Experimental Design of DNA Microarray Experiments

Richard M. Simon and Kevin Dobbin

(To Appear in Biotechniques)

Richard Simon, D.Sc.
Chief, Biometric Research Branch
National Cancer Institute
9000 Rockville Pike
MSC 7434
Bethesda MD 20892-7434
rsimon@nih.gov

Introduction

DNA microarray experiments require planning. Planning is driven by the experimental objectives. Good DNA microarray experiments have clear objectives. The objectives are not based on gene-specific mechanistic hypotheses like objectives of many other biological experiments, but it is erroneous to conceive of DNA microarray investigations as aimless data mining in search of unanticipated patterns that will provide answers to unasked questions.

The objectives of many DNA microarray experiments can be characterized as either *class comparison*, *class prediction*, or *class discovery* (4). The objective of class comparison studies is to identify the genes that are differentially expressed in cells from different types of tissue, different kinds of patients, or in cells exposed to different experimental conditions. One example of class comparison is comparing gene expression in tumor tissue for patients who respond to a given treatment to gene expression in patients with the same cancer diagnosis who don't respond to therapy (7). Another example of class comparison is comparing gene expression in kidney tissue of mice after 2 hours of ischemia to gene expression in kidney tissue of normal mice. The characteristic feature of class comparison studies is that the classes to be compared are defined independently of the expression data. The objective is to see how the expression profiles differ among the classes.

Class prediction problems are similar to class comparison problems in that the classes are defined independently of the expression data. The emphasis in class prediction

problems is in developing a multi-gene formula that can be applied to expression profiles of samples whose class is unknown, and to predict the class of the new samples. Using the example in the previous paragraph, class comparison involves identifying the genes that are differentially expressed between patients who respond to a specified treatment and those who don't respond. Developing a formula that can be used to predict whether a new patient will respond to that therapy based on the gene expression profile of his or her tumor, is class prediction. Class prediction is particularly useful in medical problems of therapy selection or diagnostic classification or prognostic prediction.

The third type of objective, class discovery, is quite different from class comparison or class prediction. In class discovery there is no classification defined independently of the expression profiles. The objective is to discover subsets (clusters) of the cases revealed by gene expression profiles and to identify the genes that distinguish the clusters. For example, Bittner et al. (1) examined expression profiles of patients with advanced malignant melanoma. The focus of the study was on attempting to identify a new taxonomy of advanced melanoma based on gene expression. No useful clinical classification existed. Class prediction also includes studies whose objective is to discover classes of genes that are co-regulated.

Levels of replication

A single dual-label DNA microarray assay provides a comparison of expression profiles for two RNA samples. The same is obtained for two Affymetrix GeneChip™ assays. With only those two expression profiles, one cannot determine whether the expression profiles in the two RNA samples differ by more than experimental variability. This is because the magnitude of all of the relevant sources of variability cannot be estimated from the data consisting only of those two expression profiles. For example, the variability in expression profiles resulting from labeling the sample cannot be estimated.

Investigators often ask, "how many replicates do I need." If you had enough RNA in the two specimens to draw aliquots used to independently label and hybridize the RNA to many arrays, you could validly determine whether the expression profiles for those two RNA samples differed. Unfortunately, that is not usually the biologically relevant question. You will probably be thinking of comparing those two RNA samples because they were collected from different tissues or from cells under different conditions. The biologically relevant question is usually whether the two types of tissue differ with regard to expression profile, or to determine the effects on gene expression of changing the experimental conditions of the tissue culture. The two RNA samples may not be representative of the two tissue types. There may be substantial biological variability in gene expression among tissues of the same type and so comparing one RNA sample of one tissue type to one RNA sample of the other tissue type does not answer the biological question. The same applies to determining the effect of experimental manipulations on gene expression for cells

grown in tissue culture. There will be variability in gene expression if the experiment is repeated because of variation in the administration of the experimental manipulations and differences in cell growth and harvesting.

For class comparison and class prediction studies, multiple biological samples are needed, not replicate arrays of the same RNA samples. It is useful to have a few technical replicates of the same RNA sample to ensure that your procedures, reagents and equipment are working properly. Technical replicates should show good agreement. Technical replicates can also be protection against bad quality hybridizations. But technical replicates are not a substitute for biological replicates; that is, having enough samples of biologically independent specimens. For tissue culture experiments, biologically independent specimens means specimens obtained from independent replications of the entire experiment.

For studies attempting to discover new taxonomies of disease, it is useful to have expression profiles from cells in different parts of the biopsy specimen of the same patient, as well as having many independent patients represented. This helps you to evaluate whether the clusters you may obtain from subsequent analysis of the expression data represent a reproducible disease taxonomy.

Pooling of Samples

Some investigators pool samples in hope that by pooling they can reduce the number of microarrays needed. For example, in comparing two tissue types, a pool of one type of

tissue could be compared to a pool of the other tissue type. Replicate arrays might be performed on each pooled sample. Although the pooled sample approach may be applicable for preliminary screening for differentially expressed genes, the approach does not provide a valid basis for statistical analysis. If only one array of each pooled sample is prepared, then there is no estimate of the variability associated with independently labeling and hybridizing the same pool onto different arrays. Even if the pools are hybridized to replicate arrays, you cannot evaluate how adequately a pool of that number of RNA specimens reflects the population of that tissue type. Unless multiple biologically independent pools of each type are arrayed, only the pooled samples themselves can be compared, not the populations from which they were derived. Selecting independent pools of samples is necessary in studying small model species where individuals must be combined in order to obtain enough RNA for assay (5).

Pairing Samples for Co-hybridization on Two-Color Microarrays

With Affymetrix GeneChips^{TM}, single samples are labeled and hybridized to individual arrays. Spotted cDNA arrays, however, generally use a dual-label system in which two RNA samples are separately reverse transcribed, labeled, mixed and hybridized together to each array. When using dual-label arrays one must decide on a design for pairing and labeling samples.

The Reference Design

The *reference design*, uses an aliquot of a common reference RNA as one of the samples hybridized to each array. This is done so that the intensity of hybridization to a spot for a sample of interest is measured relative to the intensity of hybridization to the same spot on the same array for the reference sample. This relative hybridization intensity is standardized against variation in size and shape of corresponding spots on different arrays. Relative intensity is also standardized with regard to variation in sample distribution across each array since the two samples are mixed and therefore distributed similarly. The measure of relative hybridization generally used is the logarithm of the ratio of intensities of the two labels at the spot.

The reference design is illustrated in Figure 1. Generally, the reference is labeled with the same dye on each array. Any gene-specific dye bias not removed by normalization affects all arrays similarly and does not bias class comparisons. Using a reference design, any subset of samples can be compared to any other subset of samples, hence the design is not dependent on the specification of a single mode of classification. The reference design is also convenient for class discovery using cluster analysis since the relative expression measurements are consistent with regard to the same reference.

The Balanced Block Design

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. Balanced block designs (2) are alternatives that can be used in simple situations. For example, suppose one wished to compare BRCA1 mutated breast tumors to BRCA1 non-mutated breast tumors, that equal

numbers of each tumor were available and that no other comparisons or other analyses were of interest. One could hybridize on each array one BRCA1 mutated tumor sample with one non-mutated sample (Figure 2). On half of the arrays the BRCA1 mutated tumors should be labeled with the red dye and on the other half the non-mutated tumors should be labeled with the red dye. The analysis of data for the block design is based on an analysis of variance model for channel specific background adjusted intensities (3). The 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 to accommodate n non-reference samples from each of the two classes.

Although the balanced block design is very efficient in use of arrays, it has some serious limitations. For one, cluster analysis of the expression profiles cannot be performed effectively. Without a common reference, any comparisons between expression profiles of samples on different arrays will be subject to noise resulting from variation in size and shape of corresponding spots on different arrays and variation in sample distribution patterns on individual arrays (2). Also, since it is difficult or impossible to pair the samples simultaneously with regard to all of the class comparisons of interest, the block design is most effective when there is a single type of class comparison. The block design is also less effective than the reference design when there is large inter-sample variability and when the number of samples, rather than the number of arrays, is limiting (2).

The Loop Design

Loop designs (6) are another alternative to reference designs. When cluster analysis is planned, two aliquots of each sample must be arrayed for the loop design (Figure 3). The arrays link the samples together in a loop pattern. This design uses n arrays to study n samples, using two aliquots of each sample. The loop permits all pairs of samples to be contrasted in a manner that controls for variation in spot size and sample distribution patterns using a statistical model. Contrasting two samples far apart in the loop, however, involves modeling many indirect effects corresponding to the arrays linking the two samples of interest and this adds substantial variance to many of these contrasts (2). Consequently loop designs are very inferior to reference designs for cluster analysis. Loop designs can be used for class comparisons, but are less efficient than block designs. Loop designs are also less robust against the presence of bad quality arrays; two bad arrays can break the loop apart.

Reverse Labeling

Some investigators believe that all arrays should be performed both forward and reverse labeled. In general, this is unnecessary (3).

The relative labeling intensity of the Cy3 and Cy5 may be different for different genes. Although the normalization process may remove average dye bias, gene-specific dye bias may remain. This is not important for comparing classes of non-reference samples using a reference design when the reference is consistently assigned the same label. Suppose, however, that for a group of patients we wanted to compare tumor tissue to matched normal tissue from the same individual using dual-label microarrays. One effective

design would be to pair tumor and normal tissues from the same patient for co-hybridization on the same array using the balanced block design (Figure 2). In half of these arrays the tumor should be labeled with Cy3 and in the other half of the arrays the tumor should be labeled with Cy5. It is not necessary to perform any reverse labeled replicate arrays of the tissues from the same patient (3). Gene-specific dye bias can be estimated in the balanced block design and used to adjust class comparisons without any reverse labeling arrays for the same two specimens. For a fixed total number of arrays, it is best to use the available arrays to assay tissue from new patients, using the balanced block design described, rather than to perform replicate reverse labeled arrays for single patients. The balanced block design is also best when there are n tumor tissues and n normal tissues even though the tissues are not from the same patients, or for comparing any two classes of samples. In these cases, the samples may be randomly paired, or paired based on balance with regard to potentially confounding variables such as the age of the specimens.

In some cases a reference design is used in which the primary objective is comparison of classes of the non-reference samples but comparison to the internal reference is a secondary objective. For example, there may be several types of transgenic mouse breast tumors for comparison and the internal reference may be a pool of normal mouse breast epithelium. Because the primary interest is comparison among multiple tumor models or clustering the expression profiles of the tumors, a reference design may be chosen. Use of a pool of normal breast epithelium as the internal reference, rather than a mixture of cell lines, reflects some interest in comparison of expression profiles in tumor relative to

normal breast epithelium. Comparison to a single pool of normal breast epithelium is somewhat problematic, however, for reasons described previously in the section on pooling. Nevertheless, the comparison may be of interest.

In order ensure that the comparison of tumor expression to that of the reference is not distorted by gene-specific dye bias when using a reference design, some reverse labeled arrays are needed. One can then fit a statistical analysis of variance model to the logarithms of the intensities for each channel as described by Dobbin et al. (3). A separate analysis of variance model is fit for each gene and from the model one estimates the residual dye bias after normalization. These estimates are used by the model to automatically adjust the comparison of gene expression in tumor versus reference. Not all arrays need to be reverse labeled; 5-10 reverse labeled pairs of arrays will generally be adequate. Except for this purpose of comparison to the reference using a reference design, however, we recommend against reverse labeling of the same two RNA samples forward labeled on another array.

Recommendations for dual-label designs

When experimental objectives include discovery of new classes among the samples, then we recommend the reference design. If only comparison of pre-defined classes are planned but there are several kinds of comparisons to be made, then we again recommend the reference design. If only a single kind of class comparison is of interest and the expense of the microarrays is an important issue, then the balanced block design is recommended. If there is interest in measuring expression for each diseased specimen

relative to a paired normal specimen from the same individual, then the balanced block design is again recommended. We do not recommend use of the loop design in most circumstances.

If a common reference design is used, the reference RNA need not represent a biologically relevant contrast to the experimental samples. The main purpose of the reference RNA is to enable relative expression measures to be calculated in order to avoid technical measures of variation. Many investigators use reference RNA from a mixture of cell lines so that most genes will be expressed at a level that permits both increased and decreased expression in the experimental specimens to be measured. Using the same reference RNA for all experiments of a laboratory makes it possible to compare expression among different experiments, although other sources of variation may make this difficult.

In cases where the common reference represents a pool of RNA from a source for which comparison is of interest, we recommend that the reference design (e.g. with the common reference consistently labeled with say Cy3) be supplemented by some arrays (e.g. 5-10) which represent dye swaps of the main set of arrays.

Number of Biologically Independent Samples Needed

The number of independent biological samples needed depends on the objectives of the experiment. We will describe here a relatively straightforward method for planning sample size for testing whether a particular gene is differentially expressed between two pre-defined classes. Such a test can be applied to each gene if we adjust for the number of comparisons involved (8).

This approach to sample size planning may be used for dual-channel arrays using reference designs or for single label oligonucleotide arrays. For dual-channel arrays the expression level for a gene is the log ratio of intensity relative to the reference sample; for Affymetrix GeneChip^{TM} arrays it is usually the log signal. The approach to sample size planning described here is based on the assumption that the gene-specific expression measurements are approximately normally distributed among samples of the same class. Let σ denote the standard deviation of the log expression level among samples within the same class and suppose that the means of the log expression in the two classes differ by δ for a particular gene. For example with base 2 logarithms, a value of $\delta=1$ corresponds to a 2-fold 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 *significance level* is the probability of concluding that the gene is differentially expressed between the two classes when in fact the means are the same ($\delta=0$). The significance level α will be set stringently in order to limit the number of false positive findings since thousands of genes will be analyzed. The desired statistical power will be denoted $1-\beta$.

Statistical power is the probability of obtaining statistical significance in comparing gene expression among the two classes when the true difference in mean expression levels between the classes is δ. Statistical power is one minus the false negative rate (β).

Under these conditions, the number of total samples required from different individuals or different replications of the experiment is approximately

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

where $z_{\alpha/2}$ and z_{β} denote the corresponding percentiles of the standard normal distribution (8). A standard normal distribution has mean zero and standard deviation one. The total area under the standard normal distribution (between the curve and the x-axis) is one. The area under the part of the curve to the left of the x-axis value of $z_{\alpha/2}$ is $\alpha/2$. The area to the left of the x-axis value of z_{β} is β. The n in formula (1) is the total number of experimental samples and also the number of arrays needed. 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 $(k+1)^2/4k$ compared to formula (1).

The fact that expression levels for many genes will be examined indicates that the size of α should be smaller than for experiments where the focus is on a single endpoint. We recommend planning the sample size using α=0.001 and β=0.05. In our experience, most genes are not differentially expressed. Using α=0.001 results in 10 false discoveries per 10,000 non-differentially expressed genes. This is less conservative than the multiple comparison adjustments commonly used for clinical trials, but seems reasonable for

microarray studies where findings may be followed up in other kinds of assays. Using α=0.005, however, results in 50 false discoveries per 10,000 non-differentially expressed genes, which is too large a number of false leads even for most microarray studies. For α=0.001 and β=0.05, the standard normal percentiles are $z_{\alpha/2}$ =-3.29 and z_{β} =-1.645.

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 in dual-label arrays using the reference design, we have seen the median value of σ of approximately 0.5 (using base 2 logarithms) for human tissue samples. The parameter δ represents the size of difference between the two classes we wish to be able to detect. For $\log_2$ ratios, δ=1 is often considered reasonable as it corresponds to a 2-fold difference in expression level between classes. Using α=0.001, β=0.05, δ=1 and σ=0.50 in the above formula gives a required sample size of approximately 25 total samples.

The within class variability depends on the type of specimens; human tissue samples have greater variability than inbred strains of mice or cell lines. In experiments studying microarrays of kidney tissue for inbred strains of mice, the median standard deviation of log ratios for normal kidney was approximately 0.25, with little variation among genes. Using α=0.001, β=0.05, δ=1 and σ=0.25 in the above formula gives a required sample size of approximately 7 total samples. With such small sample sizes, formula (1) based on approximate normality would be more accurate if based on the t distribution, rather than the normal distribution. The constants in expression (1) corresponding to standard normal percentiles should be replaced by percentiles of the t distribution with mean zero,

and n-2 degrees of freedom, where n is the total number of samples. Since expression (1) determines n, the expression must be solved iteratively for n. In the case of α=0.001, β=0.05, δ=1 and σ=0.25, we find that a total of 12 samples, 6 from each of the two classes being compared, are required for comparing the two classes. If this were a time series experiment with more than two time points, then one should plan for 6 animals per time point in order to enable expression profiles to be compared for all pairs of time points.

When dual-label arrays are used with the balanced block design to compare either naturally paired or independent samples from two classes, a similar formulas applies:

$$n = 2(z_{\alpha/2} + z_{\beta})^2 / (\delta/\tau)^2 \qquad (2)$$

n is the total number of independent experimental samples needed, as in expression (1) for the reference design, but only n/2 arrays are needed. For the balanced block design τ represents the standard deviation of variation across arrays in the log ratio of expression levels of samples, one from each class being compared (3). Preliminary data is generally needed to estimate τ in this case. In several cases that we have examined, τ^2 was approximately equal to 2 σ^2 and hence the total number of required non-reference samples was approximately the same for the reference design as for the balanced block design. The balanced block design required half as many arrays, however.

Adequate methods for determining the number of samples required for gene expression studies whose objectives are class prediction or class discovery have not yet been developed. For such objectives the reference design is strongly recommended. The sample size formula (1) provides reasonable minimum sample sizes for class prediction studies. Often, however, developing multivariate class predictors or survival predictors involves extensive analyses beyond determining the genes that are informative individually. Also, a substantial portion of the cases may be set aside as a validation set for estimating the misclassification rate of the model developed on the test set of data. Consequently, larger sample sizes are generally needed for class prediction studies (7,9).

References

1. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor,E., Hendrix, M., Radmacher, M., Simon, R., et. al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406: 536-540.

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

3. Dobbin, K., Shih, J., and Simon R. 2003 Statistical design of reverse dye microarrays. Bioinformatics (In Press).

4. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., et. al. 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

5. Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., and Gibson, G. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nature Genetics 29:389-395.

6. Kerr, M.K., and Churchill, G.A. 2001 Statistical design and the analysis of gene expression microarray data. Genet Res 77:123-8.

7. Rosenwald, A., Wright, G., Chan, W. C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., et. al. 2002. The Lymphoma/Leukemia Molecular Profiling Project.The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. N Engl J Med 346:1937-1947.

8. Simon, R., Radmacher, M.D., and Dobbin, K. 2002. Design of studies using DNA microarrays. Genetic Epidemiology 23:21-36.

9. Simon, R., Radmacher, M.D., Dobbin, K., and McShane, L.M. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95:14-18.

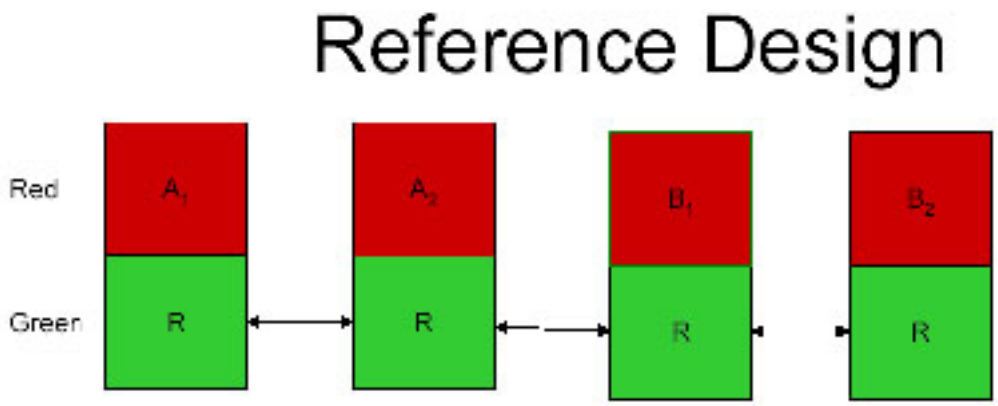

Figure 1

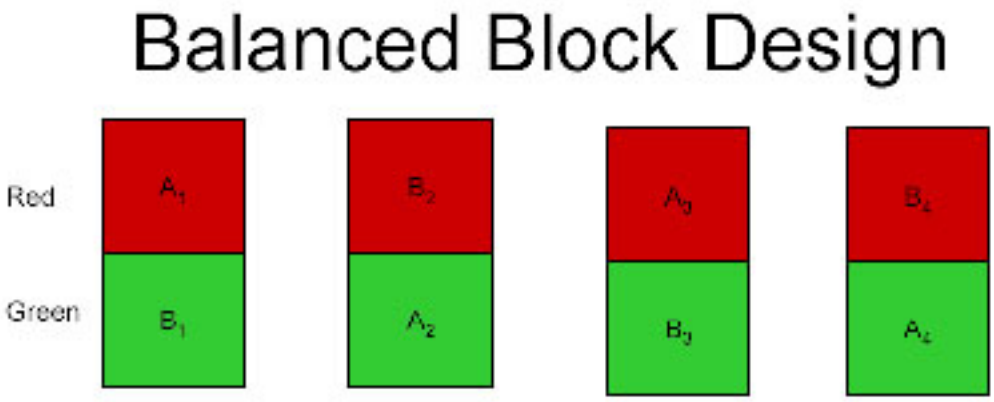

Figure 2

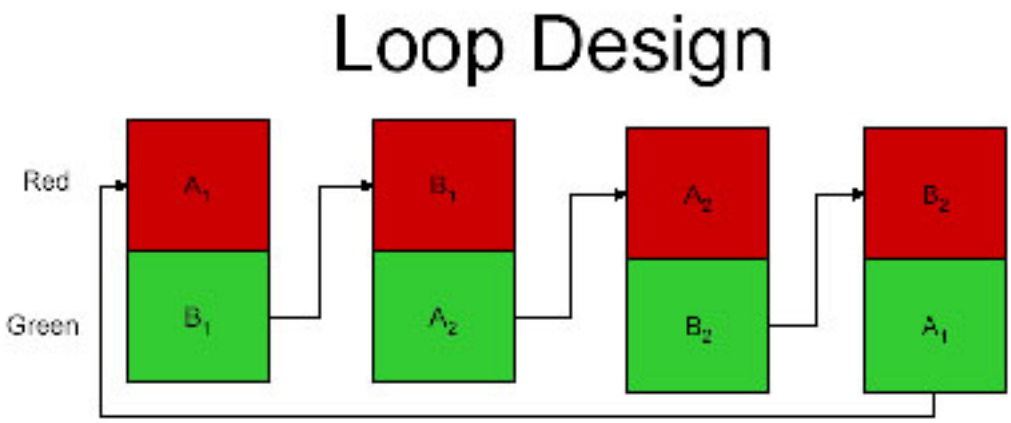

Figure 3