2

Basic Microarray Analysis

Strategies for Successful Experiments

Scott A. Ness

Summary

Microarrays offer a powerful approach to the analysis of gene expression that can be used for a wide variety of experimental purposes. However, several types of microarray platforms are available. In addition, microarray experiments are expensive and generate complicated data sets that can be difficult to interpret. Success with microarray approaches requires a sound experimental design and a coordinated and appropriate use of statistical tools. Here, the advantages and pitfalls of utilizing microarrays are discussed, as are practical strategies to help novice users succeed with this method that can empower them with the ability to assay changes in gene expression at the whole-genome level.

Key Words: Microarrays; Affymetrix; GeneChips; genomics; gene expression; transcription; clustering; normalization; data analysis; hybridization.

1. Introduction

The large-scale genome-sequencing projects have identified most or all of the genes in humans, mice, rats, yeast, and a number of other commonly used experimental systems. At the time of this writing, the publicly available human genome information available from the National Center for Biotechnology Information includes more than 2.8×10^9 nucleotides of finished, annotated DNA sequence. Although the exact number of genes continually fluctuates as annotation and gene prediction programs change and improve, the current number of human genes is nearly 43,000 (Human genome build 34, version 3). (Information about the current human genome build is available at www.ncbi.nlm.nih.gov.) Microarrays provide a means of measuring changes in expression of all the genes at once. This ability provides researchers with enormous potential to perform experiments that were impossible just a few years ago and also offers unique challenges in experimental design and data analysis.

From: *Methods in Molecular Biology, vol. 316: Bioinformatics and Drug Discovery*
Edited by: R. S. Larson © Humana Press Inc., Totowa, NJ

Microarray experiments and the laboratories that perform them can be divided into several categories. First are the laboratories that specialize in microarray technology and that perform experiments with hundreds of microarray samples. Such research groups are often responsible for developing new methods of microarray data analysis and include dedicated groups of biostatisticians and computer programmers working to improve the statistical methods and computer programs for analyzing complex data sets generated by very large microarray experiments. A second class of laboratories has performed dozens of microarray experiments and has already become familiar with the data analysis tools necessary to accomplish their goals. Such laboratories generally make use of commercial software for data analysis or "freeware" packages written by the aforementioned large groups. This chapter is geared toward the third group: the laboratories that are considering their first microarray experiments and that need help with experimental design and data analysis. New users are most likely to rely on a core facility to actually perform the microarray experiments. For that reason, I do not discuss the details of manufacturing, manipulating, hybridizing, and scanning the microarrays here. Instead, the goal is to outline the potential benefits and pitfalls that arise with microarray experiments in order to help new users avoid common mistakes and reap the most benefit from experiments that can be very expensive and time-consuming. In addition, the commercial microarray platform offered by Affymetrix (Santa Clara, CA) is the most dominant and readily available means for new users to begin performing microarray experiments. Consequently, this chapter focuses on the Affymetrix platform and its use in the academic laboratory, although most or all of the discussion also applies to custom spotted arrays produced by local microarray facilities. There is a wide variety of uses for microarrays, including detection of single nucleotide polymorphisms, analysis of alternative RNA splicing, and analysis of transcription factors binding to promoters (ChIP on a Chip). However, here the discussion is limited to the use of microarrays for gene expression analysis, the most common use of the platform and the most likely way that new users will be tempted to use microarray technology.

2. When Is a Microarray the Best Approach?

Microarray experiments are extremely powerful and provide researchers with a new and exciting means of tackling important problems on a genomewide scale. Most microarrays contain probes for 10,000–40,000 different genes, allowing researchers to assess simultaneously changes in expression of nearly all the genes in the genome. However, they are also complex, time-consuming, and often very expensive experiments, and they generate large and complicated data sets that require substantial effort to analyze and validate. For these reasons, researchers should not be lured into performing microarray experiments without spending some time considering other options or without considerable thought regarding appropriate experimental design. New users should consult extensively with their local microarray core facility before beginning to prepare samples for microarray analysis. Every microarray facility can tell stories about users who approached them with samples only to find out that unsuitable preparation or storage had resulted in RNA that was too degraded for high-quality analysis. Proper preparation and storage of the RNA is crucial to the success of microarray experiments.

This is especially true for samples derived from patients or tissues that are difficult or impossible to replace. The microarray facility should be able to guide users to the best methods for preparing samples and storing the RNA to ensure that their experiments will succeed. Because of these limitations, some experiments are better suited for microarray analysis than others.

2.1. For Better or Worse: What Can Microarrays Do?

Microarray technology has proven to be extremely powerful for following changes in gene expression that occur as synchronized cells progress through the cell cycle ***(1,2)*** or when tissue culture cells are treated with a drug ***(3,4)*** or are infected with a virus expressing a recombinant transcription factor ***(5,6)***. In such situations, all the cells in the population are responding in parallel and relatively synchronously, and the microarrays, which measure the average change in gene expression in the population of cells being studied, can detect changes in gene expression that occur simultaneously in all the cells. Because of variations in measurements, microarrays are best at detecting changes that are relatively robust—a twofold or greater change is a common benchmark—in genes that are expressed at relatively high levels. Cells from different individuals, such as different patients, can display markedly different gene expression patterns, so microarrays perform best when the samples are closely related, such as tissue culture cells or treated vs untreated cells from a single patient or animal. Because different cell types display complex differences in gene expression patterns, heterogeneous samples, such as solid tumors or tissue samples, give complex microarray results. Optimum results are obtained from homogeneous samples, such as cell lines or purified cell populations, when they are available.

Some experiments are poorly suited for microarray analysis or need a modified design to make them work. For example, many researchers would like to transfect tissue culture cells with a plasmid expressing a molecule of interest and then use microarrays to measure subsequent changes in gene expression. The problem with this approach is that transfections are often inefficient and generally only yield 5–10% of cells expressing the molecule of interest. Because microarrays measure the average changes in gene expression in all the cells in the culture, a gene would have to be induced at least 20-fold in the transfected cells to show up as twofold induced when averaged over the entire cell population. A better design would be to transfect the cells with a plasmid that expresses the protein of interest as well as green fluorescent protein or some other marker that would allow the transfected (e.g., green fluorescent protein-positive) cells to be purified by flow cytometry before performing the microarray analysis. Alternatively, recombinant adenoviruses or some other method of expressing the protein of interest in nearly 100% of the cells in the culture could be used in place of transfection ***(5,6)***. The goal is to compare the changes in gene expression in one nearly homogeneous population with those in another.

Changes in gene expression patterns have been used to provide evidence that particular biochemical, signaling, or transcription factor pathways are activated or inhibited in different cell types ***(7,8)***. Microarrays can detect subtle changes in gene expression induced by a variety of extracellular or environmental stimuli ***(9,10)***. However, such

results can be quite complicated. In general, microarray experiments should be designed with some hypothesis in mind, rather than just as a "fishing" experiment. By testing a hypothesis, it will be possible to design positive and negative controls that will greatly facilitate the data analysis. This is discussed in more detail under **Heading 4**.

3. Choosing a Microarray Platform

The first choice a new user will have to make is which type of microarray to use. Essentially, microarrays are thousands of spots or probes immobilized on a solid surface such as glass or silicon that can be hybridized simultaneously to fluorescently labeled experimental samples, referred to as targets. In the simplest scenario (**Fig. 1**), mRNA from each sample is used as template in a complementary (c)DNA synthesis reaction that includes dinucleotide triphosphates labeled with fluorescent tags, usually Cy3 or Cy5. The resulting fluorescent target cDNA is hybridized to the microarray, which contains cDNA or oligonucleotide probes for each gene of interest. Usually, a separate microarray is used for each experimental sample. After washing, a laser scanner is used to measure the fluorescence at each spot, and the data are converted into a spreadsheet format showing the relative intensity or expression of each gene. Several variations on this theme provide increased sensitivity or reproducibility. For example, in the Affymetrix GeneChip system, the target samples are labeled with biotin and are detected with fluorescent streptavidin. However, even from this simple description of microarray technology, it is apparent that the most important parameters in the assay are the quality of the samples, the efficiency of the labeling with fluorescent nucleotides, and the quality and reproducibility of the gene-specific probes on the microarray.

3.1. Glass Slide Arrays

The first microarrays were produced by using modified writing pens to spot samples of DNA directly onto glass microscope slides. After chemical or ultraviolet (UV) cross-linking to fix the DNA to the glass, the fluorescently labeled cDNAs were applied in a drop of hybridization buffer, covered with a standard cover slip, and allowed to hybridize overnight. This is still the basic process for most microarrays produced in core facilities, although the machines that make the arrays have become highly automated and new chemistries and surfaces have been developed to make the glass slides more efficient at binding the DNA and to decrease the background in the hybridization. There are also differences in what types of DNA probes are attached to the glass.

3.1.1. cDNA Arrays

The first laboratories that made extensive use of microarrays spotted libraries of cDNA clones, either polymerase chain reaction (PCR)-amplified inserts or whole plasmids, directly onto glass slides. The use of relatively long (>300 bp) cDNAs has advantages and disadvantages. The biggest advantage is that the hybridization is quite robust. Thus, point mutations or even small deletions that might occur in some individuals will have little or no impact on the results of the hybridization. This feature makes cDNA arrays quite useful for studies of large sets of human patients who might have minor differences in some of their genes. Another advantage of using cDNAs is the relatively low

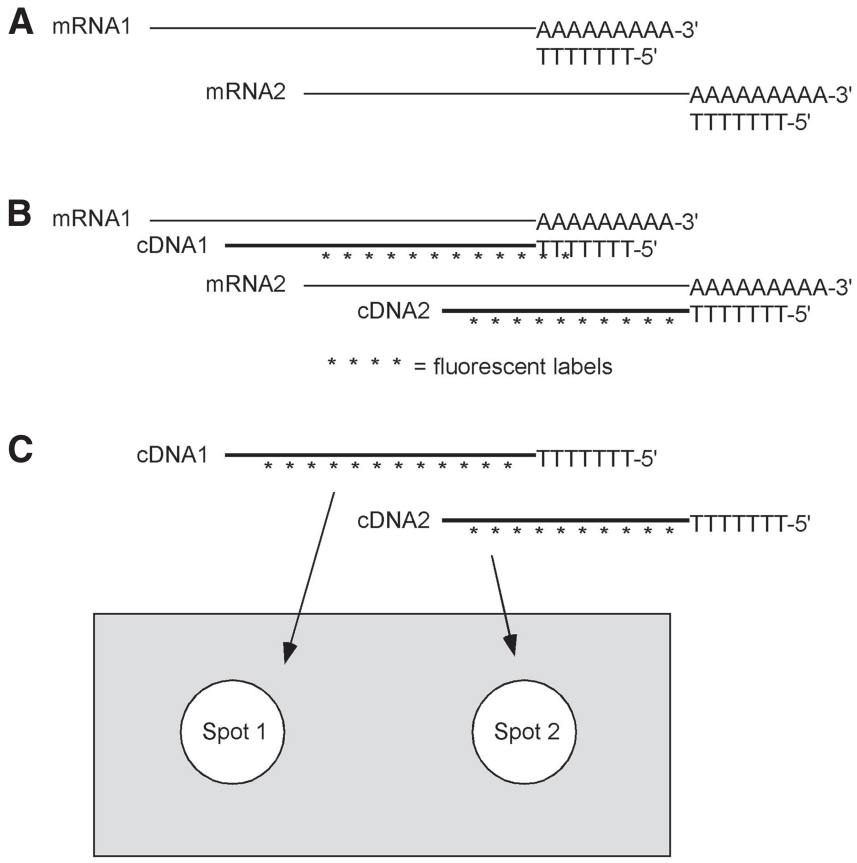

Fig. 1. Basic steps in microarray analysis. (**A**) Starting RNA. Purified mRNA is annealed with a primer (oligo-dT), ready for the reverse transcription reaction. At this point, control RNAs are often added ("spiked in") to give an internal control for the efficiency of the following steps. (**B**) Labeled cDNAs. In the simplest case, reverse transcription is performed using fluorescently tagged (e.g., Cy3 or Cy5), dinucleotide triphosphates (dNTPs), resulting in the generation of fluorescent cDNA. In the Affymetrix system, the dNTPs are biotinylated, and later detection is performed with fluorescent streptavidin. (**C**) Target hybridization. The fluorescently labeled cDNAs, referred to as "targets," are hybridized to gene-specific "probes." Each target anneals to its corresponding probe spot on the microarray. The probes can be spotted cDNAs or oligonucleotides, or oligonucleotides that were synthesized directly on the microarray surface. Although only two spots are shown, a single microarray can have probes for up to 50,000 different genes and more than a million spots per square inch. After hybridization, a laser scanner is used to detect the specific fluorescence at each spot. If all goes well, fluorescence intensity is proportional to the concentration of the relevant mRNA in the original sample.

cost. A single miniprep or PCR reaction can generate enough purified DNA to produce many thousand microarrays. Owing to their long lengths, the spotted cDNAs are likely to detect all the transcripts, such as different versions produced through alternative

promoter use or alternative RNA splicing, which could be an advantage. Major disadvantages are the cost and effort required to assemble large libraries of purified cDNAs or PCR products, all of which must be correctly identified, subjected to nucleotide sequencing, and annotated. There have been problems with collections of cDNAs provided by commercial suppliers, which contain a large fraction of clones that are improperly identified or contaminated with other plasmids that interfere with PCR amplification. In addition, cDNAs can contain repeated sequences and may hybridize to closely related transcripts (gene families) and so may not provide enough specificity for many applications. Because cDNAs vary in length and G-C content, it is difficult to ensure that all will hybridize equally well or give the same amount of background signal. These disadvantages make cDNAs difficult to work with and have contributed to their waning popularity.

3.1.2. Oligonucleotide Arrays

The most common type of glass slide microarrays use custom oligonucleotides, usually 40 to 60 mer, instead of cDNAs. The oligonucleotides, if designed properly, can overcome problems of specificity and G-C content associated with using cDNAs. There are generally fewer problems with improper identification or labeling when ordering custom oligonucleotides from commercial suppliers, although one must still trust the supplier to synthesize and purify them correctly and to put the correct oligonucleotides in each tube. The major drawbacks of using oligonucleotides are the relatively high cost of purchasing 10,000 or more custom oligonucleotides and the huge amount of bioinformatics support required to design all the necessary bits of DNA specific for each gene with matched G-C content and free of hairpins that could affect hybridization efficiency. Depending on how the oligonucleotides are designed, they might still suffer from some of the specificity problems associated with cDNAs. Complete sets of oligonucleotides are now available from commercial suppliers, greatly simplifying their use by microarray core facilities producing homemade microarrays. Nevertheless, because of the cost, it is rare for such collections to contain intentionally more than one oligonucleotide representing each gene.

3.1.3. Advantages and Disadvantages of Glass Slide Microarrays

The biggest advantage of using homemade or in-house-produced glass slide microarrays is the relatively low cost, generally less than $100 per array. However, such arrays are limited to 20,000 or fewer spots per array, so more than one array is necessary to screen an entire mammalian genome. It is also rare to have more than one spot for any gene on each array. Thus, if there are any discrepancies in the production of the arrays, such as some spots that get too little DNA or that are misshaped or smeared, there are no backup spots from that gene to confirm the hybridization results. Unfortunately, not all spots are identical on any spotted array, which makes the data subject to more variability and also makes multiple hybridizations absolutely essential. As a consequence, most users hybridize their samples to several identical arrays in order to have multiple measurements and to be able to perform statistics for each spot. This increases the cost substantially. Thus, if a single array costs $80, two arrays are necessary to rep-

Table 1
Comparison of Microarray Platforms

	Glass slide arrays	Affymetrix GeneChips
Typical cost per array	$80	$450
Arrays per 40,000 genes	2	1
Measurements per gene on each array	1	12
Arrays needed per sample to achieve at least three measurements per gene	6	1
Array cost per sample	$480	$450
Typical amount of total RNA needed per array	10 µg	0.1–1 µg
Total amount of RNA needed from each sample	>30 µg	<1 µg
Total arrays needed for a three-sample experiment (untreated, control-treated, experimental-treated) performed in duplicate	36	6
Total array cost	$2880	$2700

resent 40,000 human genes, and each sample is hybridized to three independent arrays in order to generate statistically significant measurements, each biological sample would require a total of six arrays, or a total cost of $480 just for the arrays. Thus, the apparent cost savings by using homemade arrays often disappears when the problems associated with such arrays are considered in the bigger picture (**Table 1**).

3.2. Affymetrix GeneChips

The most common commercial microarray platform is the GeneChip system from Affymetrix. GeneChips are made by synthesizing matched sets of short oligonucleotide pairs, one that matches perfectly (perfect match) and one with a single mismatch, on a silicon-based substrate using a photolithographic process similar to methods used in the computer chip industry. The newest GeneChips contain at least 12 pairs of probe sets for each gene; contain probe sets for more than 50,000 human, mouse, or rat genes; and generally cost academic users about $450 apiece. Having multiple probe sets for each gene ensures that even if part of the GeneChip surface becomes damaged or obscured by background, enough probe sets will still be readable to salvage the experiment. Multiple probe sets also allow statistical analyses to be performed, so both an expression level and a *p* value of expression can be reported for each gene. The Affymetrix system includes detailed protocols that rely on commercially available kits, automated fluidics stations for washing the arrays after hybridization, and an automated scanner and software package for analyzing the arrays. The complete system is expensive but produces very high-quality data and is relatively user-friendly, so it is the platform of choice for mainstream microarray facilities and for novice microarray users. The analysis kits from Affymetrix can be used with very small amounts of total RNA, even less than 20 ng, and new kits and specially designed GeneChips offer the ability to analyze samples extracted from paraffin-embedded clinical samples, making the analysis of gene expression in archived samples a possibility. The key feature of the Affymetrix system is the high-density GeneChips, which are available for several mammalian

species and several other common experimental organisms. Researchers studying gene expression in unrepresented organisms will need an alternative approach or will have to contract with Affymetrix (for a substantial fee) to produce customized GeneChips for their unique needs.

As shown in **Table 1**, for users who wish to screen more than 20,000 genes (which requires two spotted glass slide microarrays but only one Affymetrix GeneChip) and to have high-quality data (which requires at least three glass slide microarrays but only one GeneChip), experiments with Affymetrix GeneChips can be less expensive than using glass slide microarrays.

4. Types of Microarray Experiments

Most microarray experiments can be classified as one of three types. The first is the comparison of a single cell line, micro-organism, or animal strain before or after some defined treatment. The second type is a comparison of organisms (micro-organisms, cell lines, or inbred animals) that are isogenic except for one or a limited number of genetic changes, such as a single overexpressed or mutated gene. The third type is the comparison of normal or tumor tissues from multiple individuals, such as breast tumors or leukemia samples from different patients. Each of these types of experiments can be addressed with great success using microarray assays, provided that certain pitfalls can be avoided.

4.1. Treatment Comparisons

Treating a cell line or micro-organism with a specific treatment condition such as UV light or a drug that blocks a signal transduction cascade generates immediate and rapid changes in gene expression that can be detected with microarray assays. This is the simplest type of microarray experiment to analyze, because all the cells should behave similarly and relatively synchronously following the treatment. Nevertheless, there are several things to consider about such an experiment, such as the time course or duration of the treatment and the dose, etc., that can have dramatic effects. For example, the gene expression changes that occur 2 h after UV treatment of a human tissue culture cell line could be completely different from the changes observed 8 h after treatment. In addition, cells that are synchronized in the cell cycle could show significant differences compared with cells that are growing asynchronously or are density arrested. Thus, new users are encouraged to spend some time thinking about exactly what type of gene expression changes are expected, and in what type of cells those changes would be best detected.

An example from our laboratory illustrates this point. We developed recombinant adenovirus vectors expressing the c-Myb transcription factor. The c-Myb virus or a control virus was used to infect human MCF-7 mammary cells, primary lung epithelial cells, or primary lung fibroblasts. After 16 h, microarray assays were used to detect changes in gene expression. In each case, the c-Myb transcription factor caused specific changes in gene expression. However, the genes that were affected were completely different in each of the three cell types, suggesting that c-Myb transcriptional activity was strongly affected by cellular context ***(5)***. In this case, if we were trying to identify genes that were regu-

lated by c-Myb, we would have obtained completely different results in each cell type. Similarly, UV light or drug treatments could cause different gene expression changes in different cell types. Thus, it is crucial to study induced gene expression changes in the most relevant cell type available.

4.2. Analysis of Genetic Differences

A second type of experiment involves comparing otherwise isogenic organisms that differ at a single genetic locus, through either overexpression or mutation. Such experiments are especially common with yeast, cell lines, and genetically altered mice, or with cell lines derived from mouse knockout strains. These experiments differ from the ones described in **Subheading 4.1.** because the gene expression changes are at steady state. For example, researchers might use microarrays to compare the gene expression patterns in cells that differ by a mutation at a single genetic locus. However, if the cells compensate for the loss of one gene by upregulating other genes, the observed results could be quite complex. In this case, although the gene expression changes are a result of the mutation, the genes that are affected may be regulated by pathways that have nothing to do with the gene that was mutated, but are affected through secondary compensatory pathways. A better design might be to reexpress transiently the wild-type gene in the mutant cells to follow short-term changes in gene expression that are more directly affected by the gene of interest. This example points out that interpreting microarray data can be quite complicated, because gene expression pathways are influenced by so many regulatory interactions. Microarray experiments are relatively easy to perform, but poor experimental design may yield results that are difficult or impossible to interpret.

4.3. Comparison of Patient Samples

Microarray assays offer a rapid and sensitive means of comparing the gene expression profiles in tumors from different individuals and offers the promise of being used as a clinical tool to identify tumors that might respond better to a particular treatment or for identifying patients with better or worse prognoses. Such information could be extremely valuable for helping clinicians make decisions about which therapeutic options are most appropriate. Many investigators have access to dozens or even hundreds of clinical samples and see microarrays as a means of analyzing them for common patterns of gene expression. Several laboratories have been successful at identifying patterns of gene expression that correlate with clinical outcome or define classes of tumors, similar to other cytogenetic markers ***(7,11,12)***. However, these studies invariably require quite complex data and statistical analyses including methods, such as hierarchical clustering, support vector machines, and other advanced approaches ***(13,14)***. For this reason, novice users should consult with experts in complex data analysis before beginning such a study. In addition, successful clinical studies require balanced cohorts designed by qualified biostatisticians to avoid common pitfalls and artifacts (*see* **Subheading 5.5.**).

5. Planning and Experimental Design

Microarray experiments generate large and complicated data sets that pose special problems for statistical analysis and researchers trying to interpret the results. This section

discusses the most common problems faced by novice users beginning microarray experiments and approaches for eliminating them.

5.1. The Problem With Statistics

Most statistical methods depend on the comparison of replicates to estimate experimental variability in order to determine whether an observed difference is statistically significant. In general, such methods work better as the number of replicates increases. Thus, the best types of data for normal statistical analyses have relatively few variables (rows) and many replicate measurements (columns). However, microarray experiments generate data of exactly the opposite type, with many thousands of variables (genes) and, because of the high cost, very few replicates. As a consequence, the usual statistical methods have trouble dealing with microarray data. For example, it is impossible to use a *t*-test statistic on data that have fewer than three replicates. Yet, few researchers can afford to perform more than two or three replicates of microarray experiments that may cost $1000 per sample. Some specialized data analysis methods have been developed to get around the problems posed by microarray data. These methods often analyze the variation in other genes as pseudoreplicates in order to calculate the levels of variation among the genes in a data set. An example of such a method is the Cross-Gene Error Model used by the popular microarray analysis software program GeneSpring (Silicon Genetics, Redwood City, CA), which calculates a trust score for each gene based on its level of expression and the variation among other genes in the data set expressed at similar levels. These specialized data analysis methods can be quite effective and work best when the samples being compared are similar, such as from the same tissue culture cell line. The cross-gene methods have more difficulty when the samples being compared display more dramatic differences in gene expression patterns, such as when tumors from different patients are compared.

One of the problems with microarray data analysis is that the results of the experiments are generally reported only as fold change. This is necessary because different genes are expressed at widely different levels. If one tried to analyze microarrays using only raw expression-level scores, one would end up paying attention only to the genes that were expressed at high levels. However, in biological terms, the most highly expressed genes are often the least interesting, sometimes called "housekeeping" genes. The more interesting regulatory genes are often expressed at moderate or low levels. Thus, fold change measurements are necessary in order to emphasize the changes in gene expression, instead of the total abundance of individual transcripts. Unfortunately, reporting only fold change measurements introduces serious problems when discussing genes that are expressed at low levels. For example, using Affymetrix GeneChips, it is not uncommon for replicate measurements of a single gene in two identical samples to vary by as much as 1000 raw fluorescence units. An error of 1000 U is an inconsequential 5% change for a gene expressed at a level of 20,000 fluorescent units. However, for a gene expressed at approx 200 U, a 1000-U variation represents a sixfold change. Consequently, it is much more difficult to measure statistically significant changes in gene expression for genes that are expressed at low levels. In publications, the raw fluo-

rescence-level numbers for individual genes are almost never reported, making interpretation of the fold change measurements difficult. However, the relative change in total fluorescence units must be considered when determining the significance of an observed fold change.

5.2. Why Replicates Are Absolutely, Positively Required

One of the most common questions raised by new users, especially after calculating the high cost of a proposed microarray experiment, is: are replicates really necessary? After all, publications almost never show replicates of Northern or Western blots, two conventional methods of analyzing changes in gene expression. Why are replicates required for microarray experiments?

The differences are that Northern and Western blots seldom try to measure changes in gene expression that are as low as twofold and do not use statistical filters to identify gene products of interest. When microarray assays are used to measure gene expression patterns in two independent samples that should be identical, the data usually have a correlation coefficient higher than 0.97. This is a very high correlation coefficient for biological studies. However, it means that for any filter used to analyze the microarray data, up to 3% of the genes that pass through the filter will do so solely owing to apparently random fluctuation in the measurements. For an experiment measuring 40,000 genes, this noise could contribute to the improper identification of up to 1200 genes, a number far too large to be tolerated. However, if the fluctuation is random, different genes should be improperly identified in each sample. Thus, by performing duplicate analyses and requiring that genes pass through the filter in both replicates, the number of genes improperly identified should be only $0.03 \times 0.03 = 0.009$, or only 36 genes out of 40,000. Applying the filters to independent triplicate samples should eliminate all but one or two "false-positives," or improperly identified genes. For these reasons, new users should be counseled that replicate microarray assays are absolutely required. If costs are a concern, duplicate assays will generally suffice, but independent triplicate assays, if possible, are best.

5.3. Hybridization and Analysis Controls

Before starting a microarray experiment, it is important to consider the controls that should be included. Microarrays should be designed to allow the inclusion of "spiked" control mRNAs in the samples to be analyzed. These are most often a set of bacterial or artificial mRNAs generated by in vitro transcription, and mixed in predefined ratios representing low-, medium-, and high-abundance transcripts, that can be added to all the experimental samples and that hybridize to their own special spots on the array. Spiked controls are an excellent means of following the efficiency of the entire microarray analysis process, from reverse transcription through labeling to hybridization, detection, and quantitation. For Affymetrix GeneChips, premade sets of control RNAs are available as a kit. Including such controls is highly recommended because it requires very little additional effort or cost and adds significantly to the quality of the data.

5.4. The "Day Effect"

Microarrays are quite capable of detecting systematic problems in the analysis or preparation of samples. This is sometimes referred to as the "day effect," detected when techniques, such as Principle Component Analysis, are applied to large data sets containing samples that were analyzed on different days. The samples analyzed on the same day often correlate with each other better than samples analyzed on different days. The causes of the day effect are unknown but presumably have to do with batches of enzymes or reagents that differed or other systematic variations. Whatever the reason, the implication is that the samples analyzed on the same day will appear to be more similar to each other than they should, and the samples analyzed on different days will appear to be more different than they should. This has important implications for experimental design. For example, because of the day effect, it would be inappropriate to analyze all the control and untreated samples on one day, and all the treated or experimental samples on a different day. Instead, if it is impossible to analyze all the samples together, it is better to divide the samples into manageable groups, keeping both control and experimental samples in each group. For example, for a small experiment with three samples—untreated, vehicle alone, and drug-treated—it is recommended that the entire experiment be performed in duplicate, but on different days. Each set of three microarrays is analyzed together on different days, and then the data sets are compared and the analyses performed. This practice will ensure that some controls and some experimental samples are analyzed on different days, so any correlation observed between replicates will be owing to the experimental manipulations, not the systematic variations that cause the day effect.

5.5. Importance of Balanced Cohorts

A common use of microarrays is the analysis of clinical samples, with the intention of identifying patterns of gene expression that are predictive of a particular outcome. For example, researchers analyze a group of breast tumors in order to identify patterns in the microarray data that correlate with and can predict poor prognosis. However, this type of study is particularly prone to problems with experimental design related to unbalanced cohorts. In a typical study, researchers might have access to 60 tumor samples, of which 80% have good prognosis and 20% have poor prognosis. They choose two-thirds of the samples, or 40, to use as a training set and save the other 20 as the test or validation set. Microarray analysis identifies genes whose expression patterns can distinguish between the good and poor prognosis samples in the training set. When the expression patterns of those genes are analyzed in the test samples, they predict the outcome with 80% accuracy, so the experiment appears to be a success. However, if the researchers failed to balance the cohorts, they may have been misled. Because 80% of the original samples had good prognosis, a random selection of any sample would have an 80% chance of being in the good prognosis group. If the microarray analysis fails to perform better than random chance, it has not really worked. A better design would be to choose a balanced cohort, 50% with good and 50% with poor prognosis, to use as the training set. In this type of experiment, it is essential to get help from a qualified biostatistician before beginning, in order to obtain results that are valid and meaningful.

6. Basics of Microarray Data Analysis

Microarray experiments can produce complex data sets, and analyzing them can be difficult and time-consuming. The details of microarray data analysis methods are beyond the scope of this chapter. However, even if an expert performs the actual data analysis, it is important for the researchers to understand the basics of data analysis so that they can interpret the analysis summaries provided to them and ask the right questions of the expert analyst.

The analysis of microarray results has three phases. The initial analysis checks quality scores and controls in order to judge whether the labeling, hybridization, and scanning of the microarrays worked as planned and to identify problematic results that should be eliminated from the larger data set used for the final analysis. The second step is scaling and normalization, which adjusts the data obtained from individual arrays so that they can be compared. The normalization step is particularly important and dramatically affects the outcome. Choosing the correct normalization method is critical to obtaining the best results. Once the data are normalized, the third step, applying a variety of statistical tests and filters to identify genes whose expression change in the various samples is employed. There are many methods for performing this analysis, which indicates that there is no best or standard approach. Indeed, the statistical methods used for microarray data analysis are a major area of biostatistics research. For novice users, the experts in the core facility will likely choose the particular statistical methods that they are comfortable with and prefer to use, so a description of all possible methods or software packages currently in use is beyond the scope of this chapter. However, a description of some of the types of filters that can be applied to simple microarray data sets is useful to understand how the data are structured and to identify some of the pitfalls that can occur in microarray data analysis.

6.1. Initial Data Analysis

The first steps in the analysis of microarray data are to check the quality of the data obtained from each array or GeneChip; validate that all the wet-lab steps, such as reverse transcription, probe labeling, hybridization, and scanning were successful and efficient; and eliminate any data sets that are of low quality. For the novice user, these steps will usually be carried out by the core facility, which should provide the user with a report describing the overall quality of the data. The exact measurements used for judging data quality will depend on the microarray platform used and the types of controls present on the microarray. The Affymetrix GeneChip system includes a number of standard controls and quality measures that provide excellent examples of how data quality can be monitored.

6.1.1. Interpreting Affymetrix Quality Scores

Affymetrix GeneChips contain a number of control probe sets that measure the expression of housekeeping genes, such as β-actin and glyceraldehyde-3-phosphate dehydrogenase. Unlike most probe sets, which are skewed toward the 3'-end of the mRNA, in order to be less dependent on the quality of the reverse transcription reaction, the GeneChips contain several probe sets for the housekeeping controls, located at the 5'-

end, -middle, and 3'-end of the transcripts. By comparing the hybridization signals from these probe sets the researcher can get an excellent indication of the quality of the mRNA and the reverse transcription reaction used during the labeling process. For example, because the reverse transcription reaction begins at the 3'-end, if it was inefficient the probe sets from the 3'-ends of the housekeeping genes would give much stronger hybridization signals than the probe sets from the 5'-ends. In general, the ratio of the signals from the 3'-end to the 5'-end probe sets should be less than three. In addition, the housekeeping gene probe sets should have robust signals, as expected for transcripts expressed at high levels.

6.1.2. Percent Present

A second type of quality score provided by the Affymetrix system is the Percent Present statistic. Affymetrix GeneChips contain 12 or more perfect match probes and an equal number of mismatch probes for each gene. The analysis software measures the difference in the hybridization signals for the perfect match and mismatch probe pairs and then uses a statistical algorithm to determine whether the differences are significant. Based on this calculation, each gene is labeled "Present," "Marginal," or "Absent." This statistical flag is independent of the expression level and depends only on how much agreement there is among the individual probe sets for each gene. The software also calculates the fraction of genes labeled "Present" and reports this fraction as the Percent Present. In practice, the Percent Present can vary significantly, depending on the type of sample (e.g., primary cells vs transformed cell lines) being analyzed. However, within one experiment analyzing similar samples, all of the GeneChips should give a similar Percent Present. An abnormally low Percent Present is an indicator that an RNA sample was of poor quality or that the labeling or hybridization reactions were flawed.

6.1.3. Interpretation of Scaling Factors

Affymetrix also permits data from individual GeneChips to be scaled, which is similar to per-chip normalization (*see* **Subheading 6.2.1.**). Although scaling is not absolutely necessary, it does provide an additional quality statistic, the scaling factor. Scaling works by multiplying all the gene expression values by some constant, the scaling factor, which adjusts the average expression to some preset number, usually 500 or a similar integer. If scaling is used, the scaling factor provides an excellent quality measure. Poor-quality data sets invariably have larger scaling factors, because the labeling or hybridization was affected for all the genes represented on the GeneChip. Ideally, all the samples being analyzed as a group should have similar scaling factors. If Affymetrix scaling is used, there is no need to use additional per-chip normalization, discussed in **Subheading 6.2.1.**

6.2. Scaling and Normalization

Proper scaling and normalization of microarray data is extremely important and dramatically affects the results of the analysis. The two basic types of normalization are scaling, or per-chip normalization, which adjusts the average intensity of an entire micro-

array sample, and per-gene normalization, which is used to compare the relative expression of a single gene within a group of samples.

6.2.1. Per-Chip Normalization

Scaling, or per-chip normalization, is a means of adjusting the overall fluorescence of each microarray to the same average intensity, analogous to adjusting the sensitivity of the scanner so that each sample has the same overall brightness. This type of adjustment makes sense for samples that are similar, and that are expected to have similar numbers of genes expressed, mostly at similar levels. However, it may not make sense for samples that are dramatically different, such as a comparison of resting cells vs proliferating cells, because the latter may have many more genes expressed. By default, most samples are subjected to scaling or per-chip normalization. However, the details of the experiment should be considered carefully to determine whether per-chip normalization is appropriate. In particular, if samples have dramatically different levels of expression of the housekeeping genes, which contribute greatly to the average fluorescence, it might be better not to subject the samples to per-chip normalization.

6.2.2. Per-Gene Normalization

The absolute level of expression among different genes varies dramatically, from thousands to less than one transcript per cell. As a result, it is difficult to compare changes in the level of expression of specific genes among samples. As discussed in **Subheading 5.1.**, a 1000-fluorescent unit change in expression of a high-abundance transcript may represent a small change, as little as 5%, but could represent a manyfold change in expression of a gene that is expressed at low levels. Per-gene normalization is used to overcome this problem by comparing the relative expression of each gene across the various samples in an experiment, expressed as fold change. As a consequence, genes that display similar patterns of up or down changes in expression across samples can be identified despite the absolute differences in their expression levels.

The big problem with per-gene normalization is deciding what to normalize each sample to. By default, most microarray data analysis programs calculate the mean expression level for each gene, then normalize each sample against that mean, or control value. This approach works but can result in some strange results. Take the example described in **Subheading 5.4.** of a small microarray experiment containing just three conditions—untreated, vehicle treated, and drug treated—performed in duplicate. The entire experiment would consist of six microarrays, two independent measurements for each condition (**Fig. 2A**). Now, consider a gene expressed at or near zero in the untreated and vehicle-treated conditions. The software never reports an expression value of zero, so assume that the average value in the untreated and vehicle-treated samples is a low number, e.g., 200. If this gene is strongly induced by the drug treatment its expression level could go up to an average of 2000 U. Using the default per-gene normalization described above, the mean intensity across all six samples would be 800. The fold change reported for the untreated and vehicle-treated samples would be 0.25 and the fold change for the drug-treated samples would be 2.5. This gene would just barely pass

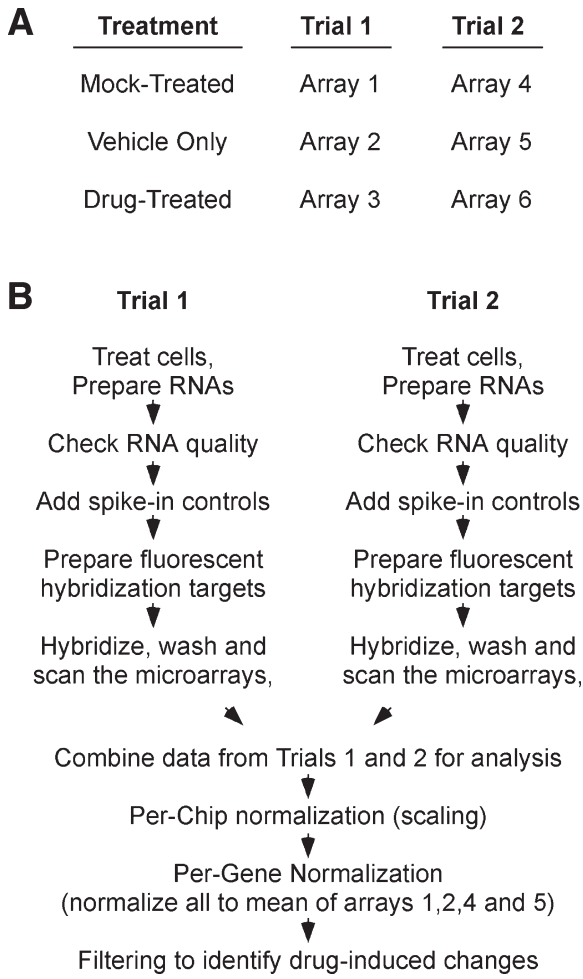

Fig. 2. Design and analysis strategy for a simple microarray experiment. (**A**) Simple microarray experiment design. A six-microarray experiment is designed to test the effect of drug treatment on tissue culture cells. Duplicate samples of the drug-treated cells will be compared with duplicates of vehicle-treated or mock-treated samples. Each sample will be analyzed with its own microarray, making a total of six assays. (**B**) Flow chart of simple microarray experiment. RNA samples from the two trials are collected and analyzed separately, and then the data are combined for the analysis. Keeping the samples separate helps to avoid day effects and other systematic problems. The data are normalized to the mean of the four control samples (mock treated or vehicle treated) to identify drug-induced changes in gene expression in the two treated samples.

a filter designed to find genes induced more than 2.5-fold by the drug. However, comparison of the raw scores shows that the average expression actually changed from 200 to 2000, which is a 10-fold change! In this case, the default normalization scheme

was inappropriate. The data should have been normalized to the untreated or the vehicle-treated samples, rather than to the average of all the samples. As the example illustrates, choosing the correct normalization scheme is extremely important and affects the results and the genes that will be identified by the analysis.

In general, if the experiment has true control samples, such as the untreated and vehicle-treated samples in the example described in **Fig. 2A**, per-gene normalization should use those samples as the controls. The result will be fold-change data that reflect the change relative to the controls, a much more logical type of result than a fold change relative to the mean of all the samples. On the other hand, when no true controls are available, such as when comparing the gene expression profiles of a number of tumor samples from different patients, normalization to the mean of all the samples may be the only available choice. In either case, it is important for the researcher to understand how the data were normalized in order to interpret the fold-change results.

6.3. The Simplest Analysis: Filtering to Identify Regulated Genes

After normalization, a variety of techniques can be used to identify genes with altered expression in one or more of the experimental conditions. This section focuses on filtering, the simplest method to identify interesting genes and one of the most useful for novice microarray users. Filtering is direct and related to the experimental design, so it is relatively easy to set up and understand. However, filtering is best used for addressing specific biological questions in relatively simple experiments. The filtering approach rapidly becomes cumbersome as the experimental design becomes more complicated and is not suitable for experiments with more than three or four types of experimental conditions. Nevertheless, a basic discussion of data analysis using filtering can point out the strengths and weaknesses in microarray data analysis and prepare users for adopting more advanced techniques, if they are necessary.

6.3.1. The Analysis Strategy

To illustrate the concepts and pitfalls of data analysis by filtering, consider the example experiment described in **Fig. 2A**, with two biological replicates each for untreated, vehicle-treated, and drug-treated samples, or a total of six microarrays. This experiment has a simple experimental design (**Fig. 2A**). Nevertheless, it is important to predict what types of results are expected in order to design the appropriate filters.

6.3.2. Filter on Flags

The first criterion is that only those genes that can actually be detected above background levels should be considered for further analysis. If a gene is expressed at such low levels that it cannot be distinguished from background in any of the samples, there is no sense in applying a filter to see whether its expression has increased. This may seem obvious but it is actually a major concern in microarray experiments that utilize normalized data, because once the data are normalized, all the information about absolute expression levels are lost. Thus, it is a common error to identify genes that are up- or downregulated based on fold change without paying attention to whether the genes are actually expressed at a level that is significant and above background. For several reasons, this problem is a special concern for users of glass spotted arrays. First, absolute

background levels are difficult to measure using glass spotted arrays. This is because the background hybridization in the areas spotted with DNA can be much greater than in the areas without DNA, because background levels can vary from probe to probe, depending on the G-C content, and because such arrays often suffer from high background and "smearing," all of which complicate background measurements. Second, glass spotted arrays often have only one spot per gene, so there is no way to do statistical calculations to determine whether an expression measurement is significantly different from background. Finally, signals using glass spotted arrays are often weak, so most spots are detected in the near-background range. Sometimes it is possible to increase the sensitivity of the scanner to alleviate this problem, but detection of low- and even medium-abundance mRNAs can nevertheless be quite difficult.

The Affymetrix GeneChip system has incorporated a number of measures to enable more accurate background detection and to permit statistical measures to be applied to determine whether each gene is expressed above background. On the Affymetrix arrays, at least 12 perfect match probes and an equal number of corresponding single nucleotide mismatch probes represent each gene. By comparing the hybridization signals for the perfect and mismatch probes, which in each case differ by only one nucleotide, a fairly accurate estimate of the difference between specific and nonspecific signals can be determined. The 12 or more independent measurements allow statistical tests to be made, and the size of the corresponding *p* value is used to calculate a Present/Marginal/Absent call. This "flag," or qualitative measure that accompanies the raw expression score, is a measure of whether the genes are statistically different from background. The flag allows the data to be filtered to exclude genes that cannot be accurately measured. In general, it is advisable to filter Affymetrix data to exclude genes that are flagged "Absent" in all the samples, which is often one-third or more of the genes on the array. It is also possible to be more selective. In our example experiment (**Fig. 2A**), if one was interested only in genes that were "off" in the controls and "on" in the drug-treated samples, one could filter for genes that were flagged "Absent" in the untreated and vehicle-treated samples and also flagged "Present" in the drug-treated samples. However, such a specific use of flags is generally unwarranted because it could be too selective.

6.3.3. Filter on Fold Change

The most basic type of filtering is the comparison of fold change. Microarray data are generally filtered to identify genes that are at least twofold different in the experimental conditions. In our example experiment, one would try to identify genes that were at least twofold up- or downregulated in the drug-treated samples compared with both the vehicle and the untreated samples.

The best approach is to combine filters to achieve the most specific result possible. For example, to identify genes that are upregulated by the drug treatment, a filter should be designed to find only genes that are flagged "Present" and also twofold or more upregulated in both of the drug-treated samples, because it makes no sense to study genes that are apparently upregulated but cannot be detected in a statistically significant manner. Whether the genes are flagged "Present" or "Absent" in the controls is irrelevant.

For downregulated genes, the measurements must be flagged "Present" in all the controls but expressed at 0.5-fold or less in both of the drug-treated samples. It is not logical to find genes that appear to be downregulated unless they were actually expressed above background in the controls.

6.3.4. Other Filters

Numerous additional filters can be applied to microarray data. To be most conservative, some users will wish to limit their analysis only to genes that are most robustly expressed, and that can be most easily detected by other methods, such as Northern blots. For that purpose, it may be useful to filter on the raw expression level, essentially setting a cutoff for minimum expression above which a gene must be expressed to be considered further. The cutoff is somewhat arbitrary and depends on the data set and the settings used for the scanners and for the normalization. Nevertheless, this approach can help identify the genes that will be simplest to study in subsequent validation experiments, at the expense of eliminating some of the most interesting genes that are expressed at lower levels, closer to the background level.

6.4. More Advanced Analysis: Clustering

Microarray data can be extremely complex, and many methods of data analysis are available. In fact, the development of new and improved methods for analyzing microarray data is a major area of research among bioinformatics specialists. The most common of these methods involves various supervised and unsupervised clustering methods that have been developed primarily for the analysis of large data sets, especially those that compare numerous samples from different individuals, such as a series of tumor vs normal samples. These methods are generally not too useful for novice microarray users performing simple experiments; their description is beyond the scope of this article, but they are discussed in **Chapter 4**. Nearly all the advanced methods use statistical tests to group genes or patients in clusters, based on their expression profiles, and do better with larger numbers of samples. However, as a general rule, it is best to filter the data first in order to limit the analysis to the smallest possible set of genes that are informative. It makes little sense to include thousands of genes that cannot be detected above background in the data set being subjected to statistical clustering. Once the data are limited to the genes that are truly flagged "Present" and that change twofold or more in the experimental samples, clustering methods may be able to divide the genes into interesting groups, especially if the experiment includes several different types of samples, such as treatments with different drugs or a time course of drug treatments.

7. Conclusion

Microarray technologies have empowered novice users with the ability to assay changes in gene expression at the whole-genome level. There is little doubt that microarray results will lead to new and entirely unexpected results, and pursuing such experiments will be worthwhile for many investigators. However, there are several concerns that should be heeded. Microarray experiments are expensive and they can be quite labor-intensive. In addition, the data that they produce are quite complex. Novice users

should seek out advice from their core facilities or collaborators to make sure that they have designed the most efficient experiment that is compatible with microarray assays. A poorly designed experiment is the most common reason that microarray experiments fail to yield results that are interpretable. In most cases, clear thinking and a discussion with an experienced microarray user, a core facility leader, or a biostatistician will lead to much better experimental design and much better data.

Acknowledgments

I thank Dr. G. G. Pickett for helpful comments on the manuscript. I am supported by grants from the USPHS/National Cancer Institute (#RO1 CA58443) and by the University of New Mexico Health Sciences Center. I am codirector of the Keck-UNM Genomics Resource, a microarray and gene expression analysis facility supported by a grant from the W. M. Keck Foundation as well as the State of New Mexico and the UNM Cancer Research and Treatment Center.

References

1. Ishida, S., Huang, E., Zuzan, H., et al. (2001) Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol. Cell Biol.* **21,** 4684–4699.
2. Spellman, P. T., Sherlock, G., Zhang, M. Q., et al. (1998) Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297.
3. Frueh, F. W., Hayashibara, K. C., Brown, P. O., and Whitlock, J. P. Jr. (2001) Use of cDNA microarrays to analyze dioxin-induced changes in human liver gene expression. *Toxicol. Lett.* **122,** 189–203.
4. Liang, G., Gonzales, F. A., Jones, P. A., Orntoft, T. F., and Thykjaer, T. (2002) Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. *Cancer Res.* **62,** 961–966.
5. Lei, W., Rushton, J. J., Davis, L. M., Liu, F., and Ness, S. A. (2004) Positive and negative determinants of target gene specificity in Myb transcription factors. *J. Biol. Chem.* **279,** 29,519–29,527.
6. Rushton, J. J., Davis, L. M., Lei, W., Mo, X., Leutz, A., and Ness, S. A. (2003) Distinct changes in gene expression induced by A-Myb, B-Myb and c-Myb proteins. *Oncogene* **22,** 308–313.
7. Ferrando, A. A., Neuberg, D. S., Staunton, J., et al. (2002) Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1,** 75–87.
8. Monks, A., Harris, E., Hose, C., Connelly, J., and Sausville, E. A. (2003) Genotoxic profiling of MCF-7 breast cancer cell line elucidates gene expression modifications underlying toxicity of the anticancer drug 2-(4-amino-3-methylphenyl)-5-fluorobenzothiazole. *Mol. Pharmacol.* **63,** 766–772.
9. Karyala, S., Guo, J., Sartor, M., et al. (2004) Different global gene expression profiles in benzo[a]pyrene- and dioxin-treated vascular smooth muscle cells of AHR-knockout and wild-type mice. *Cardiovasc. Toxicol.* **4,** 47–74.
10. Verheyen, G. R., Nuijten, J. M., Van Hummelen, P., and Schoeters, G. R. (2004) Microarray analysis of the effect of diesel exhaust particles on in vitro cultured macrophages. *Toxicol. In Vitro* **18,** 377–391.

11. Sotiriou, C., Neo, S. Y., McShane, L. M., et al. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA* **100,** 10,393–10,398.
12. Valk, P. J., Verhaak, R. G., Beijen, M. A., et al. (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med.* **350,** 1617–1628.
13. Benito, M., Parker, J., Du, Q., et al. (2004) Adjustment of systematic microarray data biases. *Bioinformatics* **20,** 105–114.
14. Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003) Regression approaches for microarray data analysis. *J. Comput. Biol.* **10,** 961–980.