New York.
53. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**: 121–167.
54. Christianini N and Shawe-Taylor J (2000) *An Introduction to Support-Vector Machines*. Cambridge University Press, Cambridge.
55. Lee Y and Lee CK (2002) Classification of multiple cancer types by multicategory support vector machines using gene expression data. Technical Report 1051. University of Wisconsin, Madison, WI.
56. Brown MPS, Grundy WN, Lin D, *et al*. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Nat Acad Sci USA* **97**: 262–267.
57. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* **7(2)**: 179–188.
58. Hastie T J, Tibshirani R and Buja A (1994) Flexible discriminant analysis by optimal scoring. *J Am Stat Assoc* **89**: 1255–1270.
59. Li W and Yang Y (2002) How many genes are needed for a discriminant microarray data analysis? In: Lin SM, Johnson KF (eds), *Methods of Microarray Data Analysis*. Kluwer Academic, Dordrecht, pp. 137–150.
60. Zhang H and Yu CY (2002) Tree-based analysis of microarray data for classifying breast cancer. *Front Biosci* **7**: 63–67.
61. George EI and McCulloch RE (1993) Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**: 881–889.
62. Clyde MA, DeSimone H and Parmigiani G (1996) Prediction via orthogonalized model mixing. *J Am Stat Assoc* **91**: 1197–1208.
63. Clyde MA and Parmigiani G (1998) Bayesian variable selection and prediction with mixtures. *J Biopharmaceut Stat* **8(3)**: 431–443.
64. Pavlidis P, Tang C and Noble WS (2001) Classification of genes using probabilistic models of microarray expression profiles. In: Zaki MJ, Toivonen H, Wang JTL, (eds), *Proceedings of BIOKDD 2001: Workshop on Data Mining in Bioinformatics*. Association for Computing Machinery, New York, pp. 15–18.

Statistical analysis of microarray time course data

20

Yu Chuan Tai and Terence P. Speed

20.1 Introduction

This chapter discusses the statistical analysis of microarray time course data, with a focus on developmental time course experiments. The methods reviewed here are generally suitable for experiments based on the most widely used kinds of microarray platforms, including single-color, fluorescently labeled, high-density short oligonucleotide arrays on silicon chips (1), radiolabeled cDNA arrays on nylon membranes, or two-color, fluorescently labeled cDNA (2, 3) or long oligonucleotide arrays on glass slides.

Microarray time course experiments typically involve gene expression measurements for thousands of genes over relatively few time points, under one (e.g. wildtype) or more biological conditions (e.g. mutant 1, mutant 2,…). The number of time points can be 3–10 for shorter and 11–20 for longer time courses. The time points at which mRNA samples are taken are usually determined by the investigator's judgement concerning the biological events of interest and are frequently irregularly spaced, although for periodic time course experiments, equally-spaced times are standard. Measurements of mRNA abundance will be based on mRNA extracted from cell lines, tissue samples or whole organisms, and in what follows we will use the general term *units* for the source of the mRNA. The major advantage of microarray time course studies is that they give us the ability to monitor the temporal behavior of a biological process of interest through the measurement of expression levels of thousands of genes simultaneously. This can be a powerful experimental design for identifying patterns of gene expression in the units of interest.

Time course experiments can be classified into two main categories which we term *periodic* and *developmental*. Periodic time courses include natural biological processes whose temporal profiles follow regular patterns. Examples are cell cycles (4–6), and circadian rhythms (7), and we expect regulated genes to have periodic expression patterns. In the literature, periodic time course experiments are frequently unreplicated, that is, they arise as a single series of microarray experiments, experimenters perhaps preferring to

use scarce resources obtaining a finer temporal resolution, rather than repeating measurements at times already observed. In what we term developmental time course experiments, gene expression levels are measured at successive times during a developing process, for example, during the natural growth and development of, or following, a treatment applied to the units. In such cases there are usually few prior expectations concerning the form of the temporal profiles. Here there are commonly two to five replicate series, but sometimes there is no replication.

We now summarize several microarray time course experiments. Tomancak *et al*. (8) conducted a study of *Drosophila* embryogenesis using microarray time course experiments as the control of RNA *in situ* hybridization. Canton S fly embryos were collected, transferred to an incubator and aged. At hours 1 to 12 post egg laying, the embryos were dechorionated and quick-frozen, yielding 12 time point samples. The same procedure was repeated on three different days, producing three replicates. Himanen *et al*. (9) reported a study on a lateral root induction system of *Arabidopsis thaliana* to characterize the early molecular regulation induced by auxin. Seeds of *A. thaliana* were germinated on media containing auxin transport inhibitor N-1-naphthylphthalamic acid (NPA). After the germination, the seeds were moved to media with auxin 1-naphthalene acetic acid (NAA) only and the samples of the root segments were collected at four time points: 0, 2, 4, and 6 h after the transfer from NPA to NAA. There were two biological replicates at each time point and cDNA microarray experiments with the reference design were performed. They identified 906 differentially expressed genes over time and grouped these genes into six major clusters. In Qi *et al*. (10), bone marrow-derived mesodermal progenitor cells (MPCs) were obtained from three donors and the gene expression profiles for MPCs or MPCs induced to the osteoblast or chondroblast lineage for 1, 2, and 7 days were monitored using cDNA microarrays. They identified 41 transcription factors differentially expressed over time, in addition to some known signaling genes, hormones, and growth factors involved in osteogenesis. A fourth example of a developmental time course study is Schwamborn *et al*. (11). These authors studied the transcriptional response of the human astrocytoma cells U373 to tumor necrosis factor α (TNFα). Again, cDNA microarray experiments were performed. Samples from both TNFα-treated and untreated U373 cells were collected at 1, 2, 4, 8, and 12 h post treatment, and each time point had three biological replicates. The temporal profiles between these two treatments were compared. More than 880 genes were shown to be responsive to TNFα. In Tepperman *et al*. (12), gene expression samples were collected at six time points, and the gene profiles of wildtype (wt) and the phytochrome B (phyB) null mutant *A. thaliana* were compared to identify genes regulated by phyB in response to continuous monochromatic red light (Rc) during the induction of seedling de-etiolation. The study of transcriptional response to corticotrophin-releasing factor (CRF) in Peeters *et al*. (13) provides an example of unreplicated time course experiments with four different treatments: DMSO, ovine CRF in DMSO, R121919 in DMSO, and CRF plus R121919 in DMSO. Samples were collected at 0, 0.5, 1, 2, 4, 8, and 24 h after these four treatments were applied to mouse AtT-20 cells, and were hybridized to Affymetrix chips.

Following standard practice in statistics (14), we further categorize time course experiments into *longitudinal* and *cross-sectional*. Longitudinal experiments are those in which the mRNA samples at different times are extracted from the *same* unit, be it cell line, tissue or individual. This allows joining of ordinate values corresponding to observations on the same unit at different times, either by straight lines or fitted curves, to give the unit's *time course* for each gene, which will also be called the *temporal pattern* or *profile*. By contrast, cross-sectional time course experiments are those in which the mRNA samples at different times are extracted from *different* sources (units). With cross-sectional data, the individual data points can also be joined across time, using averages when there are replicate measurements, but the interpretation of the resulting curve is different. It will not correspond to any particular unit, but will be thought of as a population curve. In practice there will be experiments exhibiting features of both longitudinal and cross-sectional types, e.g. Tomancak *et al*. (8). There are more cross-sectional microarray time course experiments published to date than longitudinal ones, for example, Tepperman *et al*. (12) and Himanen *et al*. (9) cited above. This is probably because it is often infeasible to carry out longitudinal experiments because of the limited availability of mRNA from individual organisms such as laboratory mice. However, Qi *et al*. (10) is an example of a longitudinal study.

In this chapter we review methods for the design and analysis of microarray time course experiments. After discussing design issues in Section 20.2, we turn to methods for identifying the genes of interest to the experimenter in Section 20.3, be they genes which change over time, or genes which change differently over time between two or more biological conditions. Depending on one's perspective, this task can be viewed as a 'filtering' of the genes to remove those which are not of interest, before turning to a different kind of analysis such as clustering, or it can be seen as identifying a small to moderate list of genes for validation and further characterization. We use the microarray time course data of the study in *Drosophila* embryogenesis in Tomancak *et al*. (8) to illustrate the concept of moderation and compare some of the statistics we describe below. We then review the literature on clustering gene expression microarray time course data in Section 20.4, and end with a few comments about alignment of time series in Section 20.5.

We close this introduction with a few remarks on why the analysis of microarray time course data is special, and not adequately covered by the enormous literature that already exists on the analysis of time series (see e.g. 15 and references therein). There are three principal reasons. One is the fact that microarray time series are usually so *short* that we cannot consider applying methods typically used to analyze time series data, for example ARMA, Fourier or wavelet methods, as in Diggle's and other time series books. A second reason is that there are typically thousands of genes and hence thousands of short time series, all sharing the common experimental conditions. It is natural to think of analyses which have elements in common for all genes, such as the empirical Bayes (EB) methods described below. While there is some literature on EB methods in time series, we know of none involving thousands of series, as is the case here. Finally, an important aspect of microarray gene expression data is the clustering of genes,

here based on their temporal profiles. As far as we know, this problem has not arisen in the traditional time series literature, at least not in the form we meet it here. A far more relevant body of literature is that on longitudinal data analysis, for when the data are longitudinal, this is precisely the right context. When the data are not longitudinal, our context is that of many related regression models.

20.2 Design

The choice of design for microarray time course experiments will depend on several factors, principally the questions the researcher wishes to address, and the available resources, including mRNA, microarrays, and related reagents. We refer to Yang and Speed (16) for a general discussion of design issues for microarray experiments.

The first and most important microarray time course design question for an investigator will be whether to carry out a longitudinal or cross-sectional study. As explained in Diggle *et al*. (14), while it is often possible to address the same biological question using either longitudinal or cross-sectional experiments, the major merit of longitudinal time course studies is they provide information about the temporal *changes* in gene expression levels *within* units, something that is not possible with cross-sectional studies. In statistical terms, the difference between longitudinal and cross-sectional experiments comes from the fact that gene expression measurements are typically correlated over time within units, and these correlations can be estimated and used to advantage in longitudinal studies. On the other hand, such biological correlations cannot be detected in cross-sectional time course studies, since mRNA is extracted from different sources at different times.

It follows from what has just been said that if temporal changes in gene expression over time are of primary interest to the experimenter, an effort should be made to carry out a longitudinal study wherever possible. We appreciate that in many, and perhaps most, cases this may be infeasible, because of the impossibility of repeatedly sampling the mRNA from the same units. Nevertheless, a good approximation to a longitudinal study can be often realized by creating parallel, identically treated units, and sampling from different ones at different times. The difference between this kind of design and true longitudinal design depends on just how similar are the parallel, identically treated units. In some cases, they can be very similar indeed, and we observe the correlations characteristic of a genuinely longitudinal design, although their origins may be different. In such cases, these designs can be more powerful than cross-sectional studies for detecting changes.

Although we might give the impression that, for design purposes, the distinction between longitudinal and cross-sectional studies is straightforward, this is not really the case. There are many contexts in which hybrid studies pose more challenging design problems. For example, we might run replicate time course experiments on plants grown in a growth chamber, where each replicate consists of a series of successively sampled plants grown together under controlled conditions. This study is cross-sectional from the perspective of plants, but longitudinal from the point of view of growth chambers. The appropriate number of full replicates, and of plants

at each time within replicates given fixed resources, will depend on the relative magnitude of the different components of variation.

The number of time points is usually decided by the biological background and the cost of the study. More time points permits a finer analysis of temporal patterns, e.g. a more accurate determination of the time of onset or decay of a gene's expression, but in some experiments accuracy of this kind is not required.

Questions of interest to an investigator might concern the temporal profile of genes for one biological condition, such as a desire to identify cell-cycle-regulated genes. Alternatively, interest might focus on comparison between gene profiles across two or more conditions. We might want to identify those genes which change over time in a wildtype organism, and similarly those genes which change over time in a mutant organism, *and* identify those genes whose temporal profiles for the wildtype and mutant are *different*. The latter may include many genes which are unchanging in one or the other of the two biological conditions. The way in which such questions can affect the choice of design is explained in Yang and Speed (16), and we will not revisit that here in detail.

A short time course experiment can be regarded as a single factor experiment with *time* as a factor (see 17, 18). What makes it different from other single factor experiments is the additional information from the natural ordering of time course samples. This natural ordering of levels will lead to certain comparisons being of greater interest to the researcher, and others of lesser interest. For example, comparisons of each time point mRNA sample with the baseline or differences between consecutive time points are likely to be of greater interest than comparisons between widely separated times. Interest might focus on specific aspects of the temporal patterns of gene responses, such as monotonicity, convexity, and linearity (16).

The design of time course experiments can be considerably more complicated in the two-color comparative experiments (e.g. cDNA arrays) in comparison with single-channel experiments (e.g. Affymetrix chips), although if a common reference design is used, the two cases are fairly similar. For short two-color comparative time courses, it is possible to enumerate all the possibilities to find the optimal design. However, for those with a much larger number of time points, like the yeast cell cycle data in Spellman *et al*. (5), this is not feasible. There is not much literature on the design of time course experiments, but recently, Glonek and Solomon (19) described a method for designing short *cDNA* time course experiments. They optimized statistical efficiency and identified so-called admissible designs, and selected efficient designs based on the effects of most interest to the biologists, the number of arrays available, and other resources. Their approach was shown to give designs better than the popular common reference design and those incorporating all possible pairwise comparisons. Optimal design for microarray time course experiments is a research topic for the future.

Replication

Replication is an important aspect of all statistical experimental design. As described in Yang and Speed (16), there can be three types of microarray

replicates: biological replicates where mRNA samples are taken from different units; technical replicates, where mRNA samples are taken from the same unit and are split and hybridized onto different arrays; and within-array replicates, where probes are spotted in replicate on the same array. These types apply equally to longitudinal and cross-sectional time course experiments. The variation between gene expression measurements taken on these three types of replicates will be different, and in general is gene-specific. Replication is a *good thing,* as it provides estimates of variability relative to which temporal changes and/or condition differences can be assessed, making analysis much more straightforward. Biological replicates are generally best, as they permit the conclusions from the experiment to be extrapolated to the wider population of units from which the experimental units were obtained, something which is not possible with only technical replicates. With unreplicated experiments, the inference to a wider population is not possible, and the analysis is less straightforward, being more dependent on unverifiable assumptions, as there is no estimate of pure error which can be used. We suggest at least three replicates at every time point. When replicates are available, it is better to use the variation between them in any analysis, rather that just average across the replicates and proceed as with a single time course experiment. Many of the methods we describe below require replicates, and are designed to be effective with the small numbers of replicates common in this context.

20.3 Identifying the genes of interest

There have been many published studies involving developmental time course experiments. As indicated above, experimenters' aims vary, but we can categorize them broadly as: (i) one-sample, where the aim is to identify genes which change over time, perhaps in some specific way; (ii) two-sample, where in addition to identifying temporally varying gene expression, interest is in comparisons across two biological conditions; and (iii) $D >$ two-sample experiments, which are as in (ii), with $D \geq$ three biological conditions. These categories apply equally to longitudinal and cross-sectional studies. Before we go on to the analysis methods, we illustrate the foregoing by assigning some of the case studies cited above to their category.

The study in Himanen *et al*. (9) on a lateral root induction system of *A. thaliana* to characterize the early molecular regulation induced by auxin is an example of a one-sample problem: only a single treatment (i.e. NPA followed by NAA) was applied to the same type of cells, and the genes whose expression levels change over time were of interest. Examples in the two-sample category include the study of Schwamborn *et al*. (11), who compared the temporal profiles between the TNFα-treated and untreated human astrocytoma cells U373, in order to elucidate the post-treatment transcriptional response. Similarly, Tepperman *et al*. (12) compared the temporal profiles of genes in wild type (wt) and phytochrome B (phyB) null mutant *A. thaliana,* to identify genes regulated by *phyB* in response to continuous monochromatic red light (Rc) during the induction of seedling de-etiolation. Both of these studies involved just two different biological conditions: treated versus control cells, and wild type versus mutant organ-

isms. In this type of problem, identifying all the genes with different temporal profiles between the two biological conditions was usually of interest, although sometimes only genes with different shapes were of interest, with those having similar shapes but different magnitude across the two conditions not being of interest. An example in category (iii) is the study of transcriptional response to CRF described in Peeters *et al*. (13). There gene expression levels were measured at 0, 0.5, 1, 2, 4, 8, 24 h after four different treatments are applied to mouse AtT-20 cells. As in the two-sample problem, genes of interest are those which have different gene expression profiles over time, either in shape and magnitude, or in magnitude only, between the four treatments.

The identification of temporally changing or differentially changing genes not only gives insight into the biological processes under study, it also provides a way of selecting a subset of genes from the entire gene set for further analysis such as clustering. As yet there are relatively few methods available for identifying the genes of interest in this context. The approaches most widely used are those for identifying differentially expressed genes for replicated microarray experiments across two or more independent sample groups (20–29). The idea here is the simple one of testing whether there are changes in a gene's expression across time by making comparisons between times, for example between all consecutive pairs of time points, or all possible pairs of time points (85). It is reasonable though not ideal to analyze time course data with these approaches, as they assume independence of the samples across different times, which is not true for longitudinal time course data. We will give some solutions to these problems for both longitudinal and cross-sectional data in the following subsections.

ANOVA and the *F*-statistic

An intuitive way to select differentially expressed genes from time course data is to use the classical ANOVA (cross-sectional) or mixed-effect ANOVA (longitudinal) model (14, 18). In the one-sample case, one includes *time* as a factor and possibly *replicate* as another factor, and calculates the *F*-statistic corresponding to time. Similarly, for two- and $D >$ two-sample cases, one includes the factors *time* and *biological condition,* and their *interaction* term in the model, and possibly *replicate* as another factor, and computes the *F*-statistic for the interaction term. Wang and Kim (30) has a one-sample example using classical one-way ANOVA. In order to obtain approximately valid *p*-values, care needs to be taken to deal with multiple testing (26 and references therein).

Park *et al*. (31) proposed a modified ANOVA approach and some variants without the normality assumption. Their basic model is like the usual two-way ANOVA with time, biological condition, and their interaction as effects. Genes which are not significant (after *p*-value adjustments) in the time $\times$ condition interaction term will be fitted with the second model, removing the interaction term. Then genes with significant time effect after *p*-value adjustments are selected. However, their method ignores the potential correlations among times from longitudinal time course data. Even if there are no biological correlations, the fact that there are usually only very

few replicates makes the estimation of gene-specific variances unstable. Romagnolo *et al*. (32) and Wang and Kim (30) used mixed-effect ANOVA to identify genes with different temporal profiles between the heat-shock-treated *let-60* and wildtype, and dauer exit and L1 starvation *C. elegans,* respectively. Himanen *et al*. (9) also used mixed-effect ANOVA for their one-sample problem. Chapter 6 of Diggle *et al*. (14) discusses some standard ANOVA methods for longitudinal data, and we refer the reader there.

A number of questions are not adequately addressed by classical ANOVA methods or the variants of Park *et al*. (31). First, to obtain an *F*-distribution for the *F*-statistic, we require that the samples at different time points are independent, an assumption easily violated by longitudinal time course data, and we also require normality of the observations. The robustness of the *F*-distribution to deviations from normality is well-studied (see e.g. 33), but for gene expression measurements on the log scale this may not be a great concern. Secondly, as a result of the very large number of genes, and relatively small number of replicates, the *F*-statistic may lead to more false positives and false negatives than would normally be the case, because of poorly estimated variances in the denominator. See Tai and Speed (34) for the results of a simulation study. This issue can be addressed using the notion of moderation (see below).

Despite these reservations concerning the *F*-statistic, it should be understood that it has been and will continue to be effective for identifying genes of interest to researchers. However, we believe that we can do this job better with alternative statistics.

Moderation

Typically, genes with large changes in gene expression levels over time relative to their replicate variances are best candidates for following up. However, given the thousands of genes in a microarray time course experiment, and the small number of replicates, the variances (in the case of cross-sectional data) or variance-covariance matrices (in the case of longitudinal data), are usually very poorly estimated. As a result, some genes which exhibit relatively small amounts of change over time and small replicate variances, may have large between-to-within time *F*-statistics because of these under-estimated denominators. We may conclude that such genes are changing over time, but if they are not, they will be false positives. For example, Jiang *et al*. (35) mentioned such genes. *Figure 20.1* gives an example of such a gene from Tomancak *et al*. (8). The *F*-statistic for this gene has a higher ranking than many other genes exhibiting greater change in expression levels over time, so we consider it a false positive. On the other hand, some genes with large amounts of changes over time, but also large replicate variances, may have small *F*-statistics because of over-estimated denominators. Such genes may be false negatives, that is may be changing over time, but not be identified as such. The gene in *Figure 20.2* is clearly changing over time, however, the expression level of experiment B at 10 h is much lower than those of experiments A and C. This single outlier leads to lower rankings for the *F*-statistic compared to all the other statistics mentioned below; however, this gene is very likely to be of interest. By moving (shrinking) gene-specific variances or variance-covariance matrices

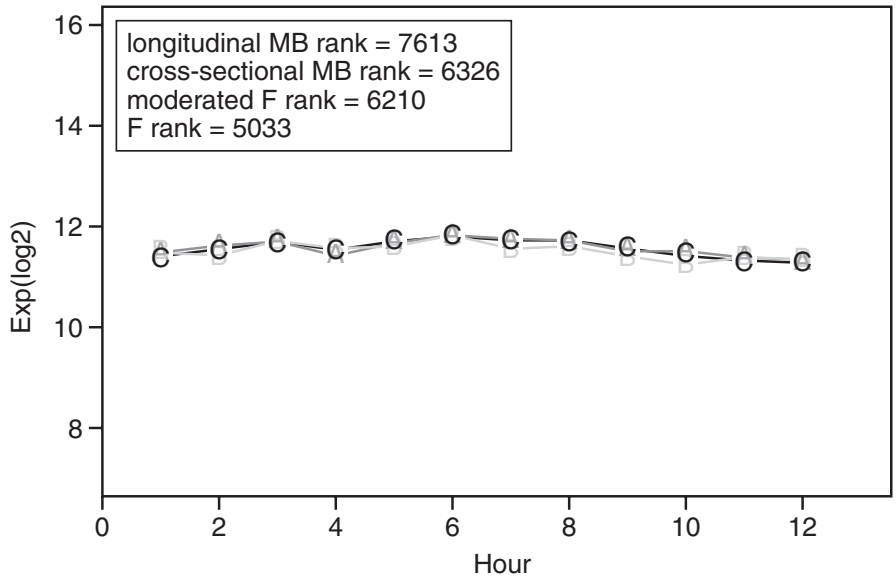


Figure 20.1.

A probable false positive gene (see text).

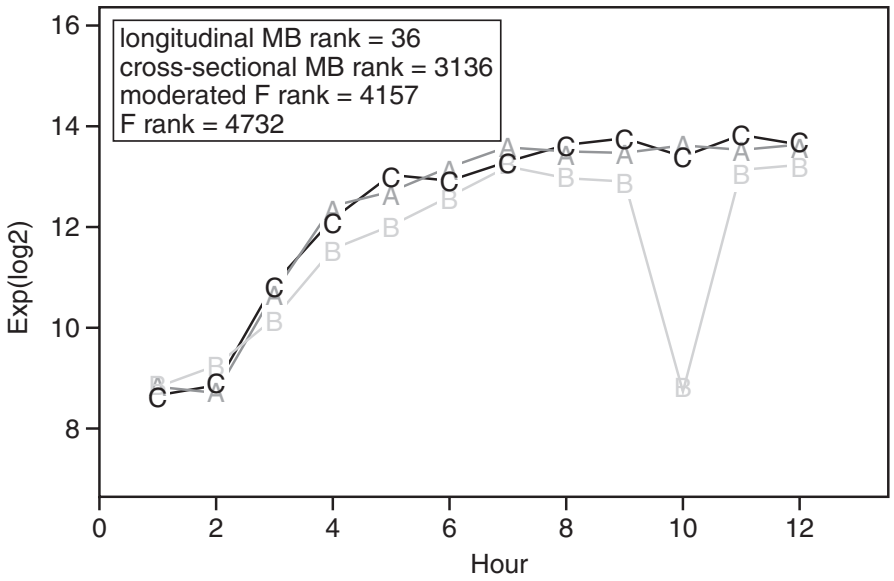


Figure 20.2.

A probable false negative gene (see text).

towards a common value estimated from the whole gene set, the total number of false positives and false negatives can usually be reduced. This is what we term *moderation*.

The idea of moderation has entered into the analysis of microarray data in different forms. Efron *et al*. (21) and Tusher *et al*. (22) tuned the *t*-statistic by adding a suitable constant to the standard deviation, the constant being estimated by a percentile of sample standard deviations, by minimizing a coefficient of variation, respectively. Their approaches are not based on any distributional theory. Lönnstedt and Speed (24) brought the idea of moderation in the univariate hierarchical Bayesian mixture model for two-channel comparative experiments by smoothing gene-specific residual sample variances toward a common value. Smyth (29) formally introduced the moderated *t*-statistic in the univariate general linear model setting by substituting the denominator of *t*-statistic with a moderated denominator. The gene-specific moderated sample variance in Smyth (29) is defined based on some nice distributional theory and the hyperparameter estimates derived there are shown to perform better than those in Lönnstedt and Speed (24). Tai and Speed (34) further extended the univariate model in Lönnstedt and Speed (24) and Smyth (29) into multivariate settings and introduced the *MB*-statistic (multivariate empirical Bayes statistic) and the $\tilde{T}^2$ statistic to rank genes in the order of differential expression in the one- and two-sample cases in the longitudinal time course context. In addition, Tai and Speed (36, 37) derived *MB*-statistics for $D >$ two samples for longitudinal and cross-sectional data.

Likelihood-based approach

Before we go on to discuss our multivariate empirical Bayes methods, here we briefly comment on the likelihood-based approaches for longitudinal time course data. Given the very few replications in this context, the gene-specific sample variance-covariance matrix or its variant may be singular and the resulting analysis can be unstable. Guo *et al*. (38) proposed the gene-specific score based on the robust Wald statistic for one-sample longitudinal data, using an approach similar to Tusher *et al*. (22), adding a small positive number times the identity matrix in the denominator.

$$w(i) = [\mathbf{L}\hat{\beta}(i)]^T [\mathbf{L}\hat{\mathbf{V}}_S(i)\mathbf{L}^T + \lambda_w \mathbf{I}_{r\times r}]^{-1} [\mathbf{L}\hat{\beta}(i)], \tag{20.1}$$

where $\mathbf{L}$ is an $r \times p$ matrix of rank r, $\hat{\beta}$ is the estimated $p \times 1$ vector of unknown regression parameters β, $\hat{\mathbf{V}}_S$ is the estimated covariance matrix for $\hat{\beta}$, and λ_w is a positive scalar. The way they estimated λ_w is exploratory. Moreover, their approach is for a one-sample problem only, and the fact that the number of subjects is usually very small makes asymptotic theory inappropriate.

Storey *et al*. (39) also proposed a likelihood-ratio based approach, assuming gene expression values are composed of population mean and individual deviates. They constructed the *F*-statistics for both longitudinal and cross-sectional data in a standard way, and presented a careful treatment of the multiple testing issue. In contrast, Tai and Speed (34) suggested the moderated LR and Hotelling T^2 statistics, with the smoothing of gene-

specific sample variance-covariance matrix making use of the replicate variability information across the whole gene set, and then they simply ranked genes according to one or the other statistic. These two statistics were shown in a simulation study to perform about as well as the *MB*-statistic described below.

Empirical Bayes

The existence of thousands of genes in the microarray time course context brings to mind the empirical Bayes (EB) approach to inference. This is a model-based way of introducing moderation into the analysis.

In Tai and Speed (34), a multivariate hierarchical normal model with conjugate priors is proposed to derive the posterior odds for differential expression in the one- and two-sample problems. This is designed for longitudinal data, and takes into account correlations across times. When all genes have the same number of replicates, the *MB*-statistic for the null hypothesis that the expected profile equals to $\mathbf{0}$, or the paired two-sample problem with the null hypothesis that the expected profiles are the same, is a monotonic increasing function of the statistic $\tilde{T}^2 = \tilde{\mathbf{t}}'\tilde{\mathbf{t}}$, where

$$\tilde{\mathbf{t}} = n^{\frac{1}{2}}\tilde{\mathbf{S}}^{-\frac{1}{2}}\overline{\mathbf{X}}, \tag{20.2}$$

is just the traditional multivariate *t*-statistic with the denominator replaced by a moderated covariance matrix $\tilde{\mathbf{S}}$. This expression incorporates the gene-specific covariances but also shares the covariance information across genes:

$$\tilde{\mathbf{S}} = \frac{(n-1)\mathbf{S} + \nu\Lambda}{n-1+\nu}. \tag{20.3}$$

Here $\mathbf{S}$ is the gene-specific sample variance-covariance matrix; $\overline{\mathbf{X}}$ is the gene-specific sample average time course vector; n is the number of replicates; ν and Λ are hyperparameters estimated from the whole gene set (see 34 for details). The *MB*-statistic or $\tilde{T}^2$ statistic for the independent two-sample problem is also derived in Tai and Speed (34), where differential expression now means that the expected profiles are different between the two biological conditions. It is shown there that the *MB*-statistic achieves the lowest numbers of false positives and false negatives, and performs about as well as the moderated Hotelling T^2 statistic. One of the values of the multivariate empirical Bayes is that it provides a natural way to estimate the gene-specific moderated sample covariance matrix, while the likelihood ratio based approach (moderated Hotelling T^2 statistic) does not.

Let $\mathbf{Y}_{di}$ be the i-th replicate for the d-th condition, and $\overline{\mathbf{Y}}$ and $\overline{\mathbf{Y}}_d$ be the overall sample average, and average for the d-th condition only, respectively. In the case that there are more than two biological conditions and all genes have the same number n of replicates within each condition, the posterior odds for difference between conditions for a conjugate normal model are proportional to

$$\left(\frac{|\,\mathbf{TSSP} + \mathbf{M} + v\Lambda\,|}{\left|\sum_{d=1}^{D} \mathbf{WSSP}_d + \sum_{d=1}^{D} \mathbf{M}_d + v\Lambda\right|} \right)^{\frac{1}{2}(nD+v)}, \tag{20.4}$$

where $\mathbf{TSSP} = \sum_{d=1}^{D}\sum_{i=1}^{n}(\mathbf{Y}_{di} - \overline{\mathbf{Y}})\,(\mathbf{Y}_{di} - \overline{\mathbf{Y}})'$ and $\mathbf{WSSP}_d = \sum_{i=1}^{n}(\mathbf{Y}_{di} - \overline{\mathbf{Y}}_d)\,(\mathbf{Y}_{di} - \overline{\mathbf{Y}}_d)'$ are the total and within-condition sums of squares and products, $\mathbf{M}d$, $\mathbf{M}$, and $v\Lambda$ are matrices involving the (condition-specific) prior means and variance-covariance matrices, respectively. This is our EB analogue of Wilks' likelihood-based Λ from MANOVA (see 36 for details).

For cross-sectional data across $D \geq$ two biological conditions, Tai and Speed (37) derive the posterior odds that the expected temporal profiles are different among biological conditions versus they are the same. Let Y_{dji} be the log_2 intensity value or log_2 ratio of this gene for the d-th biological condition, j-th time point, and i-th replicate. Y_{dji} are independent across times ($j = 1,\ldots, k$), biological replicates within conditions ($i = 1,\ldots, n_{dj}$) and biological conditions ($d = 1,\ldots, D$). The sampling times need not to be the same within and across biological conditions. For the simplest case, we assume they are, and that all genes have the same number of replicates n for all conditions and times. Again, under a conjugate normal model with unstructured means, the posterior odds are proportional to

$$\left(\frac{TSS + m + v\lambda^2}{\sum_{d=1}^{D} WSS_d + \sum_{d=1}^{D} m_d + v\lambda^2} \right)^{\frac{1}{2}(nkD+v)}, \tag{20.5}$$

where $\overline{Y}_{dj} = n^{-1}\sum_{i=1}^{n} Y_{dji}$ and $\overline{Y}_j = D^{-1}\sum_{d=1}^{D}\overline{Y}_{dj}$ denote the average log_2 (relative) expression level at the j-th time point for the d-th condition only and all the conditions, respectively; $TSS = \sum_{d=1}^{D}\sum_{j=1}^{k}\sum_{i=1}^{n}(Y_{dji} - \overline{Y}_j)^2$ and $WSS_d = \sum_{j=1}^{k}\sum_{i=1}^{n}(Y_{dji} - \overline{Y}_{dj})^2$ are the total and within sums of squares, respectively; m_d, m, and $v\lambda^2$ are quantities involving (condition-specific) prior means and variances. This is a special case of our fully moderated F-statistic, the EB analogue of the traditional F-statistic. Gene selection using either the MB-statistic or the $\tilde{T}^2$ statistic can be based on rankings.

The multivariate EB procedure in Tai and Speed (34) focuses on moderating the denominator of the multivariate t-statistic $\mathbf{t}$, and ranks genes according to the moderated statistic $\tilde{T}^2$, to reduce the number of false positives and false negatives resulting from very small or very large replicate variances or covariances. Alternatively, one could replace the numerator of the multivariate t-statistic with a robust estimate, to avoid the problem of very large $\tilde{T}^2$ resulting from outliers. Such an outliers issue can be common in the microarray time course context, when the sample sizes are typically very small (two or three). Incorporating robust methods into the analysis of microarray time course is a research topic of interest here. *Figures 20.3–20.5* gives the profiles of the top-ranked genes from Tomancak *et al*. (8) using the one-sample longitudinal MB-statistic (34), the one-sample cross-sectional MB-statistic with a fifth-degree polynomial model for the means (37), the moderated F-statistic (29), and the usual F-statistic with unstructured means.

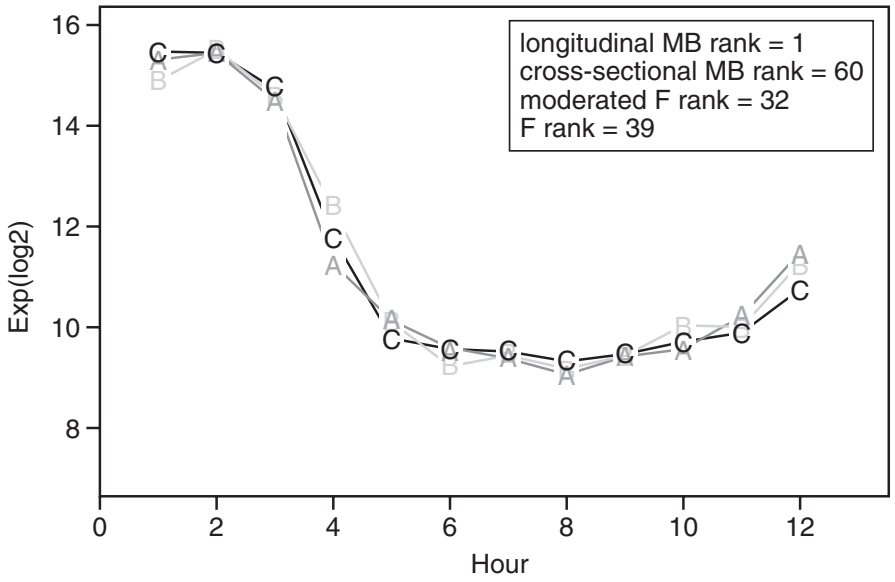


Figure 20.3.

The top gene by the one-sample longitudinal *MB*-statistic.

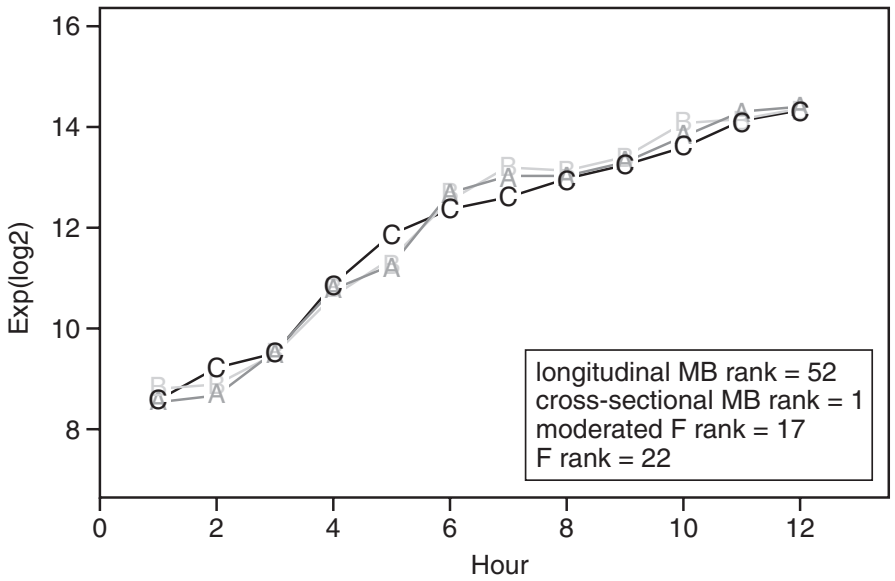


Figure 20.4.

The top gene by the one-sample cross-sectional *MB*-statistic.

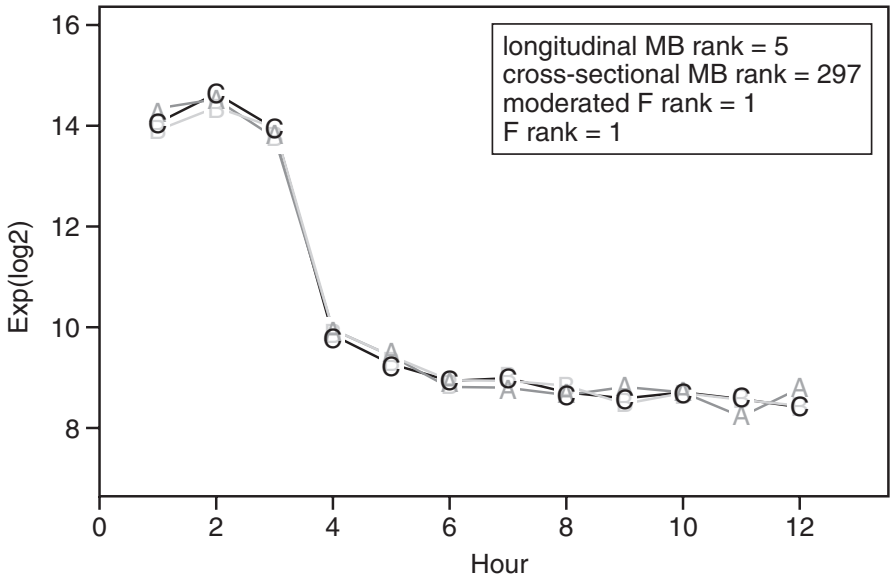


Figure 20.5.

The top gene by both the moderated *F*-statistic and the *F*-statistic.

Regression approaches, including B-splines

To date these methods have been used mainly for unreplicated time course data under a single biological condition. Zhao *et al*. (40) outlined a regression model to search for genes with transcriptional response to a stimulus. Their regression function was built to relate each gene's profile to a vector of covariates including dummies for the stimulus categories, time, and other characteristics of the sample. Their model included gene-specific parameters and parameters to model the heterogeneity across arrays. The mean vector was estimated using the technique described in Liang and Zeger (41). They further focused on the single-pulse model (SPM), which is specific for the setting when cells are released from cell cycle arrest. Xu *et al*. (42) described an application of the same kind of regression model to a time course study involving Huntington's disease.

Several researchers have suggested the use of B-splines to model gene profiles. In Bar-Joseph *et al*. (48) the expression profiles for each gene and each of two biological conditions were represented by continuous curves fitted using B-splines. A global difference between the two continuous curves and an *ad hoc* likelihood based *p*-value was calculated for each gene. Other papers using B-splines to model profiles are Luan and Li (44) and Hong and Li (45). Luan and Li (44) adopt the shape-invariant model (46, 47) for guide genes, and model the common periodic function shared by all periodically expressed genes using a B-spline basis. Such genes were identified using a false discovery rate (FDR) procedure. Both the B-spline based

approach in Bar-Joseph *et al*. (48) and Luan and Li (44) were illustrated on the yeast cell cycle datasets. They do not seem suitable for short time courses. Similarly, Hong and Li (45) proposed a B-spline based approach to identify differentially expressed genes in the two-sample case. There they modeled the expected profile as linear combinations of B-spline basis functions, and used a Markov chain Monte Carlo EM algorithm (MCEM) to estimate the gene-specific parameters and hyperparameters from the hierarchical model. They selected differentially expressed genes using empirical Bayes log posterior odds, and the posterior probability based FDR. They showed their method performed better than the traditional ANOVA model. As above, the approach in Hong and Li (45) seems more suited to longer time course data.

Contrasts

A simple but powerful tool for extracting temporal patterns is found in *contrasts:* linear combinations of gene expression measurements over time. Contrasts usually but not always have their coefficients summing to zero. An example of the use of contrasts can be seen in Lönnstedt *et al*. (49) where samples were taken from cells at 0.5, 1, 4, and 24 h after stimulation with a growth factor. Genes were regarded as *early* responders if they had large values of $< c, E >= \sum_t c_t E_t$ where $c_t = (t - 24.5)^2$ and E_t is the gene expression value at time t, while those having large values when $c_t = t^2$ were termed *late* responding genes. Smyth (29) used contrasts in the univariate linear model setting, and derived a partly moderated F-statistic for testing whether there is any change in gene expression levels over time. This approach assumes the samples are independent, and so would be appropriate for cross-sectional data. Fleury *et al*. (50) described a valuable multi-criterion optimization method called Pareto front analysis, for ranking and selecting genes of interest. In their paper they made use of contrasts to select genes with many predefined patterns. In essence, Pareto fronts and their variants (50–52) seek to identify genes with large values for all of a set of *competing* contrasts of interest.

Hidden Markov models

Yuan *et al*. (53) presented a hidden Markov model approach for selecting differentially expressed genes from replicated time course experiments with multiple biological conditions. They considered all possible equality and inequality relations among means across biological conditions as states, and the expression pattern process was modeled as a Markov chain, with either time-homogeneous or non-homogeneous transition matrices. The observations were conditionally independent given the state of the chain. In this approach, dependence between gene expression values at different times was completely described by the pattern process (i.e., the hidden Markov chain), and genes were selected based on the posterior probabilities of states of interest. This is an example of using an HMM to model time-dependence in microarray time course data, while many others have used HMMs in this context for clustering.

20.4 Clustering

The identification of differentially expressed genes narrows down the number of genes for further analysis. Clustering genes with similar temporal profiles is commonly the next phase of the initial analysis. This is done in the belief that genes with *similar* temporal profiles may well be involved in similar biological processes, for example in the same aspect of response to a treatment. Frequently there is a further hope that genes in the same cluster share common sequence motifs in their regulatory region. In clustering, the focus might be on grouping genes with particular temporal profiles, say early induced, monotonic increasing, up first and then down and so on, or it may simply be a way of partitioning all genes into automatically defined groups.

Below we briefly summarize the literature on clustering, referring to the review of Möller-Levet *et al*. (54) for a comprehensive treatment of the issues. One of the earliest examples was *hierarchical clustering* of the yeast cell cycle data by Spellman *et al*. (5) and Eisen *et al*. (55). Shortly afterwards, *self-organizing maps* (SOM) were applied to the same data, as well as a human dataset concerning hematopoietic differentiation by Tamayo *et al*. (56), while the *k-means* algorithm was used in Tavazoie *et al*. (57). This early literature used rather arbitrary criteria for reducing the number of genes prior to clustering. In the GENECLUSTER package, genes are filtered by a simple variation criteria; see (58) and GENECLUSTER 2 Reference Guide for details (http://www.broad.mit.edu/cancer/software/genecluster2/gc_ref.html). In producing a 6×4 grid SOM, Saban *et al*. (59) started with 588 genes which had to be induced at least three-fold over the initial time point in one replicate at some time point, and induced at least two-fold over the initial time point in the second replicate at that time point. Such filtering rules are not uncommon in the literature (see 60). A more recent example was the *hierarchical clustering* of 906 genes into six main groups representing three major patterns in Himanen *et al*. (9). A significance test within a mixed-model analysis was used to select 906 genes.

Different clustering algorithms and distance measures can lead to very different results. A perennial challenge with cluster analysis is the determination of the number of clusters. In recent years methods have been developed to deal with this issue, e.g. the gap statistic in Hastie *et al*. (61), see also (62).

As well as these classical clustering approaches, a number of model-based clustering algorithms have been proposed (e.g. 63). Ramoni *et al*. (64) gave a *Bayesian model-based clustering* algorithm, which represents temporal profiles by autoregressive models and used an agglomerative procedure to determine the number of clusters. Yeung *et al*. (65) used *Gaussian mixture models* in which each component corresponds to a cluster, the number of clusters being determined by the Bayesian Information Criterion (BIC). HMM clustering can be found in Schliep *et al*. (66) and Schliep *et al*. (67). There, each cluster was represented as one HMM. The method started with a collection of HMMs with typical qualitative behavior, and an iterative algorithm was used to fit these models and assign genes to clusters in such a way as to maximize the joint likelihood. This method also dealt with missing data, and was illustrated on yeast cell cycle data of Spellman *et al*.

(5), and on the fibroblast serum response data of Iyer *et al*. (68). Similarly, Ji *et al*. (69) and Zeng and Garcia-Frias (70) also used HMM approaches to cluster microarray time course data. Bar-Joseph *et al*. (43) and Luan and Li (71) did likewise, but first represented the profile for each gene by a continuous curve fitted by B-splines with gene-specific and class-specific parameters. Both papers illustrated their methods on the yeast cell cycle and fibroblasts serum response datasets. Zhang *et al*. (72) proposed a biclustering algorithm to discover genes which are co-regulated in only part of the time course. They illustrated their algorithm on the yeast cell cycle data of Cho *et al*. (4). Other noteworthy approaches were outlined in Peddada *et al*. (73) and Wakefield *et al*. (74).

Based on the above examples and our experience, we note that most model-based clustering algorithms have been effective for *periodic* time courses, but their satisfactory performance on short time-course experiments is not so clear. For *developmental* time-course data, traditional algorithms such as hierarchical clustering, SOM, or their variants based on distance measures have been more popular, probably because there are usually too few time points to allow the fitting of models. However, these approaches have the drawbacks of ignoring possible dependency across times in longitudinal studies, and generally ignoring the ordered nature of the time index. We feel that clustering methods which combine features from both the traditional and model-based approaches are urgently needed, ones which recognize the time ordering, and will deal with few time points, as well as temporal dependence and replicates where appropriate.

We end this clustering section by briefly mentioning a couple of other exploratory approaches like clustering that have used to analyze microarray time course data. These are correspondance analysis (75, 76), and singular value decomposition (SVD) (77, 78). Such graphical methods can be quite powerful.

20.5 Curve alignment

The occasional need to align gene expression profiles comes from the fact that the rates of biological processes may be different across biological or environmental conditions, or the sampling times are different between two time course datasets to be compared. Aach and Church (79) suggested a time-warping algorithm with or without interpolation, while Bar-Joseph *et al*. (43) gave a B-spline approach also for the same task, assuming that each gene's profile is fitted with gene-specific and class-specific parameters. As with so much we have mentioned, these approaches have only been illustrated on the yeast cell-cycle data. Again, this idea is well-suited to longer time series, but may not be suitable for shorter time series. There is clearly room for more work here.

20.6 Software

Many of the algorithms described in this chapter are implemented in open source software R (80), which can be downloaded from http://cran.r-project.org. The Bioconductor project (http://www.bioconductor.org) provides many software tools for the analysis of microarray data. For the

algorithms described in this chapter but not listed here, the reader should consult with the authors for software availability.

Differential expression

- The CyberT program in Baldi and Long (20) with a web interface can be downloaded from visitor.ics.uci.edu/genex/cybert/. It is also available as R code *hdarray*.
- The *MB*-statistic and $\tilde{T}^2$ statistic (34) will be implemented in the Bioconductor package *timecourse* www.stat.berkeley.edu/users/terry/Group/research/timecourse.html.
- The *B*-statistic (24, 29) and moderated *F*-statistic (29) are implemented in the Bioconductor package *limma*. The latest *limma* can be downloaded from bioinf.wehi.edu.au/limma/. *LimmaGUI* and *affylmGUI* (81) are two nice graphical user interfaces to *limma* for two-color arrays and Affymetrix chips, respectively.
- The empirical Bayes method across multiple independent groups in Kendziorski *et al*. (27) is available through the Bioconductor package *EBarrays*.

Clustering

- *SOM:* the package GeneSOM written by Jun Yan is available in R. GENECLUSTER 2.0 in Tamayo *et al*. (56) can be downloaded from www.broad.mit.edu/cancer/software/genecluster2/gc2.html.
- *hierarchical clustering:* the R function *hclust* implements bottom-up hierarchical clustering with several types of linkages.
- *k-means:* the R function *kmeans* is based on the algorithm in Hartigan and Wong (82).
- *QT_clust* proposed in Heyer *et al*. (83) is implemented by Witold Wolski. It can be accessed from www.molgen.mpg.de/~wolski/downloads/clustering/clustering.html.
- *CLARITY* proposed in Balasubramaniyan *et al*. (84) is available upon request from authors.
- The model-based clustering approach in Fraley and Raftery (63) is available as a R package *mclust:* www.stat.washington.edu/fraley/mclust/.
- The HMM clustering software *GHMM* in Schliep *et al*. (66) and Schliep *et al*. (67) is available from ghmm.org/.

Curve alignment

- Aach and Church's (79) time warping programs *genewarp* and *genewarpi* are available as DOS executables under Win 32. The download page is arep.med.harvard.edu/timewarp/pgmlicense.html.

20.7 Remarks

We have discussed some statistical issues in the analysis of microarray time course experiments, touching on their design, the identification of genes of interest, clustering, and alignment. For the practitioner, we have tried to offer some ways of addressing these issues, particularly the second.

As will be apparent from our discussion, many, perhaps most, of the methods in the literature available for choosing or clustering genes in time course experiments have been devised and tested on the yeast and human cell-cycle datasets. There is room for much more research on the analysis of what we have called developmental time course data, especially their clustering.

A less obvious bias in our coverage is the fact that almost all of the methods we have discussed have been for data generated in an experimental setting, with mRNA from cell lines, tissue samples or experimental organisms such as whole *Drosophila* embryos, or tissue from inbred strains of mice or *Arabidopsis* plants. Recently microarrays have moved to the wider clinical setting, with microarray data now being collected on human subjects over time. Such longitudinal studies present novel analytical challenges, as subject-to-subject variation, even within the same treatment group, can be substantial. The methods we have reviewed here will not be appropriate in the clinical context without modifications, for example, by including fixed or random effects for subjects. This is an important area for future research, but we can refer to Storey *et al*. (39) for a promising start.

Acknowledgments

We would like to thank all the biologists who have challenged us with their microarray time course data, specifically Suzie Grant, Jason Dugas, Mary Wildermuth, Moriah Szpara, Steve Perrin, and we particularly thank Pavel Tomancak and his colleagues for creating the *Drosophila* data set we have used in this chapter. Finally, thanks are due to Peter Diggle and Christina Kendziorski for reading and commenting on an earlier version of this chapter.

References

1. Lockhart DJ, Dong H, Byrne MC *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14(13)**: 1675–1680.
2. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA and Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14(4)**: 457–460.
3. Brown PO and Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21(1 Suppl)**: 33–37.
4. Cho R, Campbell M, Winzeler E, *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2(1)**: 65–73.
5. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **9(12)**: 3273–3297.
6. Cho R, Huang M, Campbell M, Dong H, Steinmetz L, Sapinoso L, Elledge S, Davis R and Lockhart D (2001) Transcriptional regulation and function during the human cell cycle. *Nat Genet* **27(1)**: 48–54.
7. Storch K-F, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH and Weitz CJ (2002) Extensive and divergent circadian gene expression in liver and heart. *Nature* **417**: 78–83.

8. Tomancak P, Beaton A, Weiszmann R *et al*. (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol* **3(12)**: research0088.1–0088.14 (the microarray time course data can be downloaded from www.fruitfly.org/cgi-bin/ex/insitu.pl.).
9. Himanen K, Vuylsteke M, Vanneste S *et al*. (2004) Transcript profiling of early lateral root initiation. *Proc Natl Acad Sci USA* **101(14)**: 5146–5151.
10. Qi H, Aguiar DJ, Williams SM, La Pean A, Pan W and Verfaillie CM (2003) Identification of genes responsible for osteoblast differentiation from human mesodermal progenitor cells. *Proc Natl Acad Sci USA* **100(6)**: 3305–3310.
11. Schwamborn J, Lindecke A, Elvers M *et al*. (2003) Microarray analysis of tumor necrosis factor alpha induced gene expression in u373 human glioblastoma cells. *BMC Genom* **4(1)**: 46.
12. Tepperman JM, Hudson ME, Khanna R, Zhu T, Chang SH, Wang X and Quail PH (2004) Expression profiling of *phyb* mutant demonstrates substantial contribution of other phytochromes to red-light-regulated gene expression during seedling de-etiolation. *Plant J* **38(5)**: 725–725.
13. Peeters PJ, Gohlmann HW, Van den Wyngaert I, Swagemakers SM, Bijnens L, Kass SU and Steckler T (2004) Transcriptional response to corticotropin-releasing factor in AtT-20 cells. *Mol Pharmacol* **66(5)**: 1083–1092.
14. Diggle PJ, Heagerty P, Liang K-Y and Zeger SL (2002) *Analysis of longitudinal data*. 2nd edn. Oxford University Press New York.
15. Diggle PJ (1990) *Time Series: A Biostatistical Introduction*. Oxford University Press, New York.
16. Yang YH and Speed TP (2003) Design and analysis of comparative microarray experiments. In: Speed T (ed.) *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC Press.
17. Searle SR (1997) *Linear Models*. John Wiley & Sons, New York.
18. Neter J, Kutner MH, Wasserman W and Nachtsheim CJ (1996) *Applied Linear Statistical Models*, 4th edn, McGraw-Hill/Irwin.
19. Glonek GFV and Solomon PJ (2004) Factorial and time course designs for cDNA microarray experiments. *Biostat* **5(1)**: 89–111.
20. Baldi P and Long AD (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.
21. Efron B, Tibshirani R, Storey JD and Tusher V (2001) Empirical bayes of a microarray experiment. *J Am Stat Assoc* **96**: 1151–1160.
22. Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98(9)**: 5116–5121.
23. Dudoit S, Yang YH, Speed T and Callow M (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin* **12(1)**: 111–139.
24. Lönnstedt I and Speed TP (2002) Replicated microarray data. *Stat Sin* **12**: 31–46.
25. Broberg P (2003) Statistical methods for ranking differentially expressed genes. *Genome Biol* **4(6)**: R41.
26. Ge Y, Dudoit S and Speed T (2003) Re-sampling based multiple testing for microarray data analysis. *Test* **12**: 1–77.
27. Kendziorski C, Newton M, Lan H and Gould M (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med* **22(24)**: 3899–3914.
28. Reiner A, Yekutieli D and Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19(3)**: 368–375.
29. Smyth GK (2004) Linear models and empirical bayes methods for assessing

differential expression in microarray experiments. *Stat Applic Genet Mol Biol* **3(1)**: article 3.
30. Wang J and Kim SK (2003) Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development* **130(8)**: 1621–1634.
31. Park T, Yi S-G, Lee S, Lee SY, Yoo D-H, Ahn J-I and Lee Y-S (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* **19(6)**: 694–703.
32. Romagnolo B, Jiang M, Kiraly M, Breton C, Begley R, Wang J, Lund J and Kim SK (2002) Downstream targets of *let-60 Ras* in *Caenorhabditis elegans*. *Dev Biol* **247**: 127–136.
33. Scheffé H (1959) *The Analysis of Variance*. John Wiley & Sons, New York.
34. Tai YC and Speed TP (2004) A multivariate empirical Bayes statistic for replicated microarray time course data. Technical Report 667, Department of Statistics University of California, Berkeley, CA.
35. Jiang M, Ryu J, Kiraly M, Duke K, Reinke V and Kim SK (2001) Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* **98(1)**: 218–223.
36. Tai YC and Speed TP (2005) Longitudinal microarray time course MB-statistic for multiple sample groups. *Department of Statistics, University of California, Berkeley*. In preparation.
37. Tai YC and Speed TP (2005) *Cross-Sectional Microarray Time Course MB-statistic*. Department of Statistics, University of California, Berkeley. In preparation.
38. Guo X, Qi H, Verfaillie CM and Pan W (2003) Statistical significance analysis of longitudinal gene expression data. *Bioinformatics* **19(13)**: 1628–1635.
39. Storey J, Leek J, Xiao W, Dai J and Davis R (2004) A significant method for time course microarray experiments applied to two human studies. Technical Report 232, Department of Biostatistics, University of Washington, Seattle, WA.
40. Zhao LP, Prentice R and Breeden L (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc Natl Acad Sci USA* **98(10)**: 5631–5636.
41. Liang K-Y and Zeger S (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
42. Xu XL, Olson JM and Zhao LP (2002) A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Hum Mol Genet* **11(17)**: 1977–1985.
43. Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS and Simon I (2003) Continuous representations of time-series gene expression data. *J Comput Biol* **10(3-4)**: 341–356.
44. Luan Y and Li H (2004) Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics* **20(3)**: 332–339.
45. Hong F and Li H (2004) B-spline based empirical Bayes methods for identifying genes with different time-course expression profiles. Submitted.
46. Lawton W, Sylvestre E and Maggio M (1972) Self-modeling nonlinear regression. *Technometrics* **13**: 513–532.
47. Wang Y and Brown M (1996) A flexible model for human circadian rhythms. *Biometrics* **52(2)**: 588–596.
48. Bar-Joseph Z, Gerber G, Simon I, Gifford DK and Jaakkola TS (2003) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc Natl Acad Sci USA* **100(18)**: 10146–10151.
49. Lönnstedt IM, Grant S, Begley G and Speed TP (2003) Microarray analysis of two interacting treatments: a linear model and trends in expression over time. Technical report Department of Mathematics Uppsala University Sweden.

50. Fleury G, Hero A, Yoshida S, Carter T, Barlow C and Swaroop A (2002) Pareto analysis for gene filtering in microarray experiments. In: *Proceedings XI European Signal Processing Conference,* France.
51. Hero A and Fleury G (2002) Posterior pareto front analysis for gene filtering. In *Proceedings of Workshop on Genomic Signal Processing and Statistics (GENSIPS),* Raleigh, NC.
52. Hero A and Fleury G (2004) Pareto-optimal methods for gene ranking. *J VLSI Signal Process* **38**: 259–275.
53. Yuan M, Kendziorski C, Park F, Porter JL, Hayes K and Bradfield CA (2003) Hidden markov models for microarray time course data under multiple biological conditions. *Journal of the American Statistical Association*. To be published.
54. Möller-Levet CS, Cho K-H, Yin H and Wolkenhauer O (2003) Clustering of gene expression time-series data. Technical report Department of Computer Science University of Rostock, Rostock.
55. Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95(25)**: 14863–14868.
56. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* **96(6)**: 2907–2912.
57. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* **22(3)**: 281–285.
58. Reich M, Ohm K, Angelo M, Tamayo P and Mesirov JP (2004) GeneCluster 2.0: an advanced toolset for bioarray analysis. *Bioinformatics* **20(11)**: 1797–1798.
59. Saban MR, Hellmich H, Nguyen N-B, Winston J, Hammond TG and Saban R (2001) Time course of LPS-induced gene expression in a mouse model of genitourinary inflammation. *Physiol Genom* **5(3)**: 147–160.
60. Gurok U, Steinhoff C, Lipkowitz B, Ropers H-H, Scharff *C* and Nuber UA (2004) Gene expression changes in the course of neural progenitor cell differentiation. *J Neurosci* **24(26)**: 5982–6002.
61. Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Chan W, Botstein D and Brown P (2000) 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol* **1(2)**: research0003.1–research0003.21.
62. Chipman H, Hastie TJ and Tibshirani R (2003) Clustering microarray data. In: Speed T (ed.) *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC Press.
63. Fraley C and Raftery AE (2002) Model-based clustering discriminant analysis and density estimation. *J Am Stat Assoc* **97**: 611–631.
64. Ramoni MF, Sebastiani P and Kohane IS (2002) From the Cover: cluster analysis of gene expression dynamics. *Proc Natl Acad Sci USA* **99(14)**: 9121–9126.
65. Yeung KY, Fraley C, Murua A, Raftery AE and Ruzzo WL (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17(10)**: 977–987.
66. Schliep A, Schonhuth A and Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* **19(90001)**: 255i–263i.
67. Schliep A, Steinhoff C and Schonhuth A (2004) Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics* **20**: i283–i289.
68. Iyer VR, Eisen MB, Ross DT *et al*. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* **283(5398)**: 83–87.
69. Ji X, Li-Ling J and Sun Z (2003) Mining gene expression data using a novel approach based on hidden markov models. *FEBS* **542**: 125–131.

70. Zeng Y and Garcia-Frias J (2004) A new HMM-based clustering technique for the analysis of gene expression microarray time series data. In: *Currents in computational molecular biology*. RECOMB, San Diego, CA, p. G32.
71. Luan Y and Li H (2003) Clustering of time-course gene expression data using a mixed effects model with B-splines. *Bioinformatics* **19(4)**: 474–482.
72. Zhang Y, Hongyuan Z, Wang J and Chu C-H (2004) Clustering of time-course gene expression data. In: *Currents in computational molecular biology*. RECOMB, San Diego, CA, p. G34.
73. Peddada S, Lobenhofer E, Li L, Afshari C, Weinberg *C* and Umbach D (2003) Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19**: 834–841.
74. Wakefield J, Zhou C and Self S (2003) Modelling gene expression data over time: curve clustering with informative prior distributions. In: Bernardo J, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A and West M, (eds) *Bayesian Statistics* 7. Oxford University Press, Oxford, pp. 711–722.
75. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD and Vingron M (2001) Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA* **98(19)**: 10781–10786.
76. Tan Q, Brusgaard K, Kruse T, Oakeley E, Hemmings B, Beck-Nielsen H, Hansen L and Gaster M (2004) Correspondence analysis of microarray time-course data in case-control design. *J Biomed Inf* **37(5)**: 358–365.
77. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR and Fedoroff NV (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci USA* **97(15)**: 8409–8414.
78. Alter O, Brown PO and Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* **97(18)**: 10101–10106.
79. Aach J and Church GM (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17(6)**: 495–508.
80. Ihaka R and Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* **5(3)**: 299–314.
81. Wettenhall JM and Smyth GK (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* bth449.
82. Hartigan J and Wong M (1979) A K-means clustering algorithm. *Appl Stat* **28**: 100–108.
83. Heyer LJ, Kruglyak S and Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* **9(11)**: 1106–1115.
84. Balasubramaniyan R, Hullermeier E, Weskamp N and Kamper J (2004) Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics,* page bti095.
85. Maratou K, Forster T, Costa Y, Taggart M, Speed RM, Ireland J, Teague P, Roy D, and Cooke HJ. (2003) Expression profiling of the developing testis in wild-type and Dazl knockout mice. *Molecular Reproduction and Development* **67(1)**: 26–54.

Array CGH data analysis

21

Yuedong Wang and Sun-Wei Guo

21.1 Introduction

Like the spotted cDNA microarrays (1), array comparative genomic hybridization (CGH) also uses two differentially labeled test (unknown sample to be analyzed) and reference (known to be genomically normal) DNAs which are co-hybridized, under *in situ* suppression hybridization conditions, to cloned genomic fragments with known physical locations, spotted and immobilized on glass slides. The hybridized DNAs are then detected by their different incorporated fluorophores, and the ratios of the digitized intensity values in the hybridized patterns of the DNAs onto the cloned fragments are indicative of copy-number differences between the test and the reference genomes.

The detection of genomic alterations using array CGH requires careful statistical analysis of the intensity data from the two fluorochrome, since, besides genuine differences between the two genomes, stochastic fluctuations, measurement errors or other errors of unknown origins, and consistent, region-specific variations caused by differences in hybridization characteristics of the incorporated fluorochromes and by local variation in chromosomal structures, can all cause the ratio to deviate from unity (2).

For conventional CGH, a calibration process is usually invoked, in which reference versus reference hybridizations are performed to gauge the normal range of ratio variations (3). The ratios of the test-reference hybridizations, at each chromosomal segment where the ratio is calculated, are then compared with, say, the two standard deviations (SD) outside the mean, obtained from the calibration, and a gain or loss is declared if the ratio is above or under the two SDs (presumably the nominal 95% confidence bounds without multiple comparison adjustment) (4). Sometimes a pair of fixed, global thresholds, say, 1.15 and 0.85 (5, 6), are used in lieu of two SDs.

Recognizing the variable nature of the variance of the mean ratio within and between reference:reference hybridizations and possible inequality of variances of mean ratio between the test:reference and reference:reference experiments, a *t*-like statistic incorporating reference:reference and test:reference variations to detect genomic alterations segment by segment was proposed (7). This method, however, assumes that the ratio of the variances of test:reference ratio means and of reference:reference ratio means is constant across the whole genome, which may not be true. In addition, correlation in the estimated variances and the spatial correlation of ratios in the neighboring segments are completely ignored. Spatial correlations between neighboring clones can be prominent in array CGH data, since,

once a clone exhibits alteration, its neighboring clones also tend to have alterations (8). The spatial correlation among neighboring clones is expected to be high when the regions with genomic alterations are large, or when the density of the CGH array becomes high. With high-density CGH arrays containing 30 000 (9) or even 85 000 (10) oligonucleotides on a single chip with an average resolution of 30 kb or even higher (10) on the horizon, proper handling of spatial correlations becomes a pressing issue. Proper handling of spatial correlation may also increase statistical efficiency and improve precision in estimation, which, in turn, may translate into requirement for less calibration samples.

Besides the issue of spatial correlation in analysis of array CGH data, several additional considerations are in order. First, less restrictive assumptions on variance are preferable, since the variance may depend on the chromosomal structures and thus locations of the clones. Second, the nature of variance may vary from laboratory to laboratory due to considerable differences in the execution of array CGH experiments; less distributional assumption on the ratio would be preferable. Lastly, robustness to outliers and the minimization of the dominating effect of clones with very small variance would be desirable. Our recently proposed methods (11) are well adapted to spatial inhomogeneity as in array CGH data, and have been applied successfully to the identification of genomic alterations in the endometrium of patients with endometriosis (12).

21.2 Summary

Data and standard statistics

For simplicity, we use a BAC array data set to illustrate our methods (13). Our methods apply to more complicated designs with dyes and arrays as factors (11). We shall analyze one set of data presented by Snijders *et al*. (13), GM01524, which can be downloaded from the website http://genetics.nature.com/supplementary_info/. The data result from an experiment aimed at measuring copy number changes for the cell strain GM01524 (test sample) against a normal male reference DNA (reference), which were co-hybridized on a CGH array containing 2460 BAC and P1 clones in triplicate (7380 spots) and with an average resolution of ~1.4 Mb (13). We shall only focus on chromosome 6 for ease of exposition.

Array CGH data often have systematic biases as do cDNA microarray data (11). Therefore, the first step in analysis is to remove these biases using a normalization procedure such as *lowess* (see also Chapter 17). Details can be found in Yang *et al*. (14) or Wang and Guo (11).

Our methods apply to each chromosome separately to detect copy number changes at clones on the chromosome. For simplicity, we assume, in the following discussion, that all clones to be considered are on the same chromosome. After normalization, let y_{ijk} be the kth replication of the logarithm of the dye intensity of clone i of sample j, where $i = 1,\ldots, I$ represents observed clones in a chromosome, $j = 1,2$ represents two samples (test and reference), $k = 1,\ldots, n_j$ and n_j represents the number of replications of sample j.

Assume that $y_{ijk} \overset{iid}{\sim} N(\mu_{ij}, \sigma_{ij}^2)$. Then the question of whether there are any significant copy number differences between the two samples at clone *i* can be formalized by the hypothesis H_0: $\mu_{i1} = \mu_{i2}$ vs H_1: $\mu_{i1} \neq \mu_{i2}$. Let

$$\bar{y}_{ij\cdot} = \sum_{k=1}^{n_j} y_{ijk}/n_j,\ s_{ij}^2 = \sum_{k=1}^{n_j} (y_{ijk} - \bar{y}_{ij\cdot})^2/(n_j - 1),\ \mathrm{i} = 1,\ldots,I;\ j = 1,2.$$

The standard *z*-statistic

$$Z_i = \bar{y}_{i1\cdot} - \bar{y}_{i2\cdot}, \tag{21.1}$$

and the standard *t*-statistic

$$t_i = (\bar{y}_{i1\cdot} - \bar{y}_{i2\cdot})/\sqrt{s_{i1}^2/n_1 + s_{i2}^2/n_2}. \tag{21.2}$$

Clone *i* is declared to have significantly different intensity ratio and thus copy number between the two samples when the absolute value of the *z*-statistic or of the *t*-statistic is large. It should be noted that the *z*-test ignores the heterogeneous nature of variances associated with intensity ratios. The standard *t*-test accounts for the variation in the *z*-statistic and is easy to use since t_i approximately follows a Student *t* distribution with degrees of freedom $(s_{i1}^2/n_1 + s_{i2}^2/n_2)/((s_{i1}^2/n_1)^2/(n_1 - 1) + (s_{i2}^2/n_2)^2/(n_2 - 1))$. However, it has two fundamental problems: (i) the repetition numbers n_1 and n_2 are usually small (e.g. $n_1 = n_2 = 3$ in the example) because repetitive printing of the same clone on the slide limits the total number of clones to be printed on the slide. Even if multiple slides are used, the amount of DNA extracted from the test sample is often limited and thus only a few slides can be used for hybridization. Estimates of variances s_{i1}^2 and s_{i2}^2 are unreliable when sample sizes are small (11, 15); (ii) spatial correlations between neighboring clones are ignored, rendering the methods less efficient. Our methods aim to overcome these two problems by pooling information in neighboring clones to yield more stable estimates of the variances and to detect clones with copy number changes.

Smoothing the variances

From a modeling perspective, a chromosome to be analyzed can be practically viewed as continuous, and the clones, with known physical locations, are observation points interspersed along the chromosome. Within each chromosome, it is reasonable to assume that the variance is a smooth function of clone locations. Specifically, let $s_i = \sqrt{s_{i1}^2/n_1 + s_{i2}^2/n_2}$ be the standard error and x_i be the genome position of clone i, we assume that

$$\log(s_i) = h(x_i) + e_i,\ i = 1,\ldots,I, \tag{21.3}$$

where *h* is a smooth function. We fit Equation 21.3 using the robust *lowess* method with 30% of the data used for smoothing at each position. Logarithm of standard errors and *lowess* fit to Equation 21.3 are shown in *Figure 21.1*.

We then define a modified *t*-like statistic as

$$u_i = (\bar{y}_{i1\cdot} - \bar{y}_{i2\cdot})/\exp(\hat{h}(x_i)). \tag{21.4}$$

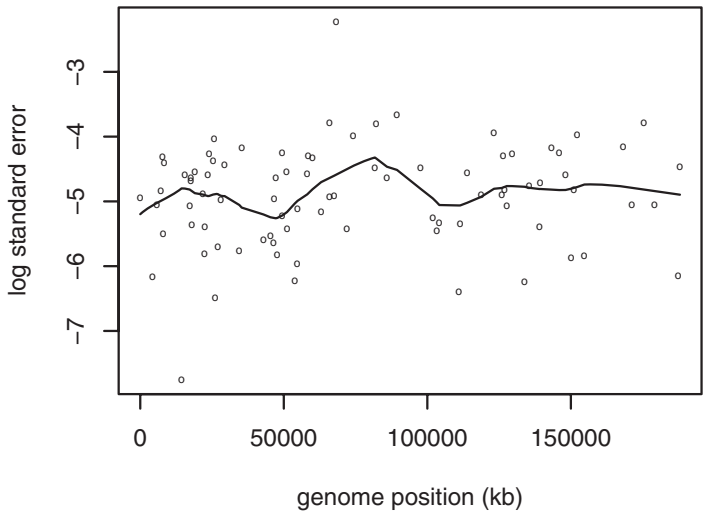


Figure 21.1.

Plot of logarithm of standard errors *vs.* genome positions as circles and the *lowess* fit as the solid line.

Replacing standard errors by their smoothed estimates also reduces the effect of outliers and prevents clones with very small variances from dominating the result.

Detecting locations and regions with different expression levels using hybrid adaptive spline

High-resolution mapping of specific regions is of crucial importance for the subsequent discovery of the disease-associated clones and thus the genes they harbor. The copy number changes in cancer often span large regions of the genome (16), although losses or gains of smaller scale and micro-deletions or micro-gains (from 100 bp to 4 Mb and not detectable by standard cytogenetic methods) are also of importance. Several methods have been proposed to take into account the spatial correlations such as likelihood based on a fixed-width window correlation structure (8), moving averages (17), CGH-Plotter (16), break point model (18) and cluster along chromosomes (19). In reality, correlations may vary with chromosomal structures and thus locations, and the sizes of segments harboring genomic alterations also vary along chromosomes. Therefore, assumptions of fixed correlation structures and difference shapes may be too restrictive.

We use the hybrid adaptive spline (HAS) that has the ability to handle a wide variety of shapes and spatial inhomogeneities (20). It is an objective approach that allows data to dictate the shape of a function. Let y_i be one of the z, t or u statistic defined in Equations 21.1, 21.2 and 21.4 respectively. We assume that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,\ldots,I, \tag{21.5}$$

where f is a function of x_i and ε_i's are random errors with mean zero and variance σ^2. For simplicity we transformed the variable x into the interval [0,1].

Since copy number changes occur in local regions, the expectation of a ratio profile along the chromosome equals zero except in some regions that harbor the changes. Thus our goal is to detect locations or regions where $f(x) \neq 0$. Since f may have discontinuous points and is spatially inhomogeneous, common nonparametric regression methods such as smoothing spline do not provide a good estimate for f. Spatially adaptive, HAS was proposed to handle spatial inhomogeneity as in array CGH data. For simplicity, we introduce the HAS procedure using cubic spline bases. The HAS procedure for general spline bases can be found in Luo and Wahba (20).

The HAS procedure

Let $\phi_1(x) = 1$, $\phi_2(x) = x - 0.5$ and $\xi_i(x) = \int_0^{min(x_i,x)} (x_i - u)(x - u)du$.

1. *Initialization:* set the maximum number of bases q ($q \geq 2$) and the inflated degrees of freedom (IDF). Start with $k = 2$ and two bases $\{\phi_1(x),\phi_2(x)\}$.
2. *Forward stepwise selection:* for $k = 3,\ldots, q$, choose the kth basis ξ_{ik} (x) to maximize the reduction in the residual sum of squares (RSS).
3. *Optimal number of bases:* choose $k \geq 2$ as the minimizer of the generalized cross-validation (GCV) score

$$GCV(k) = RSS/(1 - (2 + (k - 2) \times IDF)/I)^2.$$

4. *Backward elimination:* perform backward elimination to the selected bases. Decide the final number of bases by the Akaike Information Criteria (AIC).
5. *Fit:* fit a standard or ridge regression model to the final selected bases.

The key to spatial adaptiveness is to select bases adaptively based on data. The IDF is used to account for the added flexibility in adaptively selected bases. Luo and Wahba (20) suggested the use of IDF=1.2. For array CGH data, we found that this choice of IDF sometimes under- or over-estimates the number of bases. Our experiences suggest that the combination of a smaller IDF (1 or 1.1) with the backward elimination step provides better fits. We also found that the ridge regression step in the original HAS procedure can lead to over-smoothing for array CGH data. Therefore we recommend the standard regression using a numerically stable procedure.

We use the following bootstrap procedure to calculate p-values. Denote the HAS estimates of f and σ as $\hat{f}$ and $\hat{\sigma}$ respectively. We first generated a bootstrap sample

$$y_i^* = \hat{f}(x_i) + \varepsilon_i^*, \quad i = 1,\ldots, I,$$

where ε_i^* are sampled with replacement from residuals. Denote the HAS estimates of f and σ based on the bootstrapped sample as $\hat{f}^*$ and $\hat{\sigma}^*$ respectively. Let $D_i^* = (\hat{f}^*(x_i) - \hat{f}(x_i))/\hat{\sigma}^*$. Repeat this process B times and denote $D_i^*(b)$ as the D_i^* statistic based on the bth bootstrapped sample. We then calculate the p-values as

$$p_i = \#\{b : |D_i^*(b)| > |\hat{f}(x_i)|/\hat{\sigma}\}/B.$$

Then clones with $p_i \leq \alpha$ are significant at level a. False discovery rate (FDR) can be used to circumvent the problem of multiple comparisons (11).

Simulations in Wang and Guo (11) indicated that the modified *t*-like statistic based on smoothed variance always improves the performance. The HAS procedure is more powerful in detecting clustered locations while a separate *t*-test is more powerful in detecting isolated locations.

Results

We now apply the HAS procedure to three statistics defined in Equations 21.1, 21.2 and 21.4. *Figure 21.2* shows these statistics, the HAS fits, *p*-values and locations with significant change in copy numbers between the two samples. It is obvious that the profile of the standard *t*-statistic is rougher than those of the other two statistics in the region between 100 000 kb and 150 000 kb. Standard *t*-test would miss some of the locations in this region. For all three statistics, the HAS procedure identified the whole trisomic region 6q15-6q25 which agrees with the results in Snijders *et al*. (13). Wang *et al*. (19) identified the same region by first building a hierarchical clustering tree and then selecting the 'interesting' clusters. It is interesting to note that, besides the documented trisomy in the 6q15-25 region, the HAS procedure also identified an isolated gain in the 6p region that is previously undocumented. It is likely that the gain, due to its apparently small size, may be too small to be detected by standard cytogenetic karyotyping methods.

21.3 Concluding remarks

Our proposed methods for the identification of genomic alterations using array CGH have several advantages. First, they require much less restrictive assumptions on the variances of the test:reference ratios. Second, by smoothing variances along the genome, the smoothed *t*-like statistics are more robust to outliers. This is especially important for experiments with a relatively small number of control samples and/or ratios that follow a distribution with tails that decay much slower than the Gaussian distribution. Third, by incorporation of neighboring data with high correlations, HAS is more efficient and robust in detecting clusters of alterations. It handles nicely the inhomogeneous 'curvature' of the ratio profiles along the genome. Our inference procedure based on bootstrap does not require the normality assumption. In view of the observation that there are consistent, region-specific variations in ratio profiles, which may be caused by differences in hybridization characteristics of the incorporated fluorochromes and by local variation in chromosome structures (such as telomeres or centromeres) (7), and, in particular, the functional form of the variation as a function of clone locations is typically unknown, the HAS procedure is well suited for the array CGH data.

As a high-throughput and high-resolution genetic method for identification of whole-genome copy number alterations, array CGH holds a great potential in uncovering genes involved in disease initiation and progression in cancer and other diseases. Better and higher-resolution CGH arrays are on the horizon.

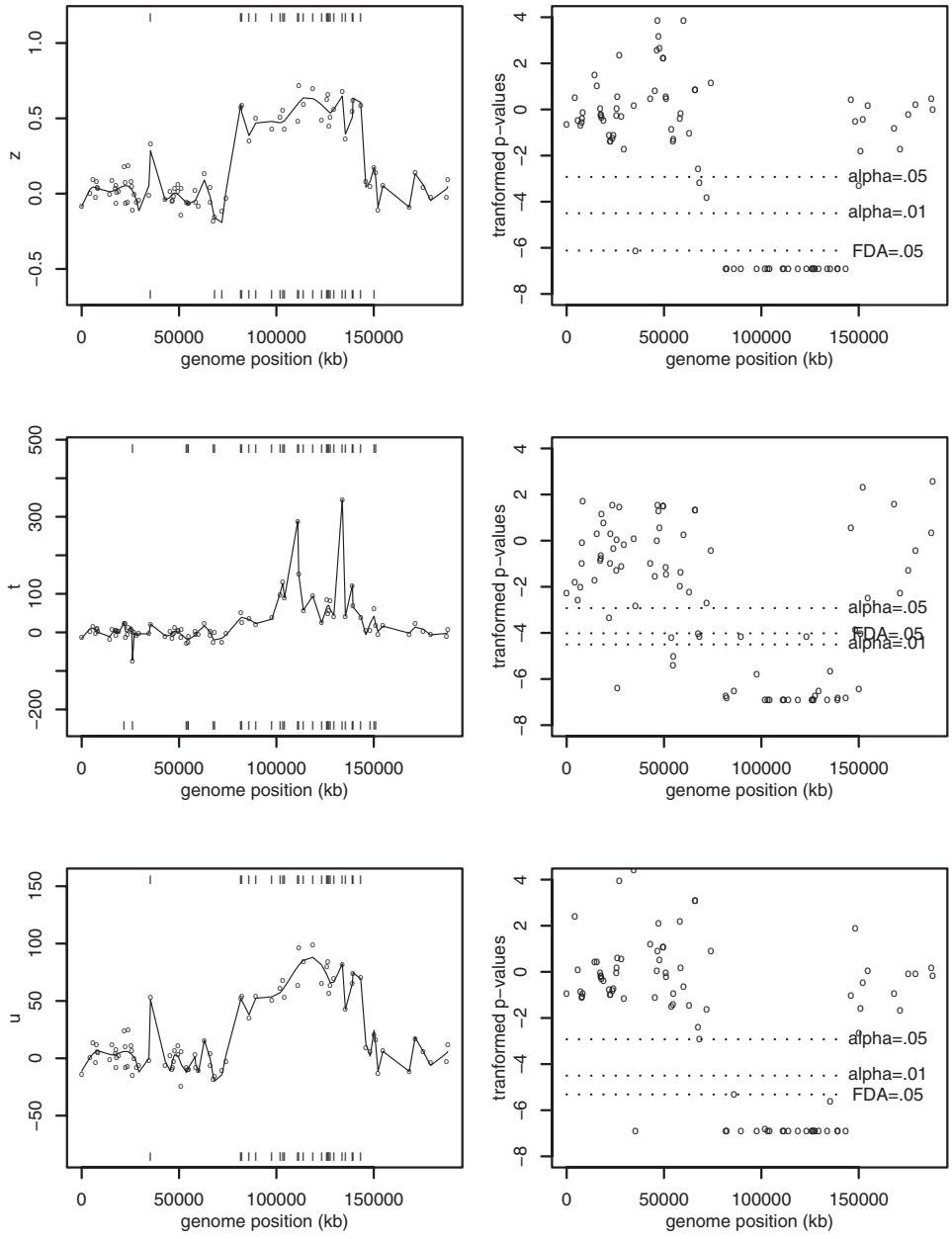


Figure 21.2.

Results for chromosome 6 of the cell strain GM01524. Three plots on the left panel are *z*-statistics (top), *t*-statistics (middle) and *u*-statistics (bottom) as circles and HAS fits as solid lines. Locations with significant change in copy numbers at 5% (1%) level based on bootstrap *p*-values with $B = 10\,000$ are marked at the bottom (top) of the plot. Three plots on the right panel are transformed *p*-values, $\log((p + a)/(1 - p + a))$ with $a = 0.001$, as circles. Regions below three dotted lines represent rejection regions with $\alpha = .01$ and $\alpha = .05$ and $FDR \leq .05$ respectively which are marked at the right end of these lines.

These technical developments may reduce some systematic biases as encountered before, but proper handling of spatial correlations becomes increasingly important. In addition, the recently documented copy number polymorphism in the human genome (21) indicates that large-scale copy number polymorphisms (about 100 kb or greater) are not uncommon in normal humans. This finding underscores the importance of experimental designs and proper choice of reference samples in array CGH studies, and poses a challenge in identification of copy number changes responsible for disease initiation and progression.

Acknowledgments

This work was supported by an NIH Grant R01 GM58533 (YW) and a grant from the Wisconsin Children's Hospital Foundation (SWG).

References

1. Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* **270**: 467–470.
2. Piper J, Rutovitz D, Sudar D, Kallioniemi A, Kallioniemi OP, Waldman, FM, Gray JW and Pinkel D (1995) Computer image analysis of comparative genomic hybridization. *Cytometry* **19**: 10–26.
3. Cher ML, Bova GS, Moore DH, Small EJ, Carroll PR, Pin SS, Epstein JI, Isaacs WB and Jensen RH (1996) Genetic alterations in untreated metastases and androgen-independent prostate cancer detected by comparative genomic hybridization and allelotyping. *Cancer Res* **56**: 3091–3102.
4. du Manoir S, Kallioniemi OP, Lichter P *et al*. (1995) Hardware and software requirements for quantitative analysis of comparative genomic hybridization. *Cytometry* **19**: 4–9.
5. Lundsteen *C,* Maahr J, Christensen B, Bryndorf T, Bentz M, Lichter P and Gerdes T (1995) Image analysis in comparative genomic hybridization. *Cytometry* **19**: 42–50.
6. Schleger *C,* Arens N, Zentgraf H, Bleyl U and Verbeke C (2000) Identification of frequent chromosomal aberrations in ductal adenocarcinoma of the pancreas by comparative genomic hybridization (cgh). *J Pathol* **191**: 27–32.
7. Moore D, Pallavicini M, Cher ML and Gray JW (1997) A *t*-statistic for objective interpretation of comparative genomic hybridization (cgh) profiles. *Cytometry* **28**: 183–190.
8. Carothers A (1997) A likelihood-based approach to the estimation of relative DNA copy number by comparative genomic hybridization. *Biometrics* **53**: 848–856.
9. Carvalho B, Ouwerkerk E, Meijer G and Ylstra B (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol* **57**: 644–646.
10. Lucito R, Healy J, Alexander J *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* **13**: 2291–2305.
11. Wang Y and Guo SW (2004) Statistical methods for detecting genomic alterations through array-based comparative genomic hybridization (CGH). *Front Biosci* **9**: 540–549.
12. Guo SW, Wu Y, Strawn E, Basir Z, Wang Y, Halverson G, Montgomery K and Kajdacsy-Balla A (2004) Genomic alterations in the endometrium may be a proximate cause for endometriosis. *Eur J Obstet Gynecol Reprod Biol* **116**: 89–99.

13. Snijders AM, Nowak N, Segraves R *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29**: 263–264.
14. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**: E15.
15. Comander J, Natarajan S, Gimbrone MA and Garcia-Cardena G (2004) Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genom* **5**: 17.
16. Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M and Kallioniemi A (2003) Cgh-plotter: MATLAB toolbox for CGH-data analysis. *Bioinjormatics* **19**: 1714–1715.
17. Kraus J, Pantel K, Pinkel D and Speicher MR (2003) High-resolution genomic profiling of occult micro metastatic tumor cells. *Genes Chromosomes Cancer* **36**: 159–166.
18. Jong K, Marchiori E, Meijer G, Van Der Vaart A and Ylstra B (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* **20(18)**: 3666–3637.
19. Wang P, Kim Y, Pollack J, Narasimhan B and Tibshirani R (2005) A method for calling gains and losses in array cgh data. *Biostatistics*. **6**: 45–48.
20. Luo Z and Wahba G (1997) Hybrid adaptive splines. *J Am Stat Assoc* **92**: 107–116.
21. Sebat J, Lakshmi B, Troge J *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

MIAME

22

Robert Wagner

22.1 Introduction

Microarray data analysis is not only a matter of statistical methods but also a question of how to interpret the results of the performed experiment. Gene expression data are large and complex and only meaningful in the context of the specific conditions under which they have been generated. In many cases publicly available data is poorly annotated, in a way that important information about the quality of the data or about the normalization and transformation procedures are missing. This makes the interpretation and comparison of the data difficult and unreliable.

In order to cope with these problems, the Microarray Gene Expression Database group (MGED; http://www.mged.org/) is developing standards for microarray data. MIAME (1), the Minimum Information About a Microarray Experiment, currently in version 1.1, aims to outline the minimum information required in order to unambiguously interpret and potentially reproduce a microarray-based gene expression experiment. It is a set of guidelines, which specifies the content that should be provided and should not be confused with a description format. A data model and exchange format, MAGE (2), which can be used to store the information covered by MIAME has also been developed by MGED and has become an Adopted Specification of the Object Management Group (OMG) standards group. Some of the terminology introduced by MAGE will be used in this chapter in order to explain the elements of MIAME.

22.2 The structure of MIAME

The MIAME structure is divided into two major sections:

- array design description;
- experiment description.

This reflects the fact that arrays are often manufactured independently of concrete experiments in which they are used. Therefore array designs can be described in a separate section, which is independent of the experimental part, and can then simply be referenced from the latter. In general, all parts of a MIAME description should be given unique identifiers in order to reference them in the corresponding sections of the experiment description.

While some elements of MIAME, especially protocols, are described as free text, most sections are required to use controlled vocabularies or external ontologies. This is essential in order to allow database queries and

automated analysis on the data. Here, the MGED group has developed the ontology MO, which is recommended to be used whenever appropriate. If MO or other external ontologies are not feasible for the data to be described, MIAME requests the user to apply their own qualifiers and values in the format of:

(*qualifier, value, source*)

triplets, where qualifier specifies the name of the element, value its current instantiation and source the reference, where the controlled vocabulary is described. For instance,

(*qualifier*: 'organism', *value*: 'homo sapiens', *source*: 'NCBI taxonomy').

In the following sections the elements of MIAME are outlined in detail giving a 'checklist' for the user to verify the completeness of a microarray experiment description according to the minimum information required. More information about MIAME and software that is related to it can be found at http://www.mged.org/miame/.

22.3 Array design description

The array design describes the characteristics of the arrays used in the experiment. The section consists of two parts describing the array as a whole and its design elements (e.g. its spots). The description of the design elements is separated into three different classes: feature, identifying the location on the array; reporter, describing the nucleotide sequence belonging to it; and composite sequence, summarizing the corresponding gene, exon or splice-variant. Additionally information about control elements is to be provided. This second part of the array design description will typically be provided as a spreadsheet or tab-delimited file. Often it will be available from the array providers, in which case they can simply be referenced. Sometimes this part might be difficult to acquire, for instance if commercial array manufacturers only provide information on the composite sequence level and not about the reporters. However, it was recently agreed that this part definitely belongs to MIAME and is necessary when providing information about the array and its elements.

1. The minimum information about the array design includes:
   - the name of the array design;
   - the platform type (e.g. spotted glass array, *in situ* synthesized array);
   - surface and coating specifications (e.g. glass, membrane);
   - physical dimension of array support (e.g. of the slide);
   - the number of features on the array;
   - availability (for commercial arrays) or production protocol for the array.
2. The minimum information about the features includes:
   - dimensions;
   - attachment (e.g. covalent, ionic);
   - a reference to the corresponding reporter(s).
3. The minimum information about the reporters includes:
   - the type of reporter (e.g. cDNA, synthetic oligonucleotide);

- single or double-stranded;
- the sequence, accession numbers in DDBJ/EMBL/GenBank and primer pair information;
- the approximate length if the exact sequence is unknown;
- clone information (id, provider, date, availability);
- a protocol describing the production of the element in case of a custom-made array;
- a reference to the corresponding composite sequence(s).

4. The minimum information about the composite sequences includes:
   - the name of the gene;
   - the reference sequence;
   - links to appropriate databases (e.g. SWISS-PROT) if relevant.
5. The minimum information about control elements includes:
   - the position of the feature (the coordinate on the array);
   - the control type (e.g. spiking, normalization);
   - the control qualifier (endogenous, exogenous).

22.4 Experiment description

An experiment is a virtual compilation of biological materials and processes performed with them (hybridizations) in order to address a biological question. This second major part of the MIAME description contains the following four subparts:

1. Experimental design.
2. Samples used, extract preparation and labeling.
3. Hybridization procedures and parameters.
4. Measurement data and specifications of data processing.

Experimental design

The experimental design characterizes the experiment as a whole and describes some general information that is necessary in order to query for an experiment by certain parameters. The minimum information required in this section is:

- contact information, authors, publishers;
- the type(s) of the experiment (e.g. compound treatment design);
- experimental factors, that is parameters or conditions tested (e.g. time, dose, response to a compound or treatment);
- the number of hybridizations performed;
- whether and which common reference was used for all hybridizations;
- quality control steps taken (e.g. replicates, dye swap);
- a brief free text description of the experiment and its goals;
- external links (URL) or database accession numbers for further information.

Samples used, extract preparation and labeling

A sample is the labeled nucleic acid that is used together with an array in the hybridization process. The MIAME section about samples describes their

biological origins as well as treatments performed on the initial material and resulting products from each intermediate step.

1. The minimum information about the biological origin of a biomaterial is:
   - the name of the organism (NCBI taxonomy);
   - the sample provider;
   - cell type;
   - sex;
   - age;
   - developmental stage;
   - organism part (tissue);
   - genetic variation (e.g. gene knockout, transgenic variation);
   - individual genetic characteristics (e.g. disease alleles, polymorphisms);
   - animal/plant strain or line;
   - disease state or normal;
   - links to clinical information (if available).
2. The minimum information about biomaterial manipulations is:
   - growth conditions;
   - *in vivo* treatments (organism or individual treatments);
   - *in vitro* treatments (cell culture conditions);
   - treatment type (e.g. small molecule, heat shock, cold shock);
   - separation technique (e.g. none, trimming, micro-dissection);
   - compound.
3. The minimum information about hybridization extract preparation is:
   - the extraction method;
   - nucleic acid extracted (total RNA, mRNA or genomic DNA);
   - amplification (e.g. none, RNA polymerase, PCR).
4. The minimum information about the labeling process is:
   - the amount of nucleic acids labeled;
   - the label used (e.g. Cy3, Cy5, 32P, 33P);
   - label incorporation method.
5. The minimum information about added external controls is:
   - the element on the array expected to hybridize to a spiking control;
   - spike type (e.g. oligonucleotide, plasmid DNA, transcript);
   - spike qualifier (e.g. concentration, expected ratio).

Hybridization procedures and parameters

Hybridization is the process of joining the complementary nucleic acid strands of the labeled target and the probes on the array. The minimum information about hybridization is:

- the labeled extract (sample) used in the hybridization;
- the array used in the hybridization;
- the solution (e.g. concentration of solutes);
- blocking agent;
- wash procedure;
- the amount of labeled extract used;
- time, concentration, volume, temperature;
- a description of the hybridization instruments;
- a detailed free-text protocol on how the hybridization was performed.

Measurement data and specifications of data processing

The section about measurements describes the data that was obtained as a result of the hybridization process(es) and consists of three parts: the raw data originating from scanning: the intensities measured by the image analysis; and the gene expression levels after normalization and transformation.

1. The minimum information about the raw data is:
   - the hardware used for scanning (make and model);
   - the software used for scanning;
   - a detailed protocol describing the scanning process including scan parameters (laser power, spatial resolution, pixel space, PMT voltage);
   - (the image file resulting from the scanning of the hybridized microarray).

There is no consensus among MGED as to whether the scanned image files are part of MIAME or not.

2. The minimum information about the measured intensities is:
   - the complete output of the image analysis software for each element;
   - used image analysis software, version and availability;
   - parameters used for the image analysis.
3. The minimum information about the data transformation process is:
   - the gene expression data matrix summarizing related elements;
   - a detailed description of the data selection and transformation procedures;
   - some reliability indicators for each data point (e.g. standard deviation).

References

1. Brazma A, Hingamp P, Quackenbush J, *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat Genet* **29**: 365–371.
2. Spellman PT, Miller M, Stewart J, *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3(9)**: RESEARCH0046.

Index