Experimental design

Kevin K. Dobbin and Richard M. Simon

Biometric Research Branch

National Cancer Institute

National Institutes of Health

Department of Health and Human Services

Bethesda, Maryland, U.S.A.

Keywords: (5 to 10 keywords) microarrays, experimental design, sample size, dye bias, reference sample

Abstract:

Experimental design issues for dual-label and single-label microarray experiments are discussed, including identification of research objectives, avoidance of confounding, allotment of samples to arrays and dyes in dual-label experiments, dye bias, pooling RNA before labeling, and determination of the number of arrays required to achieve the study objectives.

Introduction

Experimental design issues for dual-label and single-label microarray experiments are discussed, including identification of research objectives, avoidance of confounding, allotment of samples to arrays and dyes in dual-label experiments, dye bias, pooling RNA before labeling, and determination of the number of arrays required to achieve the study objectives.

1. Objectives

The first step in designing a microarray experiment is to identify the goals of the experiment. There is no one design that will be appropriate for every experiment, but the optimal design for a particular experiment will depend on the research questions being addressed. As the size and scope of microarray studies grow, so do the range of questions that researchers are asking. It is not possible to be comprehensive here in discussing study objectives, but it is useful to identify a few general types of objectives often seen in microarray research.

Class comparison objectives apply to studies of collections of specimens which come from two or more pre-defined classes, types or conditions. The defining aspect is that the classes are known ahead-of-time, independent of the gene expression data, and that the research question is: which genes are expressed differently in the different classes? For example, Hedenfalk et al. (2001) compared primary tumors from two classes of women,

those who carried the BRCA1 mutation genotype and those who carried the BRCA2 mutation genotype. One goal of the study was a class comparison goal, namely, to discover a set of genes that were expressed differently in the two genotypes. Other examples of studies with class comparison objectives include: 1) Golub et al. (1999) identified genes expressed differentially between specimens of acute myelogenous leukemia and specimens of acute lymphocytic leukemia; 2) Ross et al. (2000) compared expression profiles of cancer cell lines from different tissues of origin. In these studies, class comparison was often not the only goal, but there were other goals as well, such as class prediction.

Class prediction objectives can apply either to studies of collections of specimens from pre-defined classes, or to studies in which some clinical outcome is measured for each individual from which a specimen was obtained. The defining aspect is that the research question is: can we construct a prediction rule for the specimens that will predict, from the gene expression data alone, the likely class or outcome for this individual? The ultimate goal is usually to develop a rule that can be used on future individuals for whom the classification or outcome data is not available. The Golub et al. (1999) paper also provides an example of a class prediction objective; the differentially expressed genes discovered in the class comparison phase were used to develop a multi-gene class predictor that distinguished between acute myelogeous leukemia and acute lymphocytic leukemia, and the predictor was applied to an independent set of tumors to assess the predictive performance. Rosenwald et al. (2002) developed a molecular predictor of

survival after chemotherapy based on biopsy samples from diffuse large-B-cell lymphoma patients.

Class discovery objectives can apply either to studies in which there are no pre-defined classes for the specimens, or in which the current classification system is deemed inadequate. The defining aspect is that the research question is: can we discover a new classification system for these specimens based solely on gene expression data? Bittner et al. (2000) applied cluster analysis techniques to the gene expression profiles of a collection of melanoma samples to identify a novel classification system for this otherwise homogeneous group of specimens, suggesting a gene expression based taxonomy.

Another type of class discovery study addresses the research question: can we use what we know about the function of some genes to figure out the function of other genes? Or, what genes are co-regulated in these samples? These questions are typically approached by applying cluster analysis to the genes instead of the samples.

2. Confounding

In an ideal microarray experiment, tissue handling and cutting, RNA extraction, reverse-transcription and labeling, and array hybridization would all be performed at about the same time, under identical experimental conditions. For such an experiment in a class comparison setting, observed significant differences between the classes in gene expression could safely be attributed to real biological differences in the classes. But it is common that this type of ideal experiment is not possible, that not all microarrays used will be analyzed at the same time, that the microarray chips used may not all be uniform, that reagents may vary over the course of the experiment, etc.

If the microarray assay conditions vary, then it is still possible to construct a well designed experiment by ensuring that the assay conditions are not confounded with the goals of the experiment. For instance, suppose one has two classes of specimens one wishes to compare, that there are 10 specimens from each class, and that there are also 10 microarrays from each of two different chip versions that are to be used. An experiment that assigns the samples from one class to one chip version and the samples from the other class to the other chip version would be poorly designed because one would not know whether to attribute observed differences to chip versions or to the classes. Chip version and classes would be completely confounded, that is, their effects would be inextricably mixed together. A well designed experiment in this situation would be one in which each class is assigned five chips from each version. Then observed differences between the classes could not be attributed to the chip version, and one could separate out the effects of the different chip versions from the biological differences between the classes.

3. Dual-label microarray experiment designs

Dual-label systems present design issues not encountered by single-label systems. Each array in a dual-label system has two channels, one corresponding to each dye label. The motivation for using two labels instead of one is that it allows one to eliminate spot-to-spot variation – due to quality, size and location of spots – from comparisons of interest. The result is much greater power in class comparison experiments and greater ability to identify true clusters in class discovery than if a single dye was used. The resulting data structure, with large variation between measurements from different spots and small variation between the two measurements on the same spot, is called a block structure by statisticians, with the spots on the microarray serving as the blocking factor whose variability is blocked out of the comparisons of interest.

We will focus here on two types of designs for dual-label experiments, reference designs and balanced block designs, because for most experiments where the goal is class comparison, class prediction, and/or class discovery, one of these two designs will be optimal. Variations on these designs, such as dye swaps and technical replicates, will be discussed below. Other designs have been proposed, such as all-pairs designs (Yang and Speed, 2002), and loop designs (Kerr and Churchill, 2001).

The reference design

Reference design dual-label microarray experiments utilize multiple aliquots from a single RNA source, called the reference, which are applied to each microarray and are usually all labeled with the same dye. The role of the reference is sometimes not clearly understood. The size and location of a spot on a microarray is a major source of variability in these types of experiments. The role of the reference is to provide an estimate of this "spot" effect for every spot on every array. Consider the situation for a single gene. In all of the reference aliquots, this gene has the same expression level; therefore, differences in the expression level for this gene on different arrays can be attributed to the combination of spot size, quality, and location effects that together make up the spot-to-spot variation. Hence, for this gene, the spot-to-spot variation is well represented by the variation in the expression of the reference sample over the spots (assuming the gene is expressed at a sufficiently high level in the reference sample). This allows one to correct for this source of variability when comparing the samples in the non-reference channels. Without the reference, it would not be possible to correct for this source of variation, and the noise would effectively drown out much of the effects of interest for which one is looking.

The reference sample does not have to have a biologically meaningful interpretation, but it should be selected so that in general the genes expressed in the non-reference samples are also expressed in the reference sample. In order to get a good correction for the spot-to-spot variation, one needs some gene expression in the reference sample for that gene,

otherwise there will be a universally low gene expression in the reference channel regardless of differences in the size, quality and location of the spots, and the reference channel will not reflect the "spot" effects well. Therefore, genes with low or no expression in the reference channel will provide relatively noisy comparisons between the non-reference samples.

The balanced block design

The reference design may appear ideal because of the simple and intuitive way in which it allows one to estimate the spot-to-spot variation and eliminate it from comparisons of interest. But, in fact, spot-to-spot variation can also be estimated and eliminated from other types of designs. Although the estimates in these other types of designs may not be as intuitive as in the reference design, they are equally valid and can in fact result in considerable improvement in efficiency for class comparison experiments. The most efficient designs for class comparisons are balanced block designs. Other types of designs that have been proposed in the literature are not as efficient (Dobbin and Simon, 2002).

A balanced block design does not use a reference sample. Instead, a single aliquot from each biological sample (e.g., from each person or each mouse) is taken, and the samples are arranged so that samples from any two classes are paired together the same number of times over the microarrays. In the case of just two classes, this means that a sample from each class is paired together on each array. As discussed in the dye bias section below,

half the samples from each class should be tagged with Cy3 dye, and the other half with Cy5 dye.

Examples of a reference design and a balanced block design are given in Tables 1 and 2.

When to use a reference design and when to use a balanced block design

The reference design is the most commonly used design in dual-label microarray experiments. There are several advantages to the reference design that make it particularly appealing to scientists who may not have access to a trained biostatistician or bioinformatician to consult with on the design or analysis of their experiment: 1) Some microarray data analysis software packages assume that a reference design was used, and analyzing data from a different type of design with these packages may not be straightforward; 2) Reference designs do not require that the investigator stipulate ahead-of-time what particular comparisons are going to be of primary interest, and hence allows for greater flexibility than designs (such as the balanced block design) which do require the classes to be identified ahead of time and which may also lock the investigator into one particular type of comparison to the exclusion of other possible comparisons (for instance, one may be locked into a comparison of different tumor grades when the real interesting comparison turns out to be different tumor stages – but it may not even be possible to make this comparison after a balanced block experiment is run with grade as the comparison of interest); 3) If samples are analyzed at different times, then there may

be no easy way to adjust for these differences unless one has the reference sample to serve as a baseline for the comparisons.

If class discovery is a goal of the experiment, and the samples are to be analyzed using cluster analysis, then the reference design has been shown to be superior to other designs that have been proposed (Dobbin and Simon, 2002). The reason for this is that effective cluster analysis depends on having good estimates of the distance between every pair of samples. In a reference design, the distance between any two samples is measured with the same efficiency because that distance only involves measurement error related to two arrays, corresponding to the arrays for each of the samples; this is because the repeated reference sample on each array can be used to "connect" any two arrays. Other designs have less direct "connections" between the samples that involve more arrays and hence more measurement error. Because some of the distances in these alternative designs will be measured quite inefficiently and so will be very poorly estimated, a cluster analysis algorithm will have a hard time picking up the structure in the data, as we have shown in simulations (Dobbin and Simon, 2002). Importantly, a design which does not have any samples with multiple aliquots from that sample repeated on multiple arrays, such as the balanced block design, will have very poor class discovery performance because the spots, i.e., the sources of greatest noise in these experiments, are confounded with the individual sample effects, completely obscuring the true distances between samples on different arrays.

Things are less straightforward if class comparison is the major goal of the experiment. In this case, the best design depends on what the limiting factor is in the experiment. Arrays are the limiting factor of the experiment if one has plenty of samples (or could produce plenty of samples), but due to expense or other logistics one can only run a fixed number of arrays. On the other hand, samples are the limiting factor if one only has access to a fixed number of samples, and array expense or logistics is a relatively minor concern – one just wants to measure these samples as well as possible.

For a class comparison experiment in which arrays are the limiting factor, a balanced block design can be significantly more efficient than a reference design. This means that the resulting list of differentially expressed genes will potentially have far fewer "false positives" – genes that are not truly differentially expressed – and will be missing far fewer "false negatives" – genes that truly are differentially expressed but which don't show up on the gene list because the difference is drowned out by experimental error variation. Table 3 shows the relative efficiencies for a balanced block compared to a reference design. For two classes, the relative efficiency is 2.4, indicating that it would require more than twice as many arrays with a reference design to achieve the same efficiency as a balanced block design.

For a class comparison experiment in which the samples are the limiting factor, the efficiency differences between a balanced block and reference design are much less dramatic (see Table 3, second row). This is partly because the reference design uses twice as many arrays, and hence it will be more expensive and labor-intensive than the

balanced block design. With three or more classes, the reference design is more efficient and seems preferable overall. With two classes, the reference design is slightly less efficient, but this may not offset some of the advantages described in the first paragraph of this subsection.

4. Dye bias and the use of dye swap designs

Several definitions of dye bias have been used in the literature: 1) The tendency of one dye to appear brighter overall across genes (although this should be removed by proper normalization); 2) The tendency of spots with different overall intensities to display different relative efficiencies of the two dyes (which can be adjusted for using intensity dependent normalization such as lowess); 3) Bias for a particular subset of genes caused by an interaction between the sequence of a gene and the dye being incorporated. This creates dye bias for certain genes, but the dye bias for a gene is the same in all the samples; 4) Bias caused by an interaction between genes and the dyes that is different for different samples. We will restrict attention to definition 3), and we refer to this type of dye bias as gene-specific dye bias to distinguish it from 1) and 2). Bias of type 4), which could be called gene-and-sample-specific dye bias, could not be removed statistically from the comparisons of the different samples, so we will not discuss this type of dye bias here. Several authors have investigated gene-specific dye bias and found that it does seem to exist, but generally tends to be small in quantity (Dobbin et al., 2003a; Tseng et al., 2001).

For class comparison experiments, gene-specific dye bias will not affect comparisons between classes of samples labeled with the same dye. Hence gene-specific dye bias will not affect the identification of differentially expressed genes in single-label experiments or in dual-label reference design experiments (for comparison of classes of non-reference samples). Gene-specific dye bias may affect cluster analyses in class discovery experiments in either single-label or dual-label platforms with a correlation metric because intensity differences between genes may be influenced by sequence dependence of dye incorporation efficiency.

For class comparison experiments using a balanced block design, gene-specific dye bias can be eliminated from the class comparisons in dual-label systems by labeling half (or nearly half, if odd number) of the samples from every class with each dye, and adding a dye bias term to the model. There is no need to dye swap individual arrays (i.e., run the same two samples with the labeling reversed) to eliminate the dye bias. In fact, dye swapping individual arrays will result in a loss of efficiency compared to running new arrays with different samples (Dobbin et al., 2003a).

Finally, if, in a dual-label system, comparisons between the reference sample and the non-reference samples are of interest, then dye bias adjustment can best be made by running dye swaps on some, but not all, arrays (Dobbin et al., 2003a).

5. Pooling samples

In some situations, there is not enough RNA available from individual specimens to run the microarray assay. This problem can be overcome either by RNA amplification, or by pooling different RNA samples together until there is enough RNA for an array. Pooling samples in this way can be a viable alternative to RNA amplification. In order to perform statistical inference in a class comparison setting, it is necessary to construct several independent pools from each class or condition. Two pools are independent if there is no overlap in the sources from which the pools are constructed, e.g., if RNA from three mice is used to form each pool, then no two pools have a mouse in common.

Sometimes the motivation for pooling is not that there is inadequate RNA from individual samples to run the microarrays, but that by pooling it is hoped that the cost of the experiment will be reduced because fewer microarrays are required. For instance, an experiment with 12 samples from each of two classes would require 24 single-label microarrays; by pooling pairs of RNA samples together, one can reduce the number of arrays required to 12, 6 for each class. The pooled samples will also show less variation because, by pooling samples, the biological variation is reduced. But the improvement in power usually associated with such a reduction in variance is offset somewhat because power is also related to degrees of freedom for error, and the degrees of freedom are reduced from 22 in the 24 array experiment to 10 in the 12 array experiment. In fact, in order to get the same power as in the 24 array experiment, one will need to use more samples in the pooled experiment. This results in a tradeoff between the cost of the microarrays and the cost of sample acquisition which is displayed in Table 4. In general,

unless the samples are very available and inexpensive relative to the microarrays, pooling does not appear to be an effective way to reduce cost (Dobbin and Simon, 2003; McShane et al., 2003).

6. Sample size

Here we will focus on sample size determination for class comparison experiments. Sample size for prognostic studies was treated in Simon et al. (2002), and sample size calculations for class discovery or class prediction remain open research questions.

In class comparison problems, it is common to cycle through, gene by gene, to determine which genes are differentially expressed (although in small studies gene variance estimates may borrow information across genes (Wright and Simon, 2003)). Although multivariate permutation tests can be more effective (Simon et al., 2004), it is reasonable to power the study based on multiple univariate analyses. Assuming decisions about differential expression for individual genes are based on t-test statistics, a sample size formula for a two-class class comparison experiment is

$$Min\left\{ n : n \geq 4\sigma^2 \frac{\left(t_{n-2,\alpha/2} + t_{n-2,\beta}\right)^2}{\delta^2} \right\}$$, where Min indicates the smallest positive integer n satisfying the equation, which is found iteratively. This formula can be used either for a reference design dual-label experiment or a single-label experiment. n is the total number of arrays required, with $n/2$ for each class. σ^2 is the variance of the base 2 log-

ratios for the dual-label reference design, or the variance of the base 2 log-intensities for the single-label experiment. $t_{n-2,\alpha/2}$ and $t_{n-2,\beta}$ are the $\alpha/2$ th percentile and the β th percentile of the t distribution with $n-2$ degrees of freedom, respectively. α is the significance level of the test, and $1-\beta$ is the power to detect a difference of size δ in the class means on the base 2 log scale. Selection of an appropriate variance σ^2 is somewhat problematic because different genes are usually assumed to have different error variances, but reasonable estimates can be derived using prior data from a similar experiment (Yang and Speed, 2002).

For $n \geq 60$, the sample size formula can be based on the simpler standard normal approximation $n = 4\sigma^2 \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\delta^2}$ where $z_{\alpha/2}$ and z_{β} are the $\alpha/2$ th and β th percentiles of the standard normal distribution, respectively.

Sample size formulas for balanced block designs, and for experiments with technical replicates, have been presented elsewhere (Dobbin and Simon, in press).

References

Dobbin K., Shih, J.H., and Simon R. (2003a) Statistical design of reverse dye microarrays. *Bioinformatics*, **19**, 803-810.

Dobbin K., Shih, J.H., and Simon, R. (2003b) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *Journal of the National Cancer Institute U.S.A.,* **95**, 1362-1369.

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

Dobbin, K., and Simon, R. (in press) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics.*

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Cligiuri, M.A., Bloomfield, C.D., Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression profiling. *Science*, **286**, 531-537.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O., Borg, A. and Trent, J. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**, 539-548.

Kerr, M.K. and Churchill, G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183-201.

Mcshane, L.M., Shih, J.H. and Michalowska, A.M. (2003) Statistical issues in the design and analysis of gene expression microarray studies of animal models. *Journal of Mammary Bland Biology and Neoplasia*, **8**, 359-374.

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

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W. and Zhao, Y. (2004) *Design and analysis of DNA microarray investigations.* Springer-Verlag, New York.

Tseng, G.C., Oh, M., Rohlin, L., Liao, J., and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549-2557.

Wright, G.W. and Simon, R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448-2455.

Yang, Y.H. and Speed, T. (2002) Design isssues for cDNA microarray experiments. *Nature Reviews: Genetics*, **3**, 579-588.

Tables and captions

Array	1	2	3	4	5	6	7	8	9	10
Cy3	R	R	R	R	R	R	R	R	R	R
Cy5	A	B	C	D	E	A	B	C	D	E

Table 1: Reference design example. R is the reference sample. A, B, C, D, E are the five different classes being compared. There are two different samples from each variety.

Array	1	2	3	4	5	6	7	8	9	10
Cy3	A	C	A	E	B	D	B	C	E	D
Cy5	B	A	D	A	C	B	E	D	C	E

Table 2: Balanced block design example. A, B, C, D, E are the five different classes being compared. There are four different samples from each variety.

Relative Efficiencies		**Number of Varieties**	
		2	**3**
Limiting Factor	**Same number of arrays used**	2.4	1.8
	Same non-reference samples used	1.2	0.9

Table 3: Relative efficiencies of balanced block and reference designs. Variance ratio set to 4. Efficiency of block design divided by efficiency of reference design, so that a relative efficiency over 1 indicates block design more efficient.

Number of samples pooled on each array	**Number of arrays required**	**Number of samples required**
1	25	25
2	17	34
3	14	42
4	13	52

Table 4: Number of arrays and samples required for various pooling levels. An independent pool is constructed for each array, so that no sample is represented on more than one array. Settings were same as Table 1: $\alpha = .001$, $\beta = .05$, $\delta = 1$, and $\tau_g^2 + 2\sigma_g^2 = .25$. τ_g^2 is the biological variance within a class, and σ_g^2 is the measurement error variance. Variance ratio is $\tau_g^2 / \sigma_g^2 = 4$.