Since its development almost 2 decades ago, gene expression microarray technology has generated excitement and raised expectations for dramatic discoveries of underlying disease mechanisms. The general concept of a gene expression microarray is relatively straightforward. Messenger RNA is extracted from a biospecimen. Examples of biospecimens include cells grown in culture, tissue from animal models, and human tissue. The messenger RNAs are converted into DNA, labeled with a fluorescent dye, and hybridized to a slide containing probes corresponding to target genes of interest arranged in a matrix format. The level of gene expression is estimated from the raw intensities of the fluorescence emitted by the labeled sequence bound to probes representing genes from which messenger RNAs were transcribed. Various different platforms are available to measure gene expression levels. Some are single-channel arrays in which 1 sample is hybridized on each array; others are 2-channel arrays in which 2 samples are hybridized simultaneously on a single array with 2 different fluorescent dyes. Some platforms are all-purpose standard arrays manufactured on a large scale that typically interrogate the expression level of all known genes. There also are custom arrays available that measure the expression levels of genes known to be associated with a particular disease or genes of interest to an individual researcher. Throughout this article, the term microarray is used to refer to a gene expression microarray.

The intent of the present article is to present an overview of the statistical components related to the design and analysis of a microarray experiment. In general, the usual statistical standards of scientific research apply to microarray studies. This article reviews common study objectives and designs, describes the typical analytical techniques, and concludes with a discussion of threats to the validity of a microarray study. Because articles in this statistics series have word-count restrictions that necessarily limit their breadth and depth, readers wanting more detail are directed to more comprehensive reviews by Allison et al, Dobbin and Simon, and Dupuy and Simon.

Study Objectives
Effective use of microarray experiments requires clear objectives and effective study design. Clearly stated study objectives are needed to determine the appropriate type of specimens, an adequate sample size, and suitable analysis plans. Medical research studies using microarrays encompass a wide variety of disease-related objectives that generally fall into 1 of 3 broad categories: class comparison, class prediction, and class discovery.

Class Comparison
Class comparison is also commonly known as differential analysis or analysis of differential expression. It involves the comparison of gene expression profiles of samples from distinct, predefined groups to identify the genes that are differentially expressed among the groups. Examples include comparing treated and untreated cell lines to ascertain the effect of a new drug on gene expression levels or a comparison of healthy tissue with diseased tissue to identify genes with altered expression. Such studies yield lists of genes that are significantly altered among the groups. The purpose of the lists is to provide insight into the underlying biological mechanisms and perhaps to identify potential therapeutic targets.

Class Prediction
A class prediction study develops a classifier that uses the gene expression level of multiple genes that can be applied to the expression profile of a newly acquired biospecimen to predict its (unknown) class. As in class comparison studies, the focus is to identify genes that differ across predefined classes. However, in class prediction studies, the gene expression values are explanatory variables rather than outcome variables. Furthermore, the goal of the analysis of class prediction studies is to identify a small set of genes able to accurately discern among the distinct classes; it is not to identify all genes that differ. There are many different possibilities for classes of interest. Class prediction studies could be yield classifiers that are (1) diagnostic, eg, a classifier that distinguishes between 2 different disease states; (2) prognostic, eg, a classifier that distinguishes short-term survivors from long-term survivors; or (3) predictive, eg, a classifier that predicts whether a patient will respond to a particular drug.

Class Discovery
Class discovery differs from class comparison and class prediction studies in that the classes are not predefined.
Typically, the objective of these studies is to determine whether subsets of samples with seemingly homogeneous phenotypes can be discerned on the basis of differences in their gene expression profiles. For example, there are many diseases for which individuals with seemingly similar phenotypes have a large variability in outcomes such as survival. The hypothesis is that differences at a molecular level are the cause of the observed variability in outcome. Class discovery studies also may uncover biological features of the disease that provide insight into the disease pathogenesis; in particular, identifying molecular differences that define disease subgroups may lead to improved therapeutic agents.

**Study Design**

There have been numerous claims of gene expression profiles that can accurately diagnose disease or make an accurate prognosis, and some prominent results have not been reproduced. There are 3 main reasons for this: bias, overfitting, and generalizability. Bias poses a substantial threat to the validity of a gene expression microarray study and is the most difficult to avoid. Study bias is addressed primarily through proper study design. Another important aspect of study design is determining an adequate sample size to address the question of interest.

**Study Design Issues**

Good study design minimizes potential bias, defined here as a systematic erroneous association of a variable with a group in a way that distorts the comparison between groups. As an example of bias, suppose, in a study designed to assess the ability of a multigene expression test to discern between groups of patients with and without left ventricular dysfunction, that the samples from individuals in the left ventricular dysfunction group have been stored an average of 7 years and those in the comparison group have an average storage time of 2 years. If the test is affected by tissue storage time, the bias could account for the ability of the test to distinguish between the groups.

To avoid bias, study groups should be created that differ only with respect to the variable of interest; they should be “equal” with respect to all other variables. An experimental method such as a randomized controlled trial provides the most effective way to create equal groups and to minimize bias. Often, the experimental method cannot be used to address research questions of interest in microarray studies such as diagnosis and prognosis. However, the principle of experimental method can be applied to these observational studies, which can minimize some common biases that arise in microarray studies. Note that the issue of bias is a relatively new challenge for laboratory researchers who are used to tightly controlled conditions of experimental research. Designs for microarray studies should strive to create homogeneity across groups on all variables but the one of interest. In particular, consideration should be given to the equality of characteristics of individuals, specimen collection, handling, and storage. A mechanism to facilitate obtaining subject (biospecimen) homogeneity is the use of inclusion/exclusion criteria. Likewise, there should be similar protocols used for the collection, storage, and handling of specimens for all groups. Applying similar protocols across groups poses a challenge because disease tissues often undergo different collection, handling, and storage procedures compared with normal tissues.

**Sample Size**

Microarray studies tend to be underpowered, meaning they have too few samples for detecting features of interest. The reason cited is often the large cost of the arrays. However, from a scientific perspective, it is essential to have adequate sample sizes. One simple, common approach in the setting of class comparison is to perform power and sample size calculations based on a 2-sample *t* test. Appropriately large samples are required to avoid missing important expression differentials. Other approaches to sample size estimation would be used for prediction or discovery settings.

The approach to sample size planning for a microarray class comparison study is the same as for any other medical research study and is based on 4 parameters: α, the level of significance or probability of a type I error; 1−*β*, the statistical power (β is the probability of a type II error); δ, the minimum detectable difference in gene expression levels (usually on a log base 2 scale); and σ, the SD of a gene expression level (usually on a log base 2 scale). The investigator sets the values of the first 3 parameters. Recommended values for α and β for gene expression microarray studies are α=0.001 and β=0.05. The small α value is to protect against multiple comparisons (more below); the small β value is to provide good statistical power for identifying genes that really are differentially expressed. The value of δ differs for each gene, and investigators often set δ=1, which corresponds to a 2-fold difference on a log base 2 scale. Finally, estimates are needed for the variation of expression values among samples. This variability depends on the heterogeneity of the biospecimens, with cell line samples being the least heterogeneous, human samples being the most heterogeneous, and animal model samples falling somewhere in between. Cell line samples are derived from multiple cultures of the same cell line, 1 per culture. Furthermore, the amount of variability depends on a particular gene.

Simon et al report that the average within-class expression level variability (log 2 scale) ranges from 0.10 for cell lines to 0.25 for animal model tissue to 0.5 to 0.75 for human tissues. The Table gives general guidelines for minimum sample sizes according to the numbers above and using a 2-sided *t* test. As microarray data become more abundant, estimates of σ
should be based on previous experiments using similar biospecimens from similar populations.

Objectives of microarray experiments also include class prediction and class discovery. At the present time, there are no generally accepted methods for determining sample sizes for class prediction models; however, methods do exist in the literature such as that of Dobbin and Simon.10 The development of sample size determination for class discovery is currently an active area of research. It is generally acknowledged that these types of studies require larger sample sizes than a class comparison or class prediction study.

Analysis
Before analysis, microarray data often are transformed to the log 2 base scale. There are 2 reasons. First, gene expression levels span orders of magnitude, and using a log base 2 scale reduces the magnitude of the range of the data and often makes it more normally distributed. Second, biologists are comfortable interpreting changes on a log 2 scale in which a 1-unit change is equivalent to a doubling in the original scale. There are typically 2 steps in the analysis of microarray data. The first involves low-level preprocessing of the data; the second is a higher-order analysis that depends on the nature of the research question.

Low-Level Analysis: Data Preprocessing
Data preprocessing (low-level analysis) consists of background subtraction and normalization. The purpose of background subtraction is to make the foreground intensity more properly proportional to the abundance of bound material (ie, cDNA). There is a tradeoff between performing a background subtraction or not in terms of variance and bias. If no background subtraction is done, the variance in expression levels is the smallest, but the bias in terms of the measured intensity reflecting the amount of bound material is greatest. Performing background subtraction reduces the bias but increases the amount of variability in gene expression level. Some investigators choose not to perform any background correction because currently little is known of the biological effects of a given level of gene expression. Hence, having bias in the estimate in the absolute level of gene expression is of little consequence, and the gain is less variability, which translates to greater power. Global background correction techniques11,12 use the same constant to represent the background for all spots on the array. It is easy to perform but does not take into account background variation that exists across the array. Regional background correction13–15 provides greater flexibility than global correction in terms of accounting for variation in background across the array.

In addition to the biological variability among samples, other sources of variability or noise are introduced through the processes of biospecimen handling, RNA extraction, fluorescent labeling, hybridization to the array, and laser scanning to measure the intensities. Before the gene expression values between arrays are compared, the arrays typically are normalized to remove unwanted/systematic variability in expression levels for a gene across arrays induced by the process of RNA extraction, hybridization, and scanning while preserving the true biological variability. Many normalization procedures are based on the assumption that the expression levels of most genes do not change among the groups being compared. This is a reasonable assumption for most microarray experiments that use general-purpose arrays designed to interrogate all genes. There are 2 general types of normalization techniques: linear normalization16 and nonlinear normalization.17–19 Linear normalization uses the same normalization factor for all genes on the array. Nonlinear normalization determines the normalization factor for a spot on the basis of its level of expression. It is generally agreed that nonlinear normalization techniques are superior to linear normalization techniques.

Class Comparison
The goal of a class comparison study is to identify genes that are differentially expressed between prespecified classes. Tests used to identify genes as differentially expressed between groups need to account for variability. In early microarray studies, investigators selected genes with mean values that differed between groups by a predefined magnitude such as a ±2-fold difference. This completely ignored the amount of variability in gene expression levels, greatly increasing the probability of declaring a gene differentially expressed when it truly is not.20 Methods that account for the magnitude of the difference in mean expression between groups and level of variability of gene expression within a group include the t test, permutation test, and Wilcoxon rank-sum test. The last 2 are used when there is concern that the gene expression levels are not normally distributed. If there are >2 groups, there are natural extensions of these tests: ANOVA, permutation F tests, and the Kruskal-Wallis test, respectively.

Class Prediction
There are numerous methods for class comparison analyses, and the number keeps increasing. Methods include logistic regression,21 linear and quadratic discriminant analysis,22 nearest neighbor classifiers,23 decision trees,24 shrunken centroids,25 neural networks,26 random forests,27 support vector machines,28 and many more. It has been shown that simpler methods tend to work better,29 likely because many of the more sophisticated methods that uncover even slight structure in the data do not work well when the number of cases is small compared with the number of variables because they then tend to model the noise in the data in addition to, or rather than, the true structure.

Class Discovery
The idea of class discovery is to discover patterns in microarray data. There are 2 types of problems in this category: identifying groups of coexpressed genes and finding patterns in expression profiles of different specimens that do not have predefined profiles from among a set of samples with similar phenotypes. The techniques used for class discovery are called unsupervised methods. A primary tool used for class discovery is clustering algorithms; their goal is to form subgroups so that the objects (specimens or genes) within a subgroup are more similar to one another than objects in different groups. There are 2 main types of clustering algo-
Hierarchical algorithms derive a nested series of partitions of data points either by merging 2 clusters at each step (agglomerative hierarchical clustering) or by splitting a cluster into 2 clusters at each step (divisive hierarchical clustering). Partitional methods aim to produce a single partition of the items; there is no nesting. K-means clustering and self-organizing maps are both partitional methods.

The results of clustering algorithms are typically displayed graphically. In the case of hierarchical clustering, a tree structure or dendrogram is generated (see the Figure). Another popular display for hierarchical clustering results in a heat map (also called a color image plot). A heat map is a rectangular array of colored blocks with the block color representing the expression level of 1 gene on 1 array (specimen). Each column of boxes represents a specimen, and each row of boxes represents a gene. The columns are ordered according to how they are ordered in a hierarchical clustering dendogram. Typically, the rows also are ordered according to the dendogram that resulted from a cluster analysis of the genes (see the Figure). The results are patches of 2 different colors (typically red and green) indicating combinations of genes and specimens that exhibit high or low expression.

**Discussion**

Because of the nature of microarray data (orders-of-magnitude more variables than samples), chance often threatens the validity of the results by leading to erroneous conclusions. In the case of class comparisons, tens of thousands of hypothesis tests are performed simultaneously, 1 for each gene. Suppose that no genes truly are differentially expressed between 2 groups. If a microarray measures 30 000 genes and a significance level of 0.05 is used, 1500 genes would be falsely declared as being differentially expressed as a result of chance alone. Several approaches are used to control the number of false positives in a study. The first method is to choose a smaller level of significance, typically 0.001. In the example, this would reduce the number of false positives from 1500 to 30. Other methods used to address the multiple comparison problems are the use of Bonferroni-adjusted probability values and multivariate permutations methods. Another approach is to control the false discovery proportion, ie, of those genes that are identified as being differentially expressed, control the proportion of false positives. This average false discovery proportion is sometimes called the false discovery rate.

A problem caused by chance that is often less familiar to investigators performing gene expression microarray studies is overfitting. Overfitting occurs when a multivariable model designed to generate a classifier that discriminates between groups of patients is made to fit the data perfectly. A problem occurs when a pattern is found that fits perfectly by chance. Given that thousands of genes are typically considered when developing a multigene classifier, there is a large
likelihood that this will occur. Overfitting is not inherent in microarray studies; it can occur whenever a multivariable analysis is done to assess associations between a large number of explanatory variables and an outcome with a relatively small sample size (especially when the number of explanatory variables exceeds the sample size). Overfitting can easily be checked for by assessing reproducibility in a completely independent group of individuals. Overfitting remains a problem in microarray studies mainly because of failure to adequately carry out such checking.

Conclusions

The design and analysis of microarray studies should adhere to proven principles for all clinical research studies. Problems caused by chance and bias must be minimized at every step of a study. Bias presents the greatest difficulty; many biases are subtle. Unfortunately, in current microarray research, problems related to bias are widely ignored by investigators, reviewers, and editors, often leading to erroneous conclusions. Chance also threatens the validity of many microarray studies related to bias are widely ignored by investigators, reviewers, and editors, often leading to erroneous conclusions. Chance also threatens the validity of many microarray results. To lessen this threat, the possibility of overfitting should be explicitly assessed. Specifically, in microarray studies, claims of discrimination of classifiers should prompt the question of whether reproducibility has been assessed in an independent validation set of adequate size. If it has not, then the results should be viewed as inconclusive.

Disclosures

None.

References


Key Words: gene expression microarrays ■ genes ■ statistics
Genetics and Genomics: Gene Expression Microarrays
Karla V. Ballman

Circulation. 2008;118:1593-1597
doi: 10.1161/CIRCULATIONAHA.107.714600
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/118/15/1593

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/