Genetic Epidemiology 23:21–36 (2002)

Design of Studies Using DNA Microarrays

Richard Simon,* Michael D. Radmacher, and Kevin Dobbin

Biometric Research Branch, National Cancer Institute, Bethesda, Maryland

DNA microarrays are assays that simultaneously provide information about expression levels of thousands of genes and are consequently finding wide use in biomedical research. In order to control the many sources of variation and the many opportunities for misanalysis, DNA microarray studies require careful planning. Different studies have different objectives, and important aspects of design and analysis strategy differ for different types of studies. We review several types of objectives of studies using DNA microarrays and address issues such as selection of samples, levels of replication needed, allocation of samples to dyes and arrays, sample size considerations, and analysis strategies. Genet. Epidemiol. 23:21–36, 2002. © 2002 Wiley-Liss, Inc.

Key words: gene expression; bioinformatics; biomarkers; computational biology

INTRODUCTION

The DNA microarray is an assay that can be used to measure the level of expression in a collection of cells for thousands of genes. Nearly all the cells of an organism carry the same genome. The phenotypic differences among cells of different types are determined by differences in the level of expression of the genes. Consequently, an assay that can measure the level of expression of thousands of genes simultaneously provides genome wide insight into the workings of cells. As a result, DNA microarrays are finding use in a wide variety of studies. The experimental procedures used in DNA microarrays are described by Shalon et al. [1996] and by Lockhart et al. [1996].

The abundance of data resulting from a single microarray assay has sometimes fostered a distorted view that microarray data can be collected in a relatively unplanned manner. The naive expectation is that the sheer volume of data generated will suffice for algorithmically determining important and unanticipated patterns in

*Correspondence to: Richard Simon, Biometric Research Branch, National Cancer Institute, MSC 7434, 9000 Rockville Pike, Bethesda, MD 20892-7434. E-mail: rsimon@nih.gov
Received for publication 9 January 2002; Revision accepted 8 March 2002
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.0202

the data. This is not a good plan for using microarray technology; microarray studies require careful planning and development of analysis strategies.

There is a tension in biology research between "hypothesis driven" research and "descriptive" research. Descriptive research is often suspect because it is not designed to answer specific questions and therefore the usefulness of the data collected may be questioned. DNA microarrays have been seen as a tool for descriptive research because they provide a survey of gene expression rather than a focus on mechanistic aspects of the workings of a small number of genes. Microarrays generally are not best suited for testing gene-specific mechanistic hypotheses because other more sensitive assays are available for measuring expression of a specific gene. Nevertheless, most effective microarray-based studies have a clear objective and answer well-defined questions, although generally not gene-specific mechanistic questions. Clear identification of the objective of a microarray study is important for designing the study and constructing an appropriate analysis strategy.

In the next section, we will describe some of the objectives that are being addressed using DNA microarrays. The section entitled Sources of Variation and Levels of Replication identifies various important sources of variability and relates these sources to levels of replication in microarray studies. The section Sample Selection and Experimental Design discusses specimen selection and, for cDNA array studies, designs for allocating specimens to arrays. The section entitled Comparing Expression Profiles Among Predefined Classes addresses design issues, including sample size determination, for studies attempting to compare expression profiles and identify differentially expressed genes in comparing predefined classes of specimens. The section entitled Developing Prognostic Models addresses design issues for studies attempting to identify gene-expression-based predictors of a time-to-event variable. The section on Class Discovery addresses design issues for studies attempting to discover whether specimens of a disease are homogeneous with regard to expression profiles or whether there are subsets of the disease with characteristic expression profiles.

OBJECTIVES OF DNA MICROARRAY STUDIES

Because DNA microarrays are useful for such a wide variety of experimental studies, it is not possible for us to be comprehensive in discussing study objectives. It is useful, however, to identify a few generic types of objectives that are often seen in microarray research.

Class Comparison

One type of study involves comparison of expression profiles obtained from different predefined classes of specimens. For example, Hedenfalk et al. [2001] compared expression profiles in breast cancer specimens containing BRCA1 mutations, breast cancer specimens containing BRCA2 mutations, and breast cancer specimens from spontaneous tumors with neither type of mutation. Golub compared expression profiles between specimens of acute myelogenous leukemia and specimens of acute lymphocytic leukemia [Golub et al., 1999]. Ross et al. [2000]

compared expression profiles of cancer cell lines of different tissues of origin. Spellman et al. [1998] compared expression profiles of yeast cells synchronized after blockage of the cell cycle. They collected specimens at various times after removing the cell cycle block. In pharmaceutical research there is interest in characterizing expression profiles of each major human tissue type in order to identify tissue-specific therapeutic targets. This involves comparing expression profiles from specimens of different tissue types. One might also be interested in comparing expression profiles of a given tissue that has not been exposed to a specified drug with those profiles of tissues of the same type after exposure to the specified drug. This type of study enables the effects of a drug on tissues to be studied to identify potential toxicities.

Three types of objectives are often of interest in studies comparing expression profiles for predefined classes. The first is to determine whether the expression profiles differ among the classes. The second objective is to identify which genes are differentially expressed among the classes and to identify the patterns of differential expression across the classes. A third objective is sometimes of interest; developing a multivariate predictor of class membership based on the level of expression of selected genes. When class prediction is of interest, it is important to provide an unbiased estimate of the misclassification rate and to establish that this misclassification rate is lower than expected when there is no relation between expression profile and class.

Prognostic Prediction

Some microarray studies are designed to determine whether there is a relationship between expression profile and clinical outcome and to develop a prognostic predictor of outcome based on the level of expression of selected genes. When outcome is binary, then this can be considered a class prediction problem as discussed above. For example, some pharmacogenomic studies attempt to predict which patients will experience toxicity to an effective treatment for a disease. In many cases, however, outcome will not be categorical. One example is the development of a prognostic model for predicting duration of survival for patients with large cell lymphoma [Lymphoma/Leukemia Project, in press]. In that study, the data were continuous and right censored. In addition to thousands of gene expression candidate predictors, clinical predictors were also available for consideration. Some studies being developed attempt to predict which patients will develop cancer in a specified organ based on expression profiles of biopsies of that organ at baseline.

Class Discovery

Another type of microarray study involves the identification of novel subtypes of specimens within a population. This objective is based on the idea that important biological differences among specimens that are clinically and morphologically similar may be discernible at the molecular level. Cancers are usually classified based on the organ in which the tumor originates. Subclassification is based on the cell type of the cell in which the tumor arose if that can be discerned. Often the cell of origin cannot be determined based on standard morphological and histological criteria.

Many microarray studies in cancer have the objective of developing a taxonomy of cancers that originate in a given organ site in order to identify subclasses of tumors that are biologically homogeneous and whose expression profiles either reflect different cells of origin or other differences in disease pathogenesis [Alizadeh et al., 2000; Bittner et al., 2000]. These studies may uncover biological features of the disease that pave the way for development of improved treatments by identification of molecular targets for therapy.

SOURCES OF VARIATION AND LEVELS OF REPLICATION

The initial cDNA microarray studies involved the competitive hybridization of one mRNA sample labeled with one fluorescent dye to a second mRNA sample labeled with a second fluorescent dye on a single microarray. This type of study left many investigators believing that no replication was needed. It also led to the publication of a variety of statistical methods for comparing the expression levels in the two channels at each gene on a single microarray. There are several serious problems with this approach. For example, the relative level of red to green intensities may represent dye bias, that is, the higher affinity of one of the fluorescent dyes for DNA compared to the other dye or some artifact of the labeling reaction. The relative level could represent random variation in the diffusion of sample with one label relative to the sample with the other label to a given area on the microarray. The relative level could reflect an artifact in the RNA extraction or processing step or biological variation in the organisms from which the cells were obtained. Many potentially important sources of variation are not evaluable based on data from a single microarray. Consequently, analyses based on single microarrays are generally not sufficient.

Some important sources of variation in microarray studies can be categorized as the following.

- Between corresponding probes (i.e., spots of cDNA printed on the array surface) on different arrays for the same labeled mRNA sample
- Between labeling reaction products for the same mRNA samples
- Between specimens from the same individual
- Between individuals within a population

For example, suppose we want to determine gene expression differences between breast tumors with a mutated BRCA1 gene and tumors without a mutation. If we performed array experiments on one breast tumor with a BRCA1 mutation and one without a mutation, we would not be able to draw any valid conclusions about the relationship of BRCA1 mutations to gene expression because we have no information about the natural variation within the two populations being studied. The situation would not improve even if the tumors under investigation were large enough to perform multiple mRNA extractions and run independent array hybridizations on each extraction. Sets of tumors representative of the BRCA1-mutated population and the non-BRCA1-mutated population are necessary to draw valid conclusions about the relationship of BRCA1 mutations to gene expression.

There is sometimes confusion with regard to level of replication for microarray studies. In comparing expression profiles of BRCA1 mutated tumors to expression profiles of non-BRCA1-mutated tumors, it is not necessary to have replicate arrays of the same tumors. Having such replication may improve the sensitivity of the study, but such replications are "repeated measures" and should not be confused with the crucial need for studying multiple tumors of each type. Often the variation between individuals will be much larger than the other sources of variation, and it will be inefficient to perform replicate arrays using specimens from a small number of individuals rather than performing single arrays from a larger number of individuals.

In comparing expression profiles between two cell lines, or for a given cell line under different conditions, the concept of "individual" may be unclear. Suppose, for example, we want to compare the expression profile of a cell line before treatment with the expression profile after treatment. This type of experiment may be more complex than it sounds at first because cell lines change their expression profiles depending on the culture conditions, such as nutrients provided and degree of crowding. So, growing the cell line under what are intended to be fixed conditions may still result in different expression profiles because of differences in the confluency state of the culture at the time of cell harvesting. Although the relative size of the variation between "individuals" compared to other sources may be less for experiments involving cell lines or inbred strains of model species, the biological variation should not be assumed to be zero, as it so often is. Because of these types of effects and because of the large variation among labeling reactions, RNA used as internal reference for a set of arrays should be drawn from the reaction product of a single labeling reaction of a single uniform batch of RNA.

In some cases it is useful to obtain two specimens from the same individual. For example, if you are attempting to discover a new taxonomy of a disease based on expression profile, it is useful to establish that the classification is robust to sampling variation within the same individual. This is particularly true for class discovery, where the large number of genes makes it easy to discover interesting patterns even with random data. For many studies of human tissue, however, the tissue samples will not be large enough to provide multiple specimens for independent processing. The distinction between multiple specimens from the same individual and multiple independent labelings of one RNA sample from an individual will be most meaningful when the tissue is inhomogeneous. However, even without tissue inhomogeneity, variation may be observed among multiple specimens taken from the same individual; this variation is attributable to differences in tissue handling and RNA extraction.

There are several potential motivations for performing replicate arrays with aliquots of the same RNA sample. One motivation is to provide an estimate of the reproducibility of the microarray assay, that is, the labeling, hybridization, and quantification procedures. It is useful to know that the protocols and procedures used provide reproducible results on aliquots of the same RNA sample. A second motivation is to improve precision of the estimate of the expression profile for a given RNA sample by averaging multiple arrays for that sample. A third motivation is to use a design that is based on balancing of dyes to RNA samples or some other symmetry that can be achieved only by using multiple aliquots of each RNA sample.

There is often substantial variability in the quality of hybridizations obtained with cDNA microarrays. Some arrays have heavy background levels, others have air bubbles trapped under the coverslips on the arrays, and others have large scratches on the surfaces of the arrays. Having multiple arrays for each RNA specimen permits discarding bad arrays.

For many types of human tissue, there will not be enough material available from one individual to create multiple arrays, either by sampling multiple specimens or by labeling and arraying multiple aliquots of a single RNA sample. Some investigators use amplification protocols to create large amounts of RNA, but others remain concerned that amplification introduces distortions to the expression profiles. In studying small model species, it may be necessary to pool multiple individuals in order to obtain enough RNA for assay. For example, Jin et al. [2001] compared expression profiles for RNA samples from mixtures of flies of different ages and strains. If one assumes that there is little variation among individuals, then the conclusions about differences among mixtures of different ages or strains should apply to individual flies. This assumption, although perhaps reasonable for inbred strains of flies, is much less reasonable when applied to tumors or other tissue samples in outbred populations.

For cDNA microarrays, the probes are drops of cDNA printed on the surface of the array. There is sometimes substantial variation in the size and shapes of the corresponding spots on different arrays and between different spots on the same array. The size of the spot influences the amount of probe available for sample hybridization and the shape influences the image analysis. The samples are generally distributed inhomogeneously across the surface of cDNA arrays and this also influences the intensity of labeling. For Affymetrix arrays, the probes are more uniform and the sample is circulated so these effects are less important. For cDNA arrays, however, it is advantageous to have duplicate probes for each gene, and the duplicates should be dispersed over the surface of the array.

SAMPLE SELECTION AND EXPERIMENTAL DESIGN

Major clinical trials are conducted based on a written protocol that describes the objectives and establishes a plan for patient selection, treatment, and data analysis. The written protocol is critiqued by others. Standards of good clinical trial practice have evolved. Unfortunately, few of these characteristics apply to prognostic factor studies. There is often no written protocol, no established patient selection criteria, no clear and limited objectives, no critiqued analysis plan, and no good practice standards. As a result, the literature of prognostic factor studies is inconsistent and often unreliable, and it is difficult to make progress in developing and adopting effective new prognostic classification systems. Unfortunately, many of the defects of prognostic factor studies apply to microarray research, with the additional risks of dealing inadequately with the huge multiple testing problem and the great potential to develop overfitted models that predict poorly for independent data.

The plan for specimen selection should follow from the objectives of the microarray study. Some studies may be purely exploratory and the results should be viewed as requiring confirmation. Other studies are expected to be more than

exploratory and must therefore be designed in a focused manner. For example, if the objective is to develop a predictive model for patients with well-staged stage I lung cancer treated with surgery alone, then the specimens included in the study should be from patients with well-staged stage I lung cancer who received surgery alone.

With Affymetrix arrays, single samples are labeled and hybridized to individual arrays. cDNA arrays are generally used as a two-label system in which two RNA samples are separately labeled, mixed, and hybridized together to each array. When using cDNA arrays, one must decide on a design for allocating samples to labels and to arrays. The most commonly used design uses an aliquot of a reference RNA as one of the samples for each array. This is done so that the intensity of hybridization to a probe for a sample of interest is measured relative to the intensity of hybridization to the same probe on the same array for a standard reference sample. Because the same standard reference is used for all arrays, this measure of hybridization intensity is standardized with regard to variation in size and shape of corresponding spots on different arrays and with regard to inhomogeneities of sample distribution across the array. The latter is because the two samples are mixed and therefore inhomogeneities of distribution tend to apply to both samples. The measure of relative hybridization generally used is the logarithm of the ratio of intensities of the two labels at the probe, but analysis of variance models based on logarithm of intensities for each channel can also be used.

The reference design described above has advantages and disadvantages. The relative hybridization intensity of any sample can be contrasted to that of any other sample in a manner that is protected from variation because of spot size or sample distribution patterns. These two sources of variation can be substantial. Other designs can achieve this objective only by arraying subaliquots of each RNA specimen on two arrays, and this limits their efficiency. Because any two samples can be contrasted in this manner using a reference design, any subset of samples can be compared to any other subset of samples. For example, in studying BRCA1 mutated and BRCA1 nonmutated tumors, one might be interested in comparing samples based on their mutation status, comparing all samples based on their estrogen receptor status, or comparing samples based on the stage of the patient. The ability to contrast any pair of samples in this way is also necessary for cluster analysis of the samples. Cluster analysis is appropriate when one is attempting to identify new taxonomies of the disease. In order to do cluster analysis, one must be able to compute a distance metric between all pairs of samples. The reference design is robust to loss of arrays resulting from poor quality hybridization, although the loss of one array may entail the complete loss of information about one nonreference sample. With more complex designs in which arbitrary pairs of samples can be contrasted only indirectly through chains of comparisons, the loss of two arrays may cause a break in the chain. Finally, if a laboratory uses reference designs with the same reference sample for all of their arrays, even those for different experiments, then all of their expression profiles can be more adequately compared. Consequently, expression signatures of different tissues studied in different experiments can be sensitively compared. This latter advantage even can extend to comparisons of expression profiles made by different laboratories using reference designs with the same reference sample.

There is sometimes confusion about the role of the reference sample in reference designs. Some investigators believe that the analysis is always based on combining single array determinations of whether the Cy5 (red) label is more or less or equally intense compared to the Cy3 (green) label for a given spot on a given array. Consequently, they assume that the reference sample must be biologically relevant for comparison to the nonreference samples. In fact, the reference sample does not need to have any biological relevance. The analysis will usually involve quantitative comparisons of average logarithm of intensity ratios for one set of arrays to average log ratios for another set of arrays. More complex analyses based on analysis of variance models can also be used, but the reference sample plays the same role.

When a reference design is used, the reference sample is generally labeled with the same label on each array. There may be gene-specific dye bias not removed by the normalization process, and this would bias comparisons between reference and nonreference samples. If such comparisons are of interest, then some reverse-labeled arrays are needed. The presence of dye bias does not directly bias contrasts of relative intensity between sets of nonreference samples but may limit the ability to identify differentially expressed genes.

It is desirable that most of the genes be expressed in the reference sample but not expressed at so high a level as to saturate the intensity detection system. It is not possible to obtain control of the variations due to spot size and sample distribution for genes that are not expressed in the reference sample. Often, the reference sample consists of a mixture of cell lines so that nearly all genes will be expressed to some level. It is also important that a single batch of reference RNA is used for all arrays in a reference design. Different batches of reference RNA may have quite different expression profiles. When assaying samples collected over a long period, it is generally best to freeze the RNA samples and to perform the microarray assays at one time when all reagents can be standardized.

A disadvantage of the reference design is that half of the hybridizations are used for the reference sample, which may be of no real interest. Block designs [Dobbin and Simon, 2001] or loop designs [Kerr and Churchill, 2001] are alternatives that can be used in simple situations and analyzed using analysis of variance or mixed model [Wolfinger et al., 2002] methods. For example, suppose one wanted to compare BRCA1 mutated breast tumors to BRCA1 nonmutated breast tumors, that equal numbers of each tumor were available, and that no other comparisons or other analyses were of interest. Then one could hybridize on each array one BRCA1 mutated tumor sample with one BRCA1 nonmutated sample. On half of the arrays the BRCA1 mutated tumors should be labeled with the red dye and on the other half the BRCA1 nonmutated tumors should be labeled with the red dye. This block design can accommodate n samples of each type using only n microarrays. No reference RNA is used at all. The reference design would require $2n$ arrays. This design has limitations, however. If one wants to cluster the expression profiles for the different arrays, one must be able to compare (measure the distance between) each sample to each other sample. Because no reference sample is used and because different tumors even within a mutation class will generally have very heterogeneous expression profiles, one cannot effectively contrast arbitrary pairs of samples. The contrasts computed will generally be imprecise because of variation in size and shape

of corresponding spots on different arrays and variation in sample distribution patterns on different arrays. This can be avoided if two aliquots of each sample are arrayed. For example the first array would consist of a BRCA1 mutated sample from the first tumor (B_1) labeled red and hybridized with the first nonmutated tumor sample (A_1) labeled green. The second array would consist of a second aliquot of A_1 labeled red hybridized with a second BRCA1 tumor sample labeled green (B_2). The third array would consist of a second aliquot of B_2 labeled red and hybridized with a nonmutated tumor from a second patient (A_2). This loop continues and concludes with an aliquot of the last nonmutated tumor (A_n) labeled red and hybridized with a second aliquot of the first mutated tumor B_1 to complete the loop. This uses $2n$ arrays to study n mutated and n nonmutated samples, using two aliquots of each sample. The loops permit all pairs of samples to be contrasted in a manner that adjusts for variation in spot size and sample distribution patterns, but the number of indirect terms adds substantial variance to many of these contrasts [Dobbin and Simon, 2001]. Loop designs are less robust against the presence of bad quality arrays and they require enough RNA be available for each sample to hybridize to multiple arrays.

Block and loop designs can effectively reduce the number of arrays required for a given number of nonreference samples, but they do not possess many of the other advantages of the reference design described above. Because availability of nonreference samples is the limiting feature of tissue-based studies, getting the most information from assay of those tissues may be more important in many instances than reducing the number of arrays that must be purchased.

In some studies one may wish to compare disease tissue to normal tissue and also to compare disease samples to each other. One might collect normal tissue from each patient and hybridize disease tissue versus normal tissue from the same patient on each array, half of the arrays giving the normal tissue red label and the other half giving the normal tissue green label. This block design may be very effective for comparing disease to normal tissue in settings where we expect substantial differences in the normal tissues of different patients. This design might be reasonable for comparing subsets of disease tissue or for clustering the disease tissue if (1) we are only interested in measuring gene expression for disease tissue relative to expression for the normal tissue of the same patient and (2) we expect enough genes to be expressed in the normal tissue samples.

If we are interested in comparing subsets of disease tissue with regard to gene expression relative to some common standard, rather than relative to normal tissue expression in the same patient, then a reference design might be more appropriate. One could use pooled normal RNA as the reference if enough genes were expressed in normal tissue. However, if there is still some interest in comparing the pooled normal sample to the samples from diseased tissue, it would be necessary to perform a forward and reverse label hybridization for at least some of the specimens, because otherwise the disease-to-normal comparison may be contaminated by dye bias. Alternatively, a reference sample optimized to express many genes could be used as the reference sample. In this situation, performing array hybridizations of the individual normal samples with the reference in addition to hybridizations of the individual disease samples with the reference would allow for the comparison of disease and normal samples.

COMPARING EXPRESSION PROFILES AMONG PRE-DEFINED CLASSES

Although our purpose here is to discuss study design issues, the analysis plan cannot be ignored in the design of a study. Many investigators do not realize that cluster analysis of the expression profiles of the samples is not generally an appropriate analysis strategy for comparing predefined classes. Cluster analysis is "unsupervised" in the sense that it does not use the information of which sample comes from which class. Cluster analysis is based on a metric for measuring distance between expression profiles. This metric is generally based on the complete set of genes measured on the microarray or on a subset showing greatest variability across the arrays. It is not based on information about which genes are informative for discriminating among the predefined classes. Consequently, cluster analysis is not very powerful for distinguishing classes that differ with regard to a relatively small number of genes.

Some investigators are probably attracted to cluster analysis because if it separates the predefined classes without using the class membership information, then the separation is most likely valid. Conversely, if a supervised method distinguishes the classes, then one must demonstrate that the separation would be valid for independent data and is not just the result of multiple testing or using the same data to identify the genes and to demonstrate the separation. This is an important concern, but it is better to use a supervised analysis performed in a way that takes account of the multiple testing and overfitting problems. Dudoit et al. [2002] compared a variety of supervised methods for predicting membership in predefined classes.

In comparing expression profiles among predefined classes, the greatest interest is usually in identifying the genes that distinguish the classes and in quantifying how adequately class membership can be predicted based on expression profile. Multiple testing and model overfitting concerns must be considered in addressing both of these objectives. The multiple testing problem can be addressed in several ways. One is by use of global tests that document that the expression profiles are significantly different among the classes before attempting to determine which genes account for the differences. This approach was used in comparing expression profiles of chronic lymphocytic leukemia specimens with and without immunoglobulin gene mutations [Rosenwald et al., 2001]. They used as a test statistic the number of genes significantly different between the mutation classes at a nominal 0.001 level based on t-tests of individual genes. They generated the null permutation distribution of the test statistic by permuting the labels of which samples were mutated and which were nonmutated and recomputing the t-tests and test statistic (i.e., the number of genes significant at a nominal $P<0.001$ level) for each permutation.

There is a large literature on multiple testing methods but, for microarray studies, controlling the experiment-wise type 1 error rate is generally viewed as too conservative. However, because the genes reported as significantly discriminating predefined classes will usually be subject to confirmatory studies that may be based on a different assay methodology requiring development of gene specific reagents such as antibodies or DNA probes, there is a strong motivation to not allow too large a proportion of false positives in comparisons. This can be accomplished to some extent by comparing the classes one gene at a time using two-sample or

k-sample tests based on a stringent statistical significance level. For example, if 5,000 genes are tested, then the expected number of false positives statistically significant at the $P<0.001$ level is no greater than 5. If there are 50 genes that are significant at that level, then the expected false-discovery proportion is no greater than 10%, which is generally not a problem. More refined methods of controlling the false discovery rate have been developed that can account for the correlation structure among the genes [Tusher et al., 2001] and can control the false discovery rate probabilistically, not just in expectation [Korn et al., 2001].

To date there have been few reports of methods for planning sample size for the varied objectives of microarray-based studies. We will describe here a relatively straightforward method for planning sample size for testing whether a particular gene is differentially expressed between two predefined classes. Because such a test could be applied to each gene, this approach provides some information for planning the size of microarray studies and may be useful until more comprehensive methods become available.

The following simple guide to sample size planning may be used for two-color arrays using reference designs or for single-label oligonucleotide arrays. Suppose that some function of the expression levels, for example log ratios for cDNA arrays, are approximately normally distributed in the two classes. Let σ denote the standard deviation of the expression level among samples within the same class and suppose that the means of the two classes differ by δ. For example, with base 2 log ratios or log intensities, a value of $\delta = 1$ corresponds to a twofold difference between classes. We assume that the two classes will be compared with regard to level of expression of each gene and that a statistically significant difference will be declared if the null hypothesis can be rejected at a significance level α. The level α will be set stringently in order to limit the number of false-positive findings because thousands of genes will be analyzed. The desired statistical power for detecting statistical significance when the true difference in mean expression levels between the classes is δ will be denoted 1-β. This requires

$$\mathrm{n} = 4(z_{\alpha/2} + z_{\beta})^2/(\delta/\sigma)^2 \tag{1}$$

total tissue samples, where $z_{\alpha/2}$ and z_{β} denote the corresponding percentiles of the standard normal distribution [Desu and Raghavarao, 1990]. If the ratio of sample sizes in the two groups is k:1 instead of 1:1, then the total sample size increases by a factor of $(\mathrm{k}+1)^2/4\mathrm{k}$. The fact that expression levels for many genes will be examined indicates that the size of α and β should be smaller than for experiments where the focus is on a single endpoint. The expected number of false-positive genes identified as differentially expressed between the two classes is αN, where N is the number of genes equally expressed in the two classes. N could be as large as the number of genes on the array. In order to limit the expected number of false-positive results to 1, we require that $\alpha = 1/N$. Because N will generally be in excess of 1,000, we should use an α value no greater than 0.001. Similarly, the expected number of false-negative conclusions for genes that are actually differentially expressed between the two classes by δ-fold is βM, where M is the number of such genes. If we want the number of false negatives to be F, then $\beta = F/M$. Hence, β should equal the proportion of the differentially expressed genes that one is willing to tolerate missing. In general, α and β should not exceed 0.001 and 0.05, respectively.

The parameter σ can usually be estimated based on data showing the degree of variation of expression values among similar biological tissue samples. For log-ratio expression levels, the value of σ is often very low for many clones that are either not expressed or not differentially expressed in any of the samples, and in the range of 0.25–1 (using base 2 logarithms) for the remaining clones. The within-class variability depends somewhat on the type of specimens (human tumor samples have greater variability than cell lines) and on the heterogeneity of the classes. The parameter δ represents the size of difference between the two classes we want to be able to detect. For log2 ratios, $\delta = 1$ is often considered reasonable because it corresponds to a twofold difference in expression level between classes. Using $\alpha = 0.001$, $\beta = 0.05$, $\delta = 1$ and $\sigma = 0.75$ in the above formula gives a required sample size of approximately 56 total tissue samples. With the same parameters except $\sigma = 0.5$, approximately 26 total samples are required from this perspective. To detect smaller differences, many more samples may be required. For example, in order to detect a 1.5-fold difference between the classes ($\delta = \log 2(1.5) = 0.585$) 72 samples are required for $\sigma = 0.5$.

In addition to identifying genes that are differentially expressed between predefined classes, sometimes it is also of interest to estimate the accuracy with which one can predict class membership based on expression profile. There is a large literature of methods for multivariate prediction of class membership, but few of these methods were developed in the context of studies where the number of candidate predictors is at least one order of magnitude larger than the number of cases. This is the situation with microarray data, however, and one important consequence is that resubstitution estimates of misclassification rate are likely to be extremely biased. Consequently, it is imperative that some type of training/test sample stratification, cross-validation, or bootstrapping be used to estimate misclassification rate. Radmacher et al. [2002] describe a paradigm for estimating misclassification rate using leave-one-out cross-validation. They also recommend performing a permutation test using the cross-validated misclassification rate as test statistic in order to establish that the predictive accuracy is significantly greater than could be achieved by chance.

There is no generally accepted theory for planning sample size for developing multivariate predictors with numerous candidate variables. Rules of thumb that are sometimes used, such as having 5–10 cases for each candidate variable, would suggest that tens of thousands of cases are needed for such microarray studies. This is clearly not practical and the sample size surely depends on a more precise statement of the adequacy criteria. With microarray studies of practical size, it is very difficult to select an optimal set of predictive variables. This is illustrated in Table I, which is based on a normal linear regression rather than a classification, but the same modeling issues apply. In Table I there are m predictor genes with nonzero regression coefficients and n noise genes. The size of the true regression coefficients for the good predictors was computed to provide 95% statistical power at a 0.001 significance level in a univariate analysis with 26 cases. The data were generated with all gene expression levels being independent normal random variables with mean zero and variance 1. The last column shows a simulation estimate of the probability that all of the m truly prognostic genes and none of the noise genes are simultaneously univariately significant at $P < 0.001$. Univariate significance levels

TABLE I. Probability of Selecting All Good Predictor Genes in Presence of "Noise" Genes

Noise genes	True predictors	Number of specimens	Probability of selecting correct model
1,000	2	25	0.03
1,000	2	50	0.62
1,000	2	75	0.91
1,000	3	75	0.54
1,000	3	100	0.92
1,000	4	100	0.49
1,000	4	150	0.95

were used for the table; no model selection was performed. Although the regression coefficients are set so that the truly prognostic genes have power 0.95 with only 26 cases, much larger sample sizes are needed to identify the optimal model. This result is intuitive when one considers the number of parameters that must be estimated to identify an optimal model: a model containing few variables requires not only the accurate estimation of the regression coefficients for the true predictor genes, but also the accurate estimation of a large number of zero regression coefficients for the noise genes. The multiplicity of the problem is determined by the number of candidate predictors, not the number of variables in the optimal model.

In most cases, there will be many genes that are differentially expressed between the predefined classes, and it will be sufficient to identify some number of these genes and to provide an unbiased estimate of the predictiveness of multivariate models that use these variables without attempting to identify an "optimal" model. For example, the compound covariate predictor described by Radmacher et al. [2002] and used by Hedenfalk et al. [2001] is a very simple model based on earlier ideas of Tukey [1993], who suggested this approach in situations with many candidate predictors. Not only are optimal models difficult to assess, but they also may have little clinical relevance. The DNA microarray is an efficient assay for screening genes, but is not necessarily the best assay platform for use in clinical diagnostic situations. There are other assays that are more sensitive or easier to use in pathology departments. Quantification of mRNA transcripts or their protein products may be easier for some genes than for others. Consequently, it may not be realistic to put too much effort into developing optimal models from gene expression arrays.

DEVELOPING PROGNOSTIC MODELS

Many of the considerations described in the previous section for comparing predefined classes apply equally to developing prognostic models. In some cases, however, the clinical endpoint is continuous and right censored, and the details of sample size planning require modification. The analog to expression (1) for sample size planning is [Hsieh and Lavori, 2000]:

$$D = (z_{\alpha/2} + z_{\beta})^2/(\tau \ln \delta)^2. \quad (2)$$

In expression (2), τ denotes the standard deviation of the log ratio or log intensity level of the gene over the entire set of samples, because there are no predefined

classes. δ denotes the hazard ratio associated with a one-unit change in the log ratio or log intensity x, and ln denotes the natural logarithm. Note that we are assuming that the log ratios or log intensities are based on logarithms to the base 2, so a one-unit change in x represents a twofold change. If $\tau = 0.5$ and $\delta = 2$, then 203 events are required for a two-sided significance level 0.001 and power 0.95. This makes for a very large study. The large number of events results from assuming that a doubling of hazard rate is associated with a change in log-ratio that amounts to 2 SD. A more realistic scenario is $\tau = 0.75$ and $\delta = 3$, which results in 36 events required. Differences of less than twofold are difficult to measure reproducibly with microarrays. Hence, genes that have low standard deviations across the entire set of samples would be difficult to use for prognostic prediction in clinical situations.

The development of prognostic models is often complex, with several types of models being considered. After having developed a model, one often wishes to evaluate how predictive it is. It is common to evaluate the predictiveness of the model on the same data used for developing the model. This produces biased estimates of model predictiveness, and the bias may be extremely large for microarray data where there are thousands of predictors. In some cases, less biased estimates of model predictiveness can be obtained by embedding the model development process in a cross-validation or bootstrap procedure. The model development process, however, often is not easily reduced to an algorithm that can be embedded in this way, and it will be best to use a split-sample approach. One portion of the data is used for developing the model, and a test set is reserved and not used until a fully specified model is selected based on the training set. This method was recently used in the development of a prognostic model based on expression profiles for large-cell lymphoma [Lymphoma/Leukemia Molecular Profiling Project, 2002]. One third of the total number of cases was reserved for the training set. Although there are rarely enough cases even for the training set, often the worst risk is not to reserve a separate investigation test set. Without a test set or some other valid method for obtaining an unbiased estimate of model predictiveness, the entire investigation becomes an exploratory study that must be independently confirmed.

CLASS DISCOVERY

Some studies are designed to determine whether tissue specimens of patients with a specified disease are homogeneous or whether new subclasses of the disease may be discovered based on gene expression profiles. These studies are exploratory, and methods of sample size planning have not been developed for them. Cluster analysis methods are generally used for these types of studies [Eisen et al., 1998]. There is a wide variety of cluster analysis algorithms, and there is little basis for selecting one approach over others. Most methods always produce clusters, and it is easy for the investigators to overinterpret the finding of clusters. Some methods have been developed to evaluate the robustness of the clusters to perturbations in the data [McShane et al., 2001; Kerr and Churchill, 2001] or to compare the degree of clustering relative to that expected with data from a nonclustered population distribution [Tibshirani et al., 2000]. It is important to use such methods. It can also be useful to have two samples from each tissue specimen, or from as many tissue

specimens as would support multiple sampling. This will provide some information about how homogeneous the tissue samples are relative to the disease clusters produced.

REFERENCES

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, and Staudt LM. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–11.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536–40.

Desu MM, Raghavarao D. 1990. Sample Size Methodology. Boston: Academic Press. p 30.

Dobbin K and Simon R. 2001. Issues in the design of a microarray experiment. Technical Report 04, Biometric Research Branch, National Cancer Institute. http://linus.nci.nih.gov/~brb/TechReport.htm.

Dudoit S, Fridlyand J, Speed TP. 2002. Comparison of discrimination methods for classification of tumors using DNA microarrays. J Am Stat Assoc 97:77–87.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863–8.

Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Dowing J, Caligiuri M, Bloomfield C, Lander E. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–7.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi O, Borg A, Trent J. 2001. Gene expression profiles of hereditary breast cancer. N Engl J Med 344:549.

Hsieh FY, Lavori PW. 2000. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. Controlled Clin Trials 21:552–60.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance is drosophila melanogaster. Nature Genet 29:389–95.

Kerr MK, Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. Genet Res 77:123–8.

Kerr MK, Churchill GA. 2001. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA 98:8961–5.

Korn EL, Troendle JF, McShane LM, Simon R. 2001. Controlling the number of false discoveries; application to high dimensional genomic data. Technical report 03, Biometric Research Branch, National Cancer Institute, http://linus.nci.nih.gov/~brb/TechReport.htm

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnol 14:1675–80.

Lymphoma/Leukemia Molecular Profiling Project. 2002. Molecular diagnosis and clinical outcome prediction in diffuse large B-cell lymphoma. N Engl J Med (In press).

McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. In press. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics

Radmacher MD, McShane LM, Simon R. In press. A paradigm for class prediction using gene expression profiles. J Computat Biol.

Rosenwald A, Alizadeh AA, Widhopf G, Simon R, Davis RE, Yu X, Yang L, Pickeral O, Powell J, Botstein D, Byrd JC, Grever MR, Chiorazzi N, Wilson WH, Kipps TJ, Brown PO, Staudt LM. 2001. Relation of gene expression phenotype to immunoglobulin mutation genotype in chronic lymphocytic leukemia. J Exp Med 194:1639–48.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JCF, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO. 2000. Systematic variation in gene expression patterns in human cancer cell lines. Nature Genet 24:227–35.

Shalon D, Smith SJ, Brown PO. 1996. A DNA micro-array system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6:639–45.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9:3273–97.

Tibshirani R, Walther G, Hastie T. 2000. Estimating the number of clusters in a dataset via the Gap statistic. J R Stat Soc B 63:63:411–23.

Tukey JW. 1993. Tightening the clinical trial. Controll Clin Trials 14:266–85.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116–21.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J Computat Biol 8:625–38.