COMMENTARY

Questions and Answers on Design of Dual-Label Microarrays for Identifying Differentially Expressed Genes

Kevin Dobbin, Joanna H. Shih, Richard Simon

The rapid growth in the use of microarrays has generated many questions about how to design experiments that use this technology effectively. Investigators need answers to questions about RNA sample selection, allocation of samples to arrays, robustness of design, dye bias, sample size, and statistical power to ensure that the experimental objectives are achieved. We address some common questions that arise in designing dual-label microarray experiments and provide statistical answers to these questions, focusing specifically on how to select optimal designs for the identification of differentially expressed genes.

BACKGROUND

The dual-label microarray measures the expression level of thousands of genes for a sample of cells. A common goal of microarray experiments is to determine which genes are differentially expressed among two or more predefined classes of biologic specimens. These types of study goals are referred to as "class comparisons" *(1).* Some examples of class comparisons are 1) identifying the differentially expressed genes in BRCA1 mutation–positive, BRCA1 mutation–negative, and sporadic cases of primary breast cancer *(2);* 2) identifying the differentially expressed genes in colon cancer cells treated with high versus low doses of camptothecin *(3);* and 3) identifying the differentially expressed genes in the prostate cancer cell line LNCaP before and after treatment with the tumor growth inhibitor, PC-SPES *(4).* Because of their widespread use, class comparison experiments will be the focus of this commentary.

A microarray generally consists of either cDNA or externally synthesized oligonucleotides that are printed or coated on glass slides. A dual-label microarray uses competitive hybridization in which nucleic acids (i.e., cDNA, cRNA, or RNA) derived from two RNA sources are hybridized to the same microarray *(5,6).* The cDNA from one source is labeled with green (Cy3) dye, and the cDNA from the other source is labeled with red (Cy5) dye, either directly or indirectly *(7).* The cDNA or oligonucleotides representing different genes are immobilized on the glass slide and are often referred to as spots. For each spot there are two corresponding measurements, one for each dye, often referred to as the two channels. The advantages of competitive hybridization for cDNA experiments have been well established *(8).* The relative intensities of two labeled specimens measured at a single spot are much less variable than the relative intensities if measured at corresponding spots on different arrays. Relative expression measurements provide a means of controlling the variability in the size and shape of corresponding spots and the effects of variation in sample concentration on the surface of the array.

The relative expression measurements compare the expression levels of labeled cDNA that have originated from two different RNA sources. cDNA from a single source is often applied to every microarray slide and is labeled with the same dye (either Cy3 or Cy5). These labeled cDNAs are referred to as the reference. If the reference is labeled with Cy3 dye, then the nonreference samples are all labeled with Cy5 dye. Comparisons between the nonreference samples are based on log-ratios of the intensity of the Cy5 dye to the intensity of the Cy3 dye for corresponding spots on different arrays. Basing comparisons between the nonreference samples on the log-ratios eliminates the sources of variability attributable to aspects of the spot that affect both channels similarly. The gene expression data from such a design, called a reference design, is easy to analyze because simple *t* tests or similar statistical methods can be applied directly to the log-ratios, and there is much existing software available for performing such tests. In addition, it is also possible to control for spot variability from designs that do not use a reference by statistical modeling. Hence, the reference design may or may not be the best choice for a particular situation.

The ability to measure expression levels for two samples on each cDNA array permits a number of design options for assigning specimens to labels and arrays. When choosing among these design options, one should consider the objectives of the experiment, the sources of variability, and the differences between dyes with regard to labeling and detection characteristics. The purpose of this commentary is to provide statistically sound advice about the design of investigations for finding differentially expressed genes using dual-label microarray platforms. We present a number of results comparing the statistical properties of different designs that we have established elsewhere. However, to keep the presentation nonmathematical, we have replaced equations presented in our earlier articles *(9,10)* with graphical displays where appropriate.

Affiliation of authors: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD.

Correspondence to: Kevin Dobbin, PhD, National Cancer Institute, 6130 Executive Blvd., EPN 8124, Bethesda, MD 20892–7434 (e-mail: dobbinke@mail.nih.gov).

See "Notes" following "References."

DOI: 10.1093/jnci/djg049

SAMPLE SELECTION

Is It Sufficient to Sample One Individual From Each Class?

The answer is no, because the goals of class comparison are to determine whether the gene expression profiles are different between the classes and to identify differentially expressed genes. Different individuals in the same class are not expected to have exactly the same gene expression level measurements. Biologic variation and measurement error will produce some differences in the gene expression profiles. If we sample only one individual from each class, then there is no way to distinguish expression differences associated with class from those associated with biologic variation or measurement error. Some genes may vary widely in their expression level from individual to individual in the same class, whereas others may display differential expression that is relatively small but is nonetheless critical for class distinction. Therefore, it is important to have multiple (and distinct) individuals from each class to obtain an estimate of biologic variation. Similarly, in studying gene expression in model organisms under different biologic conditions, it is important to have distinct applications of the conditions and harvesting of cells.

How Many Replicates of Each RNA Sample Should Be Hybridized?

Some investigators *(11)* have promoted using three or more replicate measurements for each RNA sample, and others *(12)* have suggested that at least two replicate measurements are required for each sample. These guidelines may be correct in some situations; however, they will probably not be correct for class comparison problems. When one is interested in class comparison, then replication measurements should generally be at the population level, so that each replicate represents RNA from a different individual. Intuitively, the reason that this level of replication produces the best comparisons is that, by replicating at the population level, one simultaneously reduces variability from both population heterogeneity and experimental error. When multiple aliquots are replicated from the same RNA source, one only reduces variability from experimental error. Therefore, replication of individual samples is inefficient for class comparisons.

Hybridization replicates increase the accuracy of the individual sample measurements *(11)*. However, if the number of arrays is fixed (e.g., when one only has time or resources available to run a prespecified number of arrays), then increasing the hybridization replicates requires decreasing the number of distinct RNA samples assayed. The result of this approach is a reduction in the accuracy of the class mean estimates. The relationship between sample measurement accuracy and class mean estimate accuracy as the number of hybridization replicates per sample increases for an experiment with a fixed number of 24 arrays is shown in Fig. 1. (*see* supplemental information at http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/index.shtml for details and proof). Accuracy is defined as the inverse of the variance of the mean estimate. Population parameter estimates are most accurate when hybridization replication (i.e., subsampling) is avoided, even though the accuracy of individual sample estimates is at a minimum when there is no subsampling. With less subsampling, one is better able to detect differentially expressed genes in the classes when the total number of arrays is fixed. An obvious exception to this rule is when only a limited number of valuable RNA samples are available and when one does not have access to more. Assaying each sample multiple times will clearly be preferable to assaying each sample only once.

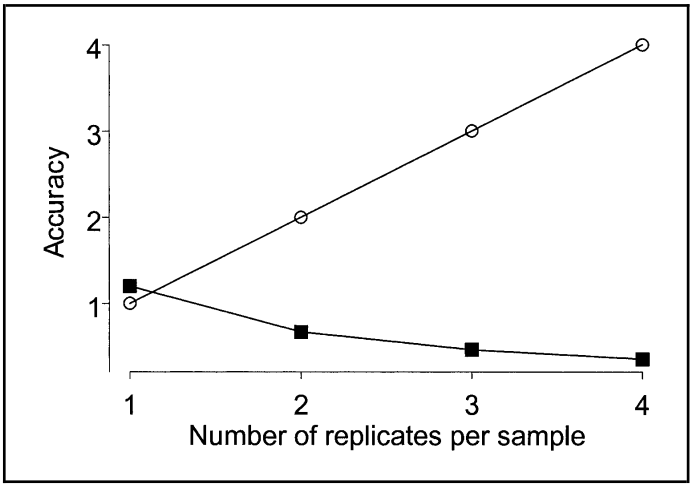

Fig. 1. Accuracy of sample and class mean estimates as a function of the number of replicates per sample. The number of arrays was fixed at 24. Accuracy is defined as the inverse of the variance of the estimate. The estimates are the difference in the class averages for class mean estimates (**solid squares**) and the average of repeated measurements on the same sample for sample estimates (**open circles**). Some parameters, such as the biologic and experimental variation, were fixed to construct the display. For further details about mathematical equations, refer to online supplemental information (*see* http//jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/index.shtml).

One might think that replicate hybridizations would help offset high measurement variability in low-quality microarray experiments that display high variation in repeated assays on the same sample. The power to detect a differentially expressed gene as a function of the number of subsamples per sample used, for example, of both a high-quality (i.e., displays low variation in repeated assays on the same sample) and a low-quality experiment, is shown in Fig. 2 (*see* supplemental information at http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/index.shtml for details and proof). The high-quality experiment is assumed to have an experimental error variance of half the biologic variance, and the low-quality experiment is assumed to have an experimental error variance twice that of the biologic variance. Although the loss of power is more dramatic for the high-quality experiment than for the low-quality experiment, the low-quality experiment also loses power when one replicates hybridizations for a fixed number of arrays.

What Are the Advantages and Disadvantages of Pooling Samples?

Pooling samples involves mixing together RNA from several sources before labeling and hybridization. Two motivations for pooling samples are 1) not enough RNA available from each individual to perform the assay, and 2) wanting to reduce the number of arrays used. Investigators sometimes hope to cut down on the number of arrays needed by comparing a single pooled sample from each class. The reasoning behind this approach is that the concentration of an mRNA molecule in a pooled sample is likely to be closer to the average concentration for the class than the concentration in a sample from a single

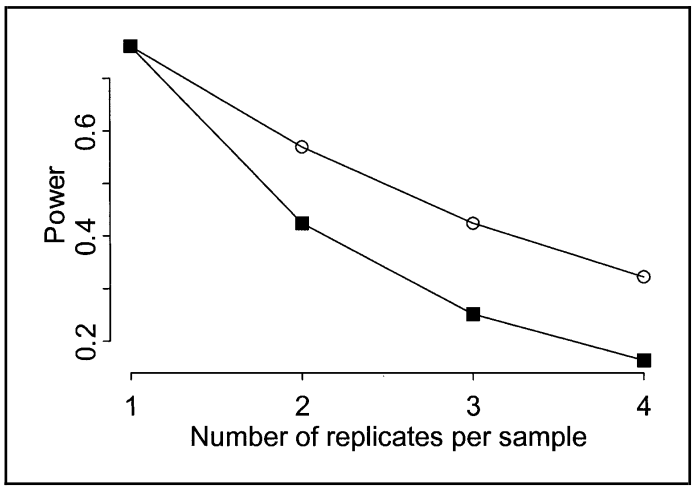

Fig. 2. Statistical power to detect differentially expressed genes as a function of the number of replicates per sample. The number of arrays was fixed at 24. The high-quality (i.e., displays low variation in repeated assays on the same sample) experiment (**solid squares**) has experimental error variance half that of the biologic variance. The low-quality (i.e., displays high variation in repeated assays on the same sample) experiment (**open circles**) has experimental error variance twice that of the biologic variance. The power is the probability of detecting a twofold change in gene expression levels for the high-quality experiment and a $2\sqrt{2}$-fold change in gene expression levels for the low-quality experiment (i.e., to make the powers comparable). For further details about mathematical equations, refer to online supplemental information (*see* http//jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/index.shtml).

individual. Unfortunately, a single pooled sample from each class will not be adequate for statistical inference, because one has no estimate of the biologic or experimental variability in the gene expression levels for pooled samples constructed from samples of the same class. Taking multiple subsamples from each pool and repeating them on multiple microarrays does not solve this problem, because variation among the subsamples will reflect only measurement error and will not include biologic variation.

It is possible to perform valid statistical comparisons between the classes with pooled data, but this approach requires multiple pooled samples from each class. Different pools of RNA should be constructed from different sets of individuals so that the pooled samples are independent and represent true replication. Comparisons of gene expression levels between classes are then straightforward. However, there are still some disadvantages to this approach. 1) It does not allow one to understand the contribution of individual RNA samples to the observed gene expression levels, which makes it impossible to identify outlier or poor-quality RNA samples. 2) A pool average is potentially biased for the class average—that is, the average expression level of a gene in the pool may differ from the average of the expression levels of the gene in the contributing samples, which can happen because of inequalities in the amounts of RNA contributed by different samples or because mixing of the RNA causes unanticipated alteration of gene expression. 3) It may be difficult or impossible to understand how gene expression is distributed in the population from pooled data and, hence, to make valid statistical inferences or predictions for individuals. In summary, pooling of samples is recommended when there is not enough RNA from individual samples to run a microarray. The use of several independent pools from each class will allow for valid statistical inference about the classes.

Pairing Samples For Co-Hybridization

What Types of Designs Should Be Considered?

Three designs have been proposed for cDNA microarray class comparison experiments (Fig. 3). The reference design is by far the most widely used because spot-to-spot variation can be eliminated in a simple way by using ratios or log-ratios. There are many other advantages to the reference design, which are explored later in this section; however, its widespread use should not preclude consideration of other alternatives. The distinctive feature of a reference design is that expression of a gene for a sample is measured relative to the expression of that gene at the same spot on the same array for a reference sample.

The ability to co-hybridize two differentially labeled samples to each array may appear to open a Pandora's box of experimental design possibilities. However, do we really need to sort through every possible design? The fact that the difference in gene expression levels between corresponding spots on different microarrays is a major source of variability makes the arrays analogous to a blocking factor in agricultural experiments. There is extensive statistical literature on the design of such experiments *(13,14),* but it cannot be applied directly to dual-labeled microarray experiments, because the error structure for microarray data is somewhat different than the agricultural analog. We have adapted the method for deriving optimal designs in the

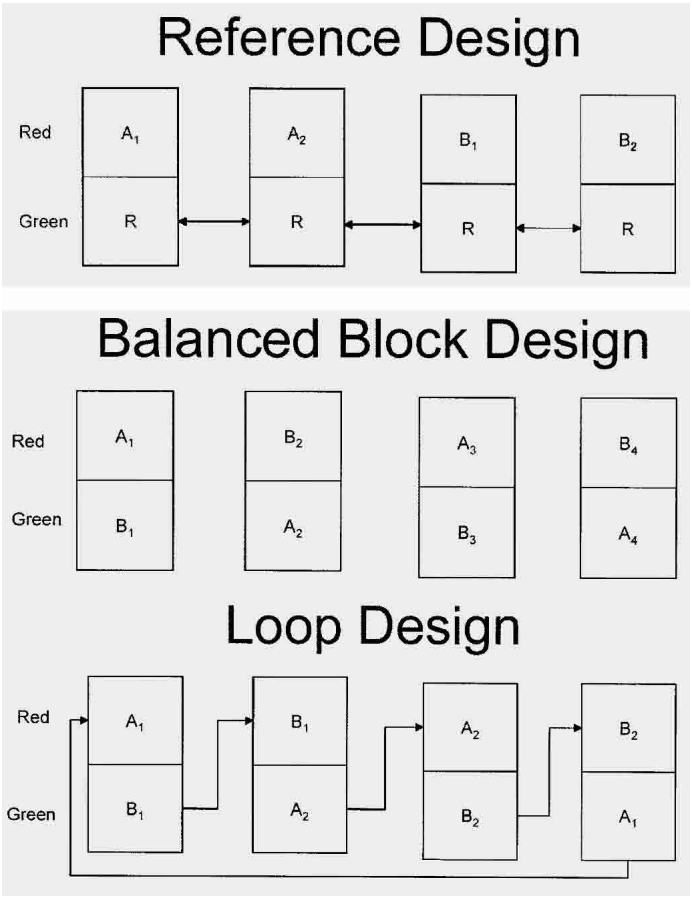

Fig. 3. Design diagrams for cDNA microarray class comparison experiments. **Rectangles** represent the arrays. A_1 is sample 1 from class A, B_1 is sample 1 from class B, A_2 is sample 2 from class A, and so on. R is the reference sample. **Arrows** connect samples repeated on multiple arrays. **Red** is the Cy5 dye and **Green** is the Cy3 dye used to label the reference and nonreference samples.

presence of a blocking factor to microarray experiments *(9)* and have established that, for many class comparison studies, the balanced block design shown in Fig. 3 is optimal. The effect of spot-to-spot variation in gene expression levels is eliminated in the balanced block design because each gene's expression level is measured at the same spot on the same array for samples from each of the two classes being compared.

The third type of design that might be considered for cDNA microarrays is one proposed by Kerr and Churchill *(15),* which they called a loop design (Fig. 3). Unlike the two other designs, the loop design requires two aliquots from each RNA sample. These aliquot pairs connect the arrays and are arranged so that the connected arrays form a loop pattern.

Class comparisons for the balanced block design and the loop design are accomplished by fitting an analysis-of-variance (ANOVA) model to the logarithm of the background-corrected channel-specific intensities *(9)* and fitting a separate model for each gene. This approach can also be used for analysis of the reference design, but the results are equal to or very similar to applying simple Student's *t* tests to the log-ratio measurements.

More elaborate designs have been proposed to achieve different experimental objectives *(15,16);* however, we will focus on the three types of designs presented in Fig. 3 because they are the most obvious choices for class comparisons. Other types of designs to consider are presented in the dye bias section.

Which Design Will Provide the Best Class Comparisons?

Balanced block, loop, and reference design experiments can all provide unbiased estimates of the differences in gene expression levels between class means, i.e., differences between the average gene expression levels. However, the three designs are not equally efficient. The efficiency of a design is based on the precision of the statistical estimates of the differences in the class means for "equivalent experiments." We define two notions of equivalent experiments that we think are appropriate to many microarray studies: 1) Two experiments are equivalent if they use the same number of microarrays, and 2) two experiments are equivalent if they use the same (nonreference) samples and subsamples.

Definition 1 is appropriate when nonreference RNA samples are abundant and the limiting factor is the amount of time or resources required to actually run the arrays. The question then might be "If I can afford to run only 20 arrays, how should I design the experiment?" Definition 2 is appropriate when the nonreference RNA resources are scarce and the cost of running the arrays is less critical. The question then might be "Given that I have only these 12 RNA samples, how should I design the experiment?"

Efficiency comparisons of the three designs for a typical experiment (with biologic variation twice that of the experimental error variation) calculated from equations presented in Dobbin and Simon *(9)* are shown in Fig. 4, A. When the number of microarrays is limited (equivalence definition 1), then the balanced block design is substantially more efficient than the reference or the loop designs. However, the efficiency gain with the balanced block design comes with some sacrifice, including robustness and difficulty in clustering samples.

When the nonreference RNA samples are limited (equivalence definition 2), then the efficiencies of the reference and balanced block design are similar (Fig. 4, B). The loop design is less efficient than the balanced block design and also suffers from the same lack of robustness. The more robust reference design appears to be better overall than the other two designs when nonreference RNA samples are limited.

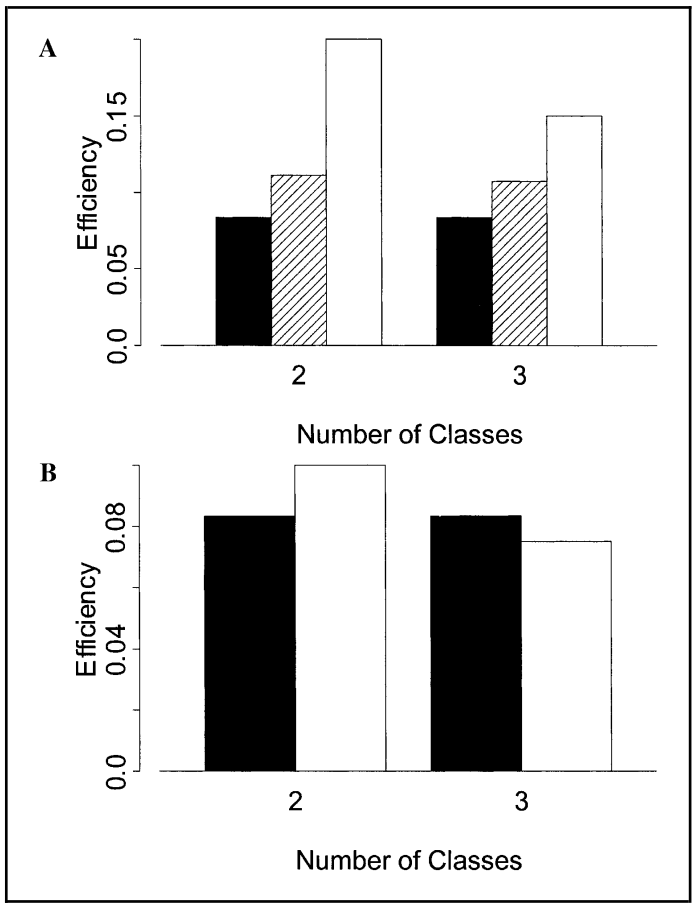

Fig. 4. Comparison of design efficiencies. **A**) Comparison of design efficiencies for the reference (**solid bars**), loop (**hatched bars**), and balanced block (**open bars**) designs when the number of arrays is fixed. **B**) Comparison of the reference (**solid bars**) and balanced block (**open bars**) designs when the nonreference RNA samples are fixed. Efficiency is the inverse of the variance of the estimated difference between the class averages. Some parameters, such as the biologic and experimental variation, were fixed to construct the display. Results are general in that the specific number of arrays or samples used does not affect the relationship between the heights of the histogram bars. The loop design was not included in the histogram because it uses a different sampling scheme. For further details about mathematical equations, refer to online supplemental information (*see* http//jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/index.shtml).

What Happens If the Class Definitions Change?

It is not unusual to have different classifications of the samples or to have corrections in the class of specific samples. The reference design is more robust to changes in the classification scheme than either the balanced block or loop designs. The reason for this increased robustness is that the reference design will remain a reference design with a new classification. In contrast, the balanced block design will probably lose its structure (i.e., it will no longer be a balanced block design). With regard to a new classification, many arrays may contain two samples from the same class, which can result in a severe loss of efficiency. It is also possible that, with a new classification, the classes cannot be compared with the balanced block design because they never appear together on any arrays. The loop design

is also subject to large efficiency loss, because under a new classification, the classes may appear together only on a small proportion of arrays.

What If We Also Plan to Perform Class Discovery on the Samples?

Class discovery is the process of finding a new classification system for a set of biologic samples on the basis of gene expression profiles when the class labels are unknown ahead of time. Cluster analysis is the most appropriate approach to use in class discovery. Of the three designs presented in Fig. 3, effective class discovery can only be performed for the reference and loop designs. Individual samples must be compared in class discovery. The balanced block design confounds spot variability with comparison of samples on different arrays because no RNA sample appears on more than one array. The arrows connecting the samples repeated on different slides in the reference and loop designs indicate why this type of confounding is not a problem in these designs—that is, connections can be made between any two samples on different arrays using the arrows.

The reference design is recommended for class discovery because cluster analysis can perform substantially better with a reference design than a loop design *(9),* particularly as the number of samples increases. An example of a cluster analysis for 10 and 20 samples, which was originally presented in Dobbin and Simon *(9),* is presented in Fig. 5. The data in that figure were generated from two true clusters (i.e., the data in each cluster were generated from a different mean gene expression vector). The number of discrepancies between the clusters found by a common cluster analysis algorithm and the true clusters for the reference and loop designs appear on the *x*-axis. The reference design finds the true clusters almost every time, whereas the loop design performs poorly for 10 samples and much worse for 20 samples. Moreover, the loop design performs even worse when there are more than 20 samples *(9).* The difference in cluster analysis performance is so dramatic that it will usually offset any relatively moderate differences in efficiency and power between the loop and reference designs. For this reason, we recommend using the reference design for class discovery experiments.

What Is Sacrificed If a Reference Design Is Not Used?

Most investigators are familiar with the reference design, and they may want to know what will be sacrificed if an alternative design such as the balanced block design is used. In addition to the issues discussed in the last two questions, there are other considerations worth mentioning. 1) The data from a balanced block or loop design may be more difficult to analyze than data from a reference design. Most microarray analysis packages assume a reference design has been used, so analyzing the experiment may require switching to different software. 2) The balanced block or loop design may be more difficult to devise than the reference design. If there are many groups being compared or many possible ways to group the samples, designing the study so that all comparisons of interest can be made may be nontrivial. 3) It may not be possible to compare data from different microarray experiments or prospective data that is analyzed by microarrays at different times. If a common reference sample is used for all experiments, then there is some foundation for the comparison of samples collected over time or samples analyzed in different experiments, a situation that is generally not possible for balanced block or loop designs.

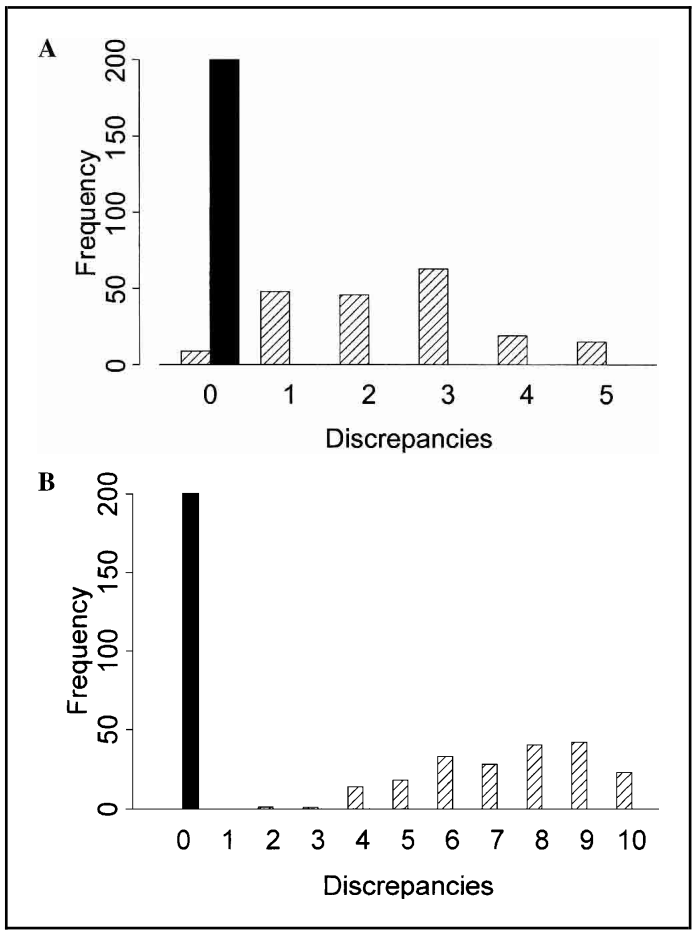

Fig. 5. Comparison of cluster analysis performance. Comparison of cluster analysis performance for the reference (**solid bars**) and loop (**hatched bars**) designs on **A**) 10 samples and **B**) 20 samples. Simulated data comes from two true (each with a different mean gene expression) clusters. One thousand genes were present in the clusters, 20 of which were differentially expressed. *x*-axis is the number of discrepancies between true clusters and closest matches. *y*-axis is the frequency of the number of discrepancies observed in 200 simulations. Simulation was based on a prostate cancer dataset [*see* Dobbin and Simon *(9)* for details].

Dye Bias

What Is the Source of Dye Bias?

Cy3 and Cy5 have different efficiencies for their labeling ability and detection characteristics. Background correction and normalization adjust for consistent dye-related differences that are not gene-specific. For example, median centering of arrays is meant to eliminate bias that is common across all genes, and intensity-dependent normalization, such as loess smoothers, adjust for bias related to overall spot intensity *(15).* Gene-specific dye bias is displayed by genes that do not fall into the overall pattern of the dye effect that characterizes the majority of genes. This bias may persist even after normalization.

Does Gene-Specific Dye Bias Exist?

To our knowledge, there has been no definitive study characterizing the nature or magnitude of gene-specific dye bias. In addition, it is not clear that gene-specific dye bias is the same

from one experiment to another or from one laboratory to another, but it is of general concern among microarray investigators. Many studies *(3,18–22)* have been designed to guard against gene-specific dye bias, whereas others *(8,17,23–25)* have made gene-specific dye bias adjustments to their statistical analysis. Some studies have attempted to eliminate gene-specific dye bias through technical innovations in labeling *(7,23,26,27)*. Although novel labeling procedures such as indirect labeling appear to reduce gene-specific dye bias, it is not clear that they eliminate dye bias.

We have observed gene-specific dye bias and provide, as an example, one reference design experiment involving transgenic mice (Green J: unpublished data). Nine distinct RNA samples from nine mice were examined, and three of these samples were run twice, once with each dye label (i.e., once with the reference labeled with Cy3 and once with the reference labeled with Cy5), for a total of 12 arrays. The intensity data were first background-adjusted to eliminate stray fluorescence signals from the slide and normalized to make the measurements on different arrays comparable. We then performed an ANOVA on the individual channel log intensities. An ANOVA model was fit separately to data for each of 8832 genes. In the ANOVA approach, the dye bias effects are called dye-by-gene interactions. Overall, we observed that there were many genes with a statistically significant dye-by-gene interaction ($P<.001$), but these effects tended to be small. The size of these effects on the base 2 log-scale is shown in Fig. 6. The average absolute value of the gene-by-dye interactions was 0.18 (standard deviation = 0.16), corresponding to a 1.13-fold change in gene expression levels. Only 10 of the 8832 genes had dye bias that corresponded to a twofold or greater change in gene expression levels. Tseng et al. *(8)* have presented similar results. Although dye bias appears to be common in these direct-labeled cDNA experiments, it appears to be fairly small in magnitude.

When Is Gene-Specific Dye Bias an Issue?

Gene-specific dye bias is a potential issue when comparisons are made between samples labeled with different dyes. Hence, it is not generally a problem in reference design experiments because they compare classes of nonreference RNA samples. Because all of the nonreference RNA samples are labeled with the same dye, the dye bias between the nonreference and reference intensities does not become a bias in comparing classes. Gene-specific dye bias is a potential problem, however, if nonreference RNA samples are compared with a common reference RNA sample. Gene-specific dye bias is also an issue for balanced block and loop designs. When gene-specific dye bias is an issue, its magnitude must be estimated for each gene, and an explicit adjustment to the statistical analysis must be made to ensure that class comparisons are unbiased. For example, in ANOVA analysis, the adjustment involves adding terms representing gene-specific dye bias to the statistical model.

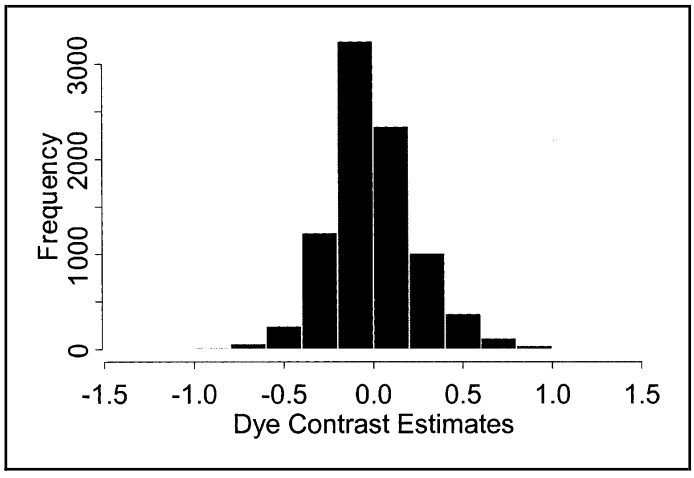

Fig. 6. Estimated dye bias contrast that was not corrected for in normalization. Estimates for dye bias were based on 8832 genes from a transgenic mouse experiment. Data were transformed to base 2 logarithms so that an estimated dye bias contrast of size 1 corresponds to a twofold change in gene expression.

How Should I Design an Experiment to Eliminate Dye Bias From the Class Comparisons?

Dye bias can be eliminated from the class comparisons in two ways: 1) by labeling all samples from all classes being compared with the same dye, and 2) by labeling half the samples with one dye and half the samples with the other dye for each class being compared.

Reference designs usually use strategy 1 to eliminate dye bias. Other designs, such as balanced block designs, often use strategy 2. Labeling exactly half the samples of a class with a dye is preferable to labeling some other fraction because it produces more accurate class comparisons and is simpler to analyze. If there is an odd number of nonreference RNA samples from each class (e.g., seven), then strategy 2 would not be able to be followed exactly (e.g., three samples labeled with red dye [Cy5] and four samples labeled with green dye [Cy3]). Dye bias can still be eliminated from such a design, but it requires a more complex weighted analysis to adjust for the dye asymmetry.

Another approach that is sometimes used to eliminate dye bias is to run a set of arrays with the reference in both channels to identify the genes that display dye bias. These genes could then be flagged as suspect if they show up as statistically significant in the class comparisons.

Some investigators *(12)* have used the existence of dye bias as a reason to run all sample pairs twice, once with each dye, to eliminate the bias. However, we *(10)* have shown that complete dye swapping is an inefficient way to adjust for the dye bias correction. If each sample is run twice in a fixed number of arrays, then the effective sample size is cut in half. The reference design or balanced block design will provide unbiased estimates of the class comparison without running any sample pairs twice. Hence, the complete dye-swapping strategy effectively halves the sample size and reduces the efficiency with no real gain as far as class comparisons are concerned. Balancing the classes with respect to the dyes is more efficient than dye swapping of individual samples for eliminating dye bias.

How Will Class Discovery Results Be Affected by Dye Bias?

Dye bias generally will not have a substantial impact on class discovery, although it may be necessary to make an explicit dye bias adjustment. In this commentary, we have focused on class comparison experiments in which we already have class labels for the samples. Class discovery can be performed on all the samples or on only the samples within a particular class. Class discovery using cluster analysis on all of the samples is sometimes performed to verify that the resulting clusters recapitulate the known classes *(28,29)*. In addition, cluster analysis within a

particular class is sometimes used to identify novel subclasses *(28,30,31).*

In the previous section we discussed two designs for class discovery—the reference design and the loop design. Dye bias generally will not affect class discovery for the reference design because all the samples being clustered are labeled with the same dye. The effect of dye bias on cluster analysis results can also be eliminated from the loop design by making a dye bias adjustment; however, we do not recommend this design because of its poor cluster analysis performance, as discussed in the previous section.

How Can Dye Bias Be Eliminated From Comparisons Between the Reference and the Nonreference Samples in a Reference Design?

One can eliminate dye bias from the comparisons between the reference and nonreference samples by including dye-swapping arrays in the design of the experiment. Consider a reference design experiment used to study a collection of tumor samples, where the reference sample consists of a mixture of normal tissue. A fairly common experimental situation is one in which the primary goal is to perform class discovery on the tumors and the secondary goal is to compare the tumors with the normal reference to identify potential tumor markers *(32,33).* Because the normal reference sample is labeled with a different dye than the tumor samples, there is potential for dye bias in the comparisons. In this case, we recommend appending the basic reference design with just enough dye-swapping arrays to allow for good statistical inference for the comparison with the reference sample. This comparison is made by ANOVA and is adjusted for dye bias; an example of such a design is shown in Fig. 7. Note, we do not recommend reversing all the arrays in this situation, because running all samples both forward and backward with the reference sample substantially reduces the efficiency of the tumor versus normal comparison (for a fixed number of arrays) and hinders the ability of the cluster analysis to identify true groupings in the gene expression data. Running dye-swapping arrays on all samples essentially sacrifices the primary goal of discovering a new taxonomy for the secondary goal of identifying potential markers; even for the secondary goal, complete dye swapping is inefficient in most cases.

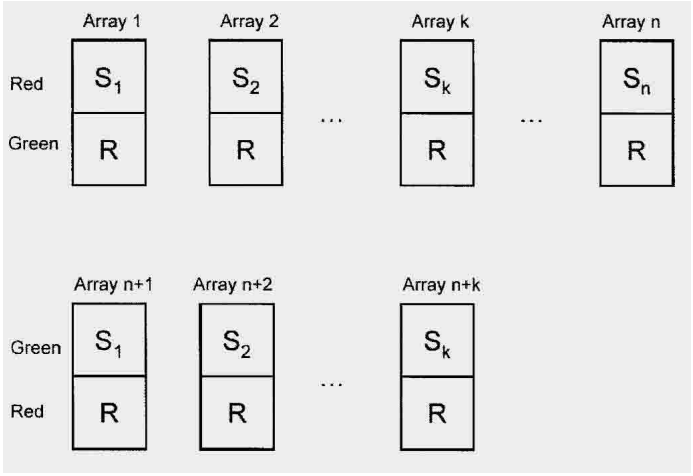

Fig. 7. Dye-swapping reference design for clustering and comparison of nonreference with reference RNA samples. **Rectangles** represent the arrays. S_1 is sample 1 from the nonreference samples, S_2 is sample 2 from the nonreference samples, and so on up to some numbered sample n (S_n). R is the reference sample. Of the n + k arrays, k is run as a dye swap on repeated samples. The **first row** of arrays represents the forward arrays and **second row** of arrays represents the reverse arrays. The reference sample is dyed green (Cy3) on the forward arrays and red (Cy5) on the reverse arrays. The resulting fixed-effects analysis of variance table has k –1 degrees of freedom for error.

Sample Size

How Many Biologic Samples Are Needed for a Reference Design?

Suppose we want to test whether a particular gene is differentially expressed in two classes. To test the null hypothesis that there is no difference in gene expression levels at the α significance level, we want to have $1 - \beta$ power to detect a difference of δ in the class mean log-ratios. Let σ be the standard deviation of the log-ratios within each class and n be the total number of arrays used, i.e., $n/2$ arrays for each class. Then the usual sample size formula *(34),* based on an assumption of normal distributions within the classes, would be:

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

The notation $z_{1-\alpha/2}$ indicates the $100(1 - \alpha/2)^{th}$ percentile of the standard normal distribution. When the samples sizes are small, the normal approximation of the test statistic may be poor, and an iterative computational procedure based on the t distribution can be used to compute the sample size. For example, we have observed an $\sigma \approx .50$ for human cancer data using log base 2 intensities on cDNA microarrays and a reference design, and we have observed $\sigma \approx .25$ with data from inbred strains on transgenic mice *(9).* A $\delta = 1$ corresponds to a twofold difference in gene expression. Setting $\alpha = .001$ guards against an excessive number of false-positive genes. For example, with 10 000 genes, $\alpha = .001$ results in an average of 10 false-positive genes. Setting $\beta = .05$ provides 95% probability of detecting a twofold change in gene expression. The resulting sample size is then 30 total samples for $\sigma = .50$ and 12 total samples for $\sigma = .25$. Because of the small sample sizes, we have used t distribution percentiles in both cases.

What Sample Size Should Be Used for a Balanced Block Design?

Suppose that two classes will be compared and that the samples from each class are independent. Again, we want to test the null hypothesis that there is no difference in gene expression levels between the classes at the α significance level and to have $1 - \beta$ power to detect a difference of δ in the class means. Let τ be the standard deviation of the log-ratios. In the balanced block design, each log-ratio involves two independent samples, one from each class. The τ parameter will tend to be larger than the σ parameter in the reference design because additional biologic variation is displayed in the log-ratios. Let n be the total number of arrays used, i.e., n arrays with n samples from each class. The sample size formula would now be:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\delta/\tau)^2}$$

Because the reference sample appears on each array in the reference design, variability among the log-ratios will be smaller for a reference design than for a balanced block design. We provide

on-line supplemental material (*see* http//jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/index.shtml for details) that shows how prior data from a reference design experiment can be used to estimate τ. For example, using our estimated standard deviation of the log-ratios from the reference design that used human samples ($\sigma = .50$) and the same parameter settings that we used for the reference design sample size calculation ($\delta = 1$, $\alpha = .001$, $\beta = .05$) results for $\tau \approx .67$, the sample size required changes from 30 arrays under the reference design to 17 arrays under the balanced block design. The reference design uses 30 arrays from 30 total samples, 15 from each class, whereas the balanced block design uses 13 fewer arrays but requires 17 samples from each class, or a total of 34 samples.

REFERENCES

(1) Golub T, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

(2) Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene-expression profiles in hereditary breast cancer. N Engl J Med 2001; 344:539–48.

(3) Zhou Y, Gwadry FG, Reinhold WC, Miller LD, Smith LH, Scherf U, et al. Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: microarray analysis of dose- and time-dependent effects. Cancer Res 2002;62:1688–95.

(4) Bonham MJ, Galkin A, Montgomery B, Stahl WL, Agus D, Nelson PS. Effects of the herbal extract PC-SPES on microtubule dynamics and paclitaxel-mediated prostate tumor growth inhibition. J Natl Cancer Inst 2002; 94:1641–7.

(5) Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999;21(1 Suppl):10–4.

(6) Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405:827–36.

(7) Manduchi E, Scearce LM, Brestelli JE, Grant GR, Kaestner KH, Stoeckert CJ Jr. Comparison of different labeling methods for two-channel high-density microarray experiments. Physiol Genomics 2002;10:169–79.

(8) Tseng GC, Oh M, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 2001;29:2549–57.

(9) Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 2002;18:1438–45.

(10) Dobbin K, Shih JH, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 2003;19:803–10.

(11) Lee MT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 2000;97: 9834–9.

(12) Goryachev AB, MacGregor PF, Edwards AM. Unfolding of microarray data. J Comput Biol 2001;8:443–61.

(13) Cochran WG, Cox GM. Experimental designs. 2nd ed. New York (NY): Wiley; 1992. p. 95–182.

(14) Scheffé H. The analysis of variance. New York (NY): Wiley; 1999. p. 55–146.

(15) Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2:183–201.

(16) Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002;3:579–88.

(17) Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res 2002;30:e15.

(18) Bayani J, Brenton JD, Macgregor PF, Beheshti B, Albert M, Nallainathan D, et al. Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. Cancer Res 2002;62:3466–76.

(19) Klebes A, Biehs B, Cifuentes F, Kornberg TB. Expression profiling of Drosophila imaginal discs. Genome Biol 2002;3:RESEARCH0038.

(20) Aharoni A, Keizer LC, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, et al. Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. Plant Cell 2000;12:647–62.

(21) Barrans JD, Allen PD, Stamatiou D, Dzau VJ, Liew C. Global gene expression profiling of end-stage dilated cardiomyopathy using a human cardiovascular-based cDNA microarray. Am J Pathol 2002;160:2035–43.

(22) Desai KV, Xiao N, Wang W, Gangi L, Greene J, Powell JI, et al. Initiating oncogenic event determines gene-expression patterns of human breast cancer models. Proc Natl Acad Sci U S A 2002;99:6967–72.

(23) Yu J, Othman MI, Farjo R, Zareparsi S, MacNee SP, Yoshida S, et al. Evaluation and optimization of procedures for target labeling and hybridization of cDNA microarrays. Mol Vis 2002;8:130–7.

(24) Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A 2001;98:8961–5.

(25) Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8:625–37.

(26) Wilson AS, Hobbs BG, Speed TP, Rakoczy PE. The microarray: potential applications for ophthalmic research. Mol Vis 2002;8:259–70.

(27) Stears RL, Getts RC, Gullans SR. A novel, sensitive detection system for high-density microarrays using dendrimer technology. Physiol Genomics 2000;3:93–9.

(28) Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.

(29) Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001;98:13790–5.

(30) Beer DG, Kardia SL, Huang C, Giorano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival in patients with lung adenocarcinoma. Nat Med 2002;8:816–24.

(31) Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.

(32) Lin Y, Furukawa Y, Tsunoda T, Yue C, Yang K, Nakamura Y. Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. Oncogene 2002;21:4120–8.

(33) Jazaeri AA, Yee CJ, Sotiriou C, Brantley KR, Boyd J, Liu ET. Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. J Natl Cancer Inst 2002;94:990–1000.

(34) Desu MM, Raghavarao D. Sample size methodology. Boston (MA): Academic Press; 1990. p. 30.

NOTES

We thank Jeff Green and Nianqing Xiao for the transgenic mouse data used for our analysis of gene-specific dye bias.

Manuscript received February 10, 2003; revised July 9, 2003; accepted July 16, 2003.