BIOINFORMATICS

Vol. 19 no. 7 2003, pages 803–810
DOI: 10.1093/bioinformatics/btg076

# Statistical design of reverse dye microarrays

*K. Dobbin*, J. H. Shih and R. Simon*

*National Cancer Institute, Biometric Research Branch, 6130 Executive Blvd., MSC 7434, Bethesda, MD 20892, USA*

Received on July 25, 2002; revised on September 26, 2002; accepted on November 13, 2002

**ABSTRACT**

**Motivation:** In cDNA microarray experiments all samples are labelled with either Cy3 dye or Cy5 dye. Certain genes exhibit dye bias—a tendency to bind more efficiently to one of the dyes. The common reference design avoids the problem of dye bias by running all arrays 'forward', so that the samples being compared are always labelled with the same dye. But comparison of samples labelled with different dyes is sometimes of interest. In these situations, it is necessary to run some arrays 'reverse'—with the dye labelling reversed—in order to correct for the dye bias. The design of these experiments will impact one's ability to identify genes that are differentially expressed in different tissues or conditions. We address the design issue of how many specimens are needed, how many forward and reverse labelled arrays to perform, and how to optimally assign Cy3 and Cy5 labels to the specimens.

**Results:** We consider three types of experiments for which some reverse labelling is needed: paired samples, samples from two predefined groups, and reference design data when comparison with the reference is of interest. We present simple probability models for the data, derive optimal estimators for relative gene expression, and compare the efficiency of the estimators for a range of designs. In each case, we present the optimal design and sample size formulas. We show that reverse labelling of individual arrays is generally not required.

**Contact:** dobbinke@mail.nih.gov.

**Supplementary information:** Supplementary material referenced in the text is available at http://linus.nci.nih.gov/~brb/TechReport.htm

## INTRODUCTION

A growing number of cDNA microarray experiments seek to compare samples labelled with red (Cy5) dye to samples labelled with green (Cy3) dye. For example, tumor samples may be co-hybridized with paired normal tissue samples on each array (Boer *et al.*, 2001; Lossos *et al.*, 2002); or, comparison with the common internal reference sample may be of interest (Zhou *et al.*, 2002; Lin *et al.*, 2002; Chu *et al.*, 1998; Jazaeri *et al.*, 2002); or, there may be no reference and several varieties (Jin *et al.*, 2001). Comparisons between differently labelled samples also typically occur in comparative genomic hybridization (Forozan *et al.*, 1997). For each gene, such comparisons use the normalized spot intensity as a proxy for the amount of cDNA that hybridized to a particular spot. Some genes have been observed to incorporate one dye more efficiently than the other (Ideker *et al.*, 2000; Wang *et al.*, 2001; Tseng *et al.*, 2001; Kerr *et al.*, 2001; Goryachev *et al.*, 2001), and therefore may generally tend to appear brighter in one color. As a result, an observed difference between red and green channel intensities for a particular gene may be due to differences in expression level between the samples or differences in dye incorporation efficiency between the dyes. For example, a low intensity spot channel reading may indicate there is a low level of the corresponding cDNA present, or that only a small proportion of the cDNA present successfully incorporated the dye and bound to the array. Gene expression may be confounded with dye incorporation efficiency in these experiments. Normalization of the data typically corrects for dye incorporation differences which affect all the genes similarly, or genes with the same intensity similarly, but not for individual genes which act differently than the rest. These gene-specific dye effects have been observed to exist for some genes (Tseng *et al.*, 2001; Zhou *et al.*, 2002).

Suppose one wishes to compare two groups of samples when some are labelled red and others green. One's ability to distinguish between genes that are truly expressed differently in the groups and genes that incorporate the dyes differently will depend on the experimental design. For example, if one wants to compare normal samples to tumor samples, labelling all the normal samples green (Cy3) and all the tumor samples red (Cy5) will result in confounding between those genes that are expressed differently in cancer tissue, and those genes that incorporate the dyes differently. On the other hand, labelling half the tumor samples green and the other half red, and similarly with the normal samples, may allow one to distinguish between these two classes of genes. The goal of this paper is to exam-

*To whom correspondence should be addressed.

ine which allocations of samples to the arrays and labels to the samples will produce the most accurate, unbiased estimates of the true differences in gene expression between two groups (varieties) of samples. Much of the paper focuses on this class comparison problem, in which the classes are defined independently of the gene expression profiles. Consideration of other goals, such as class discovery, appears in the discussion section.

### Motivation

A majority of cDNA microarray studies use a reference design, in which one aliquot from a reference sample appears on each array with a sample of interest. Usually the samples of interest are all tagged with the same color dye. This means that if a gene has a tendency to bind better to one dye than the other, this effect will not confound comparisons among groups of non-reference samples.

Why design a study in which one will need to adjust for dye effects? We will discuss three situations in which such a design may be desirable: (1) when specimens occur naturally in pairs; (2) when identification of genes expressed differently in two varieties is the only goal; (3) when identification of genes expressed differently in the reference sample and the non-reference samples is desired. In situations (1) and (2), a design that uses a reference, and tags all non-reference samples with the same dye, will be less efficient than a design that avoids the use of a reference. In situation (3), one is clearly forced to compare samples tagged with different dyes.

Some examples of paired samples are: (1) a collection of patients from whom a sample of normal tissue and a sample of tumor tissue was drawn, with the goal of identifying genes expressed differently in tumor and normal tissue (on average, across individuals); (2) a collection of paired tumor samples, in which one member of each pair was taken before treatment and the other was taken after treatment, with the goal of determining the effect of treatment on gene expression in the tumors; (3) a collection of RNA samples from two conditions which have been paired based on covariate or clinical information. For paired samples, the quantity of interest is the difference in expression between the two members in each pair. For a fixed number of arrays, a design which places each member of a pair on a separate array with a reference will be less efficient than one which runs the pair together, forward on one array and reversed on the other (to guard against potential dye bias). Comparing the two members of a pair will then require comparing samples tagged with different dyes.

Another situation arises when the goal of an experiment with unpaired samples is focused on comparing two varieties to identify differentially expressed genes. For instance, one may wish to identify genes differentially expressed in estrogen receptor positive and estrogen receptor negative breast tumor specimens. In this case, it has been shown that one can get equivalent results with fewer arrays by placing one sample from each variety on each array than by using a reference design (Dobbin and Simon, 2002; Kerr *et al.*, 2001; Cochran and Cox, 1992). Such a design is referred to as a balanced complete block design, and necessitates comparison of samples across dyes. (These designs are not optimal in other respects, only for identifying differentially expressed genes.)

Sometimes researchers desire to compare the non-reference samples to the reference sample. For example, a mixture of normal RNA is commonly used as a reference for tumor tissue. Comparing the normal mixture reference to the non-reference samples may indicate which genes are differentially expressed in the tumor tissue, and suggest potential tumor markers. In such cases, comparison with the reference may be a primary or secondary goal of the experiment. In either case, one will clearly be forced to compare samples tagged with different color dyes.

## METHODS AND RESULTS

Our approach to design comparisons utilizes analysis of variance (ANOVA) models (Kerr *et al.*, 2001; Wolfinger *et al.*, 2001; Lee *et al.*, 2000). For each gene, a separate ANOVA model is fit to the arrays; most effects in the model are not of interest, but are included because this automatically adjusts the estimates of interest to take into account these other sources of variability. Our main yardstick for comparing designs will be *efficiency*. Efficiency has a quantitative definition for statistical models, which intuitively corresponds to the notion of amount of output for a fixed input. The input will be the number of arrays used in the experiment. The output will be the accuracy of the estimated differences in average gene expression between the classes. If design A is twice as efficient as design B, then it will require twice as many arrays under design B to obtain the same accuracy as under design A.

Our model differs from the model given in Kerr *et al.* (2001) in the following respects: (1) they assumed a common variance for all genes, whereas we allow each gene to have it's own variance; (2) they did not incorporate variation among samples of the same variety, whereas we include such effects; (3) they did not have gene-specific dye effects in their model. Some of these differences may be attributable to the fact that the authors restricted attention to an experiment with just two arrays.

### Non-reference designs for paired samples

Paired samples typically consists of 'before treatment' and 'after treatment' RNA samples for each individual, or 'tumor' and 'normal' samples from each individual in a study. The main interest is in understanding the average

effect the 'treatment' or disease has on gene expression. This can help identify genes which are affected by the treatment, while at the same time eliminating person-to-person population variation in expression levels. Since the most accurate comparisons between samples in a cDNA microarray experiment are made between the two channels on a single slide, it is desirable that cancer and normal tissue from the same individual always appear together on a single array. To simplify presentation, throughout this section the two varieties are represented by 'normal' and 'tumor', although more general paired samples as described in the motivation section are also implied.

In order to correct for gene-specific dye bias, we will need to run some arrays forward and others reverse. Balancing the dyes and the varieties, so that each variety is tagged with each dye in half the samples, will minimize the variance of the variety effect contrasts; so we will want half the 'normal' samples tagged red and the other half green; and similarly for the 'tumor' samples. One can either run the same individuals both forward and backward, or one set of individuals forward and a different set backward, or some compromise between these two. It will generally be suboptimal to have RNA from the same individual on more than two arrays (e.g. once forward and once reverse), because for a fixed number of arrays, the more samples one has from each individual the fewer the number of individuals and the larger the variance of estimated population parameters. The loss in efficiency from replicating individuals on arrays instead of collecting new samples will be greatest when the population variance is large relative to the experimental error. But even in poor quality microarray experiments in which experimental variance is much larger than population variance, one will still lose some efficiency by repeating samples (see supplementary material, Appendix F).

These considerations lead to a range of design options represented by Table 1. Throughout the text, $k$ will represent the number of samples that are run both forward and backward on different arrays, $n - k$ the number of samples appearing only once on an array, and $m = n + k$ the total number of arrays used.

*ANOVA Formulation* To simplify presentation, we assume that the intensity data have been background adjusted and normalized (e.g. the two channels on each array have been median centered, and all the arrays have been median centered).

Let $r_{gadvp}$ be a background-adjusted, normalized log-intensity. In the subscripts, $g$ indexes the genes, $a$ indexes the array, $d$ indexes the two dyes. The $v$ indicates the varieties. The $p$ indexes the individual participants involved in the study. We propose the model

$$r_{gadvp} = G_g + GA_{ga} + GD_{gd} + GV_{gv} + GP_{gp} + GVP_{gvp} + \epsilon_{gadvp} \quad (1)$$

For an individual spot on a particular array, this model postulates that the observed background-adjusted, normalized log-intensity is a result of additive effects of the amount of RNA in the sample, the size and quality of the spot, the dye effects, and random error. Included in the random error are inhomogeneities in the RNA sample and technical issues in the measurement, extraction, and reverse-transcription and labelling reactions.[†] Differential gene expression is represented in the $GV$ interaction, which is the term of interest. Further discussion of the model appears in section 1 of the supplementary material.

The analysis of variance table for paired samples is given in Table A of the supplement. We present three ANOVA tables because there are often very few or no degrees of freedom for estimating the sample-specific effects $GP$ and $GVP$. Further discussion of when these effects should be excluded from the model appears in the supplement. In fact, a single design appears most efficient for all three cases.

*Results for paired samples.* Assume the total number of arrays is fixed at $m = n + k$. In the supplementary material we show that in each of the three cases given in supplement Table A, a design that runs each sample once on an array, and balances the samples with respect to the dyes, will be most efficient for paired samples. This design minimizes the variance of the main estimated contrast of interest, $\widehat{GV}_{g1} - \widehat{GV}_{g0}$. The sample size formula appears in the sample size section in what follows.

## Non-reference designs for unpaired samples

Sometimes the research question has a focused goal of comparing two varieties with each other, e.g. to identify differentially expressed genes, but there is no clear way to pair the samples. A reference design may be used in this case, although a non-reference design has been shown to be more efficient (Dobbin and Simon, 2002; Kerr *et al.*, 2001). Each array should contain one sample from each variety.

The arrays should be balanced with respect to the dyes, so that half the arrays are run forward and half reverse, because this will produce minimum variance estimates of the variety contrast. In general, we want to minimize the number of times the same sample occurs on an array because this results in loss of replicates at the population level and loss of efficiency in comparing varieties; on the other hand, repeating samples on multiple arrays may give more accurate estimates of gene-specific dye bias than avoiding such replication altogether. These considerations give rise to a collection of designs given by Table 2a.

[†] Our model assumes a single RNA extraction for each sample.

**Table 1.** Paired Samples Design.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Array 1** | **Array 2** | **…** | **Array** $k$ | **Array** $k+1$ | **…** | **Array** $n$ |
| **Green** | Normal 1 | Normal 2 | … | Normal $k$ | Normal $k+1$ | … | Cancer $n$ |
| **Red** | Cancer 1 | Cancer 2 | … | Cancer $k$ | Cancer $k+1$ | … | Normal $n$ |
| | **Array** $n+1$ | **Array** $n+2$ | **…** | **Array** $n+k$ | | | |
| **Green** | Cancer 1 | Cancer 2 | … | Cancer $k$ | | | |
| **Red** | Normal 1 | Normal 2 | … | Normal $k$ | | | |
| | ↑ | ↑ | | ↑ | ↑ | | ↑ |
| | Individual 1 | Individual 2 | … | Individual $k$ | Individual $k+1$ | … | Individual $n$ |

'Normal 1' indicates a sample of normal tissue from individual 1, and 'Cancer 1' a sample of tumor tissue from individual 1. Array 1 represents a forward experiment for participant 1, and Array $n+1$ a backward experiment for the same individual. $k$ represents the number of samples that are run both forward and reverse. $n-k$ represents the number of samples that are run only once; these are assumed to be balanced, so that $\frac{n-k}{2}$ are run forward, and $\frac{n-k}{2}$ are run backward ($n-k$ is assumed even).

**Table 2.** Other designs.

Design for comparing two unpaired varieties[a]

| | | | | | | |
|---|---|---|---|---|---|---|
| | **Array 1** | **…** | **Array** $k$ | **Array** $k+1$ | **…** | **Array** $n$ |
| **Green** | Cancer 1 | … | Cancer $k$ | Cancer $k+1$ | … | Normal $2n$ |
| **Red** | Normal $n+1$ | … | Normal $n+k$ | Normal $n+k+1$ | … | Cancer $n$ |
| | **Array** $n+1$ | **…** | **Array** $n+k$ | | | |
| **Green** | Normal $n+1$ | … | Normal $n+k$ | | | |
| **Red** | Cancer 1 | … | Cancer $k$ | | | |

Reference design when comparison of the non-reference samples among themselves is the main goal[b]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Array 1** | **Array 2** | **…** | **Array** $k$ | **Array** $k+1$ | **…** | **Array** $n$ |
| **Green** | Reference | Reference | … | Reference | Reference | … | Reference |
| **Red** | Sample 1 | Sample 2 | … | Sample $k$ | Sample $k+1$ | … | Sample $n$ |
| | **Array** $n+1$ | **Array** $n+2$ | **…** | **Array** $n+k$ | | | |
| **Green** | Sample 1 | Sample 2 | … | Sample $k$ | | | |
| **Red** | Reference | Reference | … | Reference | | | |

Reference design when comparison with the reference is the main objective[c]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Array 1** | **Array 2** | **…** | **Array** $k$ | **Array** $k+1$ | **…** | **Array** $n$ |
| **Green** | Reference | Reference | … | Reference | Reference | … | Sample $n$ |
| **Red** | Sample 1 | Sample 2 | … | Sample $k$ | Sample $k+1$ | … | Reference |
| | **Array** $n+1$ | **Array** $n+2$ | **…** | **Array** $n+k$ | | | |
| **Green** | Sample 1 | Sample 2 | … | Sample $k$ | | | |
| **Red** | Reference | Reference | … | Reference | | | |

[a]Tumor tissue samples from $n$ individuals and normal tissue samples from $n$ different individuals. $k$ sample pairs are run both forward and reverse. The $n-k$ samples which are run only once are balanced with respect to the dyes, so that half are run forward and half reverse.
[b]Unreplicated samples are all run forward to optimize class discovery and robustness of comparisons among the non-reference samples (Dobbin and Simon, 2002).
[c]Unreplicated samples are run half forward and half reverse to preserve balance and produce the most efficient comparisons between the reference sample and non-reference variety.

*ANOVA formulation.* Our ANOVA model is as follows:

$$r_{gadvf} = G_g + GA_{ga} + GD_{gd} + GV_{gv} + GF_{gf} + \epsilon_{gadvf}.$$

This is the same as the model of the last section except that we have replaced the $GP$ gene by participant interaction with a $GF$ gene by sample interaction. Note that there is a conceptual shift here, because instead of having two varieties for each individual (cancer and normal), now each individual is associated with just one variety, and sample effects $GF$ are nested in variety effects $GV$. This implies that it makes no sense to have an interaction ($GVF$) between sample and variety. For simplicity, we assume we have just two varieties, $GV_{g1}$ and $GV_{g2}$. There seems to be no *a priori* reason to think the inter-sample variability will be equal in the two varieties, so we will allow each population to have its own inter-sample variation, and we will denote these parameters $\tau^2_{g1}$ and $\tau^2_{g2}$ respectively.

*Results for unpaired samples* In Appendix C of the supplement, the minimum variance linear unbiased estimator is derived, and it is shown that the variance of the estimator is minimized when $k = 0$. For a fixed number of arrays, the most efficient design will have a different pair of samples on each array, and the dyes and varieties will be balanced—so that each variety has half the samples tagged red and the other half green. The sample size formula corresponding to the most efficient design is given in the sample size section below.

## Reference designs for comparing a common reference to non-reference samples

We now turn to the situation in which a reference design is to be used, and one desires to compare the common reference to the non-reference samples. For example, the reference sample may be a mixture of normal tissue and the non-reference samples RNA extracted from different tumors, so that the comparison would give some indication of genes expressed differently in the tumors. In this type of experiment, we are really interested in testing the null hypothesis $H_0 : \mu_g = \nu_g$ versus $H_1 : \mu_g \neq \nu_g$ where $\mu_g$ is the population mean for tumor samples and $\nu_g$ is the population mean for normal samples. But we cannot test this hypothesis because we only have a single sample from the normal tissue (even if it is a mixture), so we have no way to estimate the variation in the normal tissue; we need such an estimate to test the hypothesis. Since we are not able to test the hypothesis of interest, we instead test a similar hypothesis. We test the hypothesis $H_0 : \mu_g = \bar{y}_g$ versus $H_1 : \mu_g \neq \bar{y}_g$ where $\bar{y}_g$ represents the average expression level for this gene in the reference mixture. The results of this hypothesis test may be of biological interest, but may also be problematic. For instance, unless the reference pool is a homogeneous mixture from a large number of RNA samples, the $\bar{y}_g$ may not be close to the population parameter $\nu_g$, and the hypothesis test not a good approximation to the one in which we are really interested. Throughout this section, we assume that the reference RNA is homogeneous, so that variation among the measurements on the reference RNA sample is small compared to variation among the non-reference sample measurements.

*ANOVA formulation.* Let $r_{gadvf}$ represent background-adjusted, normalized log intensity as before. We propose the model

$$r_{gadvf} = G_g + GA_{ga} + GD_{gd} + GV_{gv} + GF_{gf} + \epsilon_{gadvf}. \quad (2)$$

One 'variety' here consists of the non-reference samples, and the other 'variety' of the reference sample. The error term $\epsilon$ is assumed normally distributed with mean zero and variance $\sigma_g^2$. The ANOVA table for these data is given in Table B of the supplementary material. The rightmost column represents the degrees of freedom when no sample pairs are repeated on the arrays.

The data are examined by fitting the model of Equation 2 for each gene. The $GD$ interaction term is the potential source of bias. The $GV$ term is the effect of interest. Variation among the $GF$ effects for the non-reference samples represents biological variation among samples of the same variety in the population from which the non-reference RNA was drawn, and variation in the RNA extraction and reverse transcription process. The variance of the variety contrast estimate will depend on the variation among the $GF$ terms, so to compare estimates some assumption about this variation must be made. We will assume that for a given gene, the $GF$ terms are independent and normally distributed with mean 0 and variance $\tau_g^2$.

*Results when comparison with reference is a secondary goal.* Often, the main objective of a microarray experiment is comparison of the non-reference samples, either supervised analysis to compare different types of tumors or unsupervised analysis to identify new taxonomies for the tumors. When this is the case, the most efficient design will be different than when comparison with the reference is the primary goal; in particular, it will generally be suboptimal to balance the varieties and the dyes. Reference designs in which most or all of the samples are tagged with the same dye have many advantages in these situations, they tend to be robust, relatively simple to analyze, and produce better cluster analysis results than other designs (Dobbin and Simon, 2002). For these reasons, when comparison of non-reference varieties is the main objective of the experiment, one may wish to restrict attention to reference design experiments with chiefly forward arrays, but appended by enough reverse arrays to allow good comparison of the reference to the non-reference. An example of this design is given in Table 2b.

In Appendix D of the supplement, we derive the minimum variance linear unbiased estimator of the variety contrast between the reference variety and the non-reference variety. Hence, we are here considering how to optimize the experiment with respect to the secondary goal of efficient comparison with the reference. The variance of this contrast estimate for $k > 0$ is

$$\text{var}\left(\widehat{GV}_{g1} - \widehat{GV}_{g0}\right) = \frac{n+3k}{(n+k)^2}\tau_g^2 + \frac{n^2+3k^2}{k(n+k)^2}\sigma_g^2,$$

where variety subscript '0' indicates the reference variety, and '1' the non-reference variety. For fixed $m = n + k$, the $k$ which will minimize the variance is $k = \max\left(1, \frac{m\sigma_g}{\sqrt{2\tau_g^2+4\sigma_g^2}}\right)$. (We require $k > 0$ in this case because if $k = 0$, then we cannot correct the dye bias.) If the biological variation is small compared to the experimental error ($\frac{\tau}{\sigma}$ near 0), then all the samples should

be run both forward and backwards (that is, $k = \frac{m}{2}$) so that one gets the most accurate dye bias correction. On the other hand, if biological variation is large compared to experimental error ($\frac{\tau}{\sigma}$ large), then a single sample should be run both forward and reverse (that is, $k \rightarrow 1$) so that one maximizes the replication at the population levels to offset the large biological variation. In one dataset on human cell lines we examined (unpublished) the ratio from a high-quality experiment had median 2.7. Plugging this into the equation indicates that the most efficient design has approximately one-fourth of the arrays reversed and three-fourths forward.

Of course, optimizing with respect to the secondary goal of comparison with the reference may not make much sense if too great a cost to the primary goal is involved. And there is an inverse relation between number of reverse arrays and effective sample size for the primary goal. A more practical guideline is to run some minimal number of reverse experiments that will provide enough degrees of freedom for error to permit good inference for comparisons between the reference and non-reference varieties (error degrees of freedom appear on Table B of the supplement).

Sample size calculations should be based on the primary goal, i.e. comparisons of the non-reference samples. If avoiding false-positives and false-negatives in the comparison between the reference and non-reference is important, then one should run enough reverse to provide reliable F-tests. If these are of lesser importance, then one may run fewer reverse dye experiments, which will allow for better inference among the non-reference samples (for a fixed number of arrays).

*Results when comparison with reference is primary goal.* Comparison with the reference may also be the primary goal of the experiment. In this case, the varieties should be balanced with respect to the dyes, so that each variety appears tagged with each dye in half the samples, because this will minimize the contrast variance. An example of the design is given in Table 2c. Here, $n - k$ is assumed even, and half the arrays from $k + 1$ to $n$ are run forward and the other half reverse. In Appendix E of the supplement, we show that for this type of design, the variance of the estimated contrast between the reference sample and non-reference variety is minimized for a fixed number of arrays $m = n + k$ when $k = 0$, i.e. when each non-reference sample appears on exactly one array, and the varieties are balanced with respect to the dyes. The sample size formula for the most efficient design appears in the sample size section in what follows

### Sample sizes for most efficient designs

In the previous sections, we have found the most efficient non-reference designs for paired and unpaired samples, and the most efficient reference design when comparison with the reference is the goal. Here we present sample size formulas for each of these most efficient designs.

For conciseness, a single formula will be presented which can be used to determine the sample size for any of the designs. In each case, an estimate of the variance of the log-ratios under that particular design is needed to determine the sample size required. Importantly, we do not need separate estimates of the individual variance parameters $\tau_g^2$ and $\sigma_g^2$ to determine the sample size. Suppose we wish to test for differentially expressed genes at the $\alpha$ significance level, and have a sample size large enough to detect a difference of $\delta$ in the log-intensities with power $1 - \beta$. Let $V_g$ be the variance of the log-ratios under the design to be used. Note that $V_g$ is a general notation for the variance of the log-ratios, but that this variance will be different for different designs. For example, the variance may be smaller with paired samples than with unpaired samples. Let $m$ be the number of arrays. The sample size formula for all three cases can then be written in the compact form

$$m = V_g \left[ \frac{z_{\alpha/2} + z_\beta}{\delta} \right]^2 .$$

The notation $z_{\alpha/2}$ represents the $100(1-\alpha/2)$th percentile of the normal distribution. (For small sample sizes, the $t$-distribution adjustment may be used.) Derivation of the sample size formulas appear in the supplement.

The formula can be applied to determine the sample size for the most efficient design in each situation we have discussed with one exception. For a reference design in which comparison of the non-reference samples among themselves is the primary goal, and comparison of the non-reference samples to the reference is the secondary goal, sample size should be determined by the primary goal.

## DISCUSSION

Dye bias may be an issue when samples tagged with different dyes are to be compared. We have argued that in these situations, it is not necessary to run every sample pair twice so as to eliminate the dye bias. In fact, we have shown that it is often most efficient to avoid repeating sample pairs altogether, and instead balance the varieties being compared with respect to the dyes, so that each variety is tagged with each dye in half the samples. We have seen that this is true with paired samples, with unpaired samples comparing two varieties, and with reference design data (when comparison with reference is the primary goal). Figure 1 summarizes these results. It is important to note that even if dye bias exists, it is generally wasteful to run the same samples both forward and backward on separate arrays. We have also given sample size formulas based on simple statistical models.

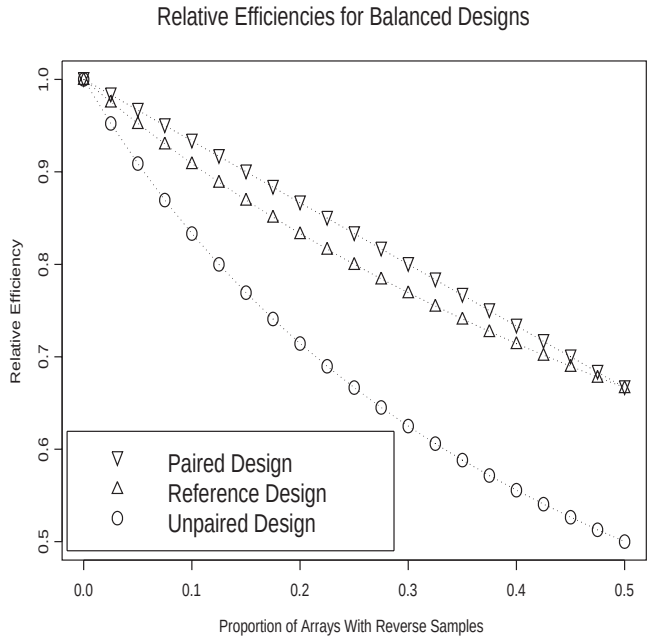


**Fig. 1.** Balanced designs comparison: relative efficiencies versus the proportion of arrays with reverse samples, i.e. proportion of arrays with the same RNA samples as forward arrays but with the direction reversed. Relative Efficiency $= \frac{\text{Efficiency}(k/m)}{\text{Efficiency}(0)}$ where Efficiency$(k/m)$ is the efficiency of the contrast estimate $\widehat{GV}_{g1} - \widehat{GV}_{g0}$ when $k/m$ is the proportion of arrays with reverse samples. Parameter settings are $\tau_g^2 = 2\sigma_g^2$ for the paired and reference designs, and $\tau_{g1}^2 = \tau_{g2}^2 = 2\sigma_g^2$ for the unpaired design. In all three cases, 0 reverse arrays maximizes the efficiency.

When comparison of the non-reference samples among themselves is the primary goal (e.g. by cluster analysis), and comparison with the reference a secondary goal, we have presented a formula and some practical guidelines for selecting the number of reverse arrays.

Since our model characterizes dye bias and experimental variation as gene-specific, one can analyze the data gene by gene. Some have assumed a common variance for all genes (Kerr *et al.*, 2001), or particular variance–covariance structure across genes (Rocke and Durbin, 2001; Tusher *et al.*, 2001; Ideker *et al.*, 2000). In general, comparing designs for complex variance-covariance structures may be problematic because the form of the contrast estimate for a particular gene may be complex and not known at the design phase. But if we assume a common variance across genes in our model, this will not affect the form of our variety contrast estimates or the variance calculation for these estimates (it would only affect the way in which the variances themselves were estimated). Hence our results will remain unchanged under a multivariate model with equal error variance.

Further discussion of the recommended designs and the ANOVA model assumptions appear in the supplement.

We would recommend using the balanced designs we have described even if one believes no gene-specific dye bias will be present. Many studies have performed reverse arrays to guard against dye bias (Bayani *et al.*, 2002; Zhou *et al.*, 2002; Klebes *et al.*, 2002; Aharoni *et al.*, 2000; Barrans *et al.*, 2002; Desai *et al.*, 2002), and there is abundant literature discussing dye bias adjustments (Tseng *et al.*, 2001; Yu *et al.*, 2002; Yang *et al.*, 2001; Kerr *et al.*, 2001; Wolfinger *et al.*, 2001). There is ongoing work in dye labelling technology to try to reduce or eliminate these dye effects (Wilson *et al.*, 2002; Stears *et al.*, 2000; Manduchi *et al.*, 2002; Yu *et al.*, 2002). While some of this work is promising, there is not a consensus that the problem has been 'solved' by these technologies. Besides, even in the absence of gene-specific dye effects, only in one of the four cases we described would one lose efficiency by designing the experiment as we have suggested (namely, in reference designs when comparisons with the reference is a secondary concern, in which case no arrays should be run reverse). In all other cases, one loses nothing in efficiency by following our designs, and in fact one gains the ability to detect and correct for gene-specific dye biases if any exists. One essentially gains in robustness with no loss in efficiency.

## REFERENCES

Aharoni,A., Keizer,L.C., Bouwmeester,H.J., Sun,Z., Alvarez-Huerta,M., Verhoeven,H.A., Blaas,J., van Houwelingen,A.M, De Vos,R.C., van der Voet,H.*et al.* (2000) Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *Plant Cell*, **12**, 647–661.

Barrans,J.D., Allen,P.D., Stamatiou,D., Dzau,V.J. and Liew,C. (2002) Global gene expression profiling of end-stage dilated cardiomyopathy using a human cardiovascular-based cDNA microarray. *Am. J. Pathol.*, **160**, 2035–2043.

Bayani,J., Brenton,J.D., Macgregor,P.F., Beheshti,B., Albert,M., Nallainathan,D., Karaskova,J., Rosen,B., Murphy,J., Laframboise,S. *et al.* (2002) Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. *Cancer Res.*, **62**, 3466–3476.

Boer,J.M., Huber,W.K., Sültmann,H., Wilmer,F., von Heydebreck,A., Haas,S., Korn,B., Gunawan,B., Vente,A., Füzesi,L. *et al.* (2001) Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res.*, **11**, 1861–1870.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Cochran,W.G. (1977) *Sampling Techniques*, Third Edition, Wiley, New York.

Cochran,W.G. and Cox,G.M. (1992) *Experimental Design*, Second Edition, Wiley, New York.

Desai,K.V., Xiao,N., Wang,W., Gangi,L., Greene,J., Powell,J.I.,

Dickson,R., Furth,P., Hunter,K., Kucherlapati,R. *et al.* (2002) Initiating oncogenic event determines gene-expression patterns of human breast cancer models. *Proc. Natl Acad. Sci. USA*, **99**, 6967–6972.

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

Forozan,F., Karhu,R., Kononen,J., Kallioniemi,A. and Kallioniemi,O. (1997) Genome screening by comparative genomic hybridization. *Trends Genet.*, **13**, 405–409.

Goryachev,A.B., Macgregor,P.F. and Edwards,A.M. (2001) Unfolding of microarray data. *J. Comput. Biol.*, **8**, 443–461.

Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Jazaeri,A.A., Yee,C.J., Sotiriou,C., Brantley,K.R., Boyd,J. and Liu,E.T. (2002) Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. *J. Natl Cancer Inst.*, **94**, 990–1000.

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 Genet.*, **29**, 389–395.

Kerr,M.K., Martin,M. and Churchill,G.A. (2001) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Kerr,M.K., Afshari,C.A., Bennett,L., Bushel,P., Martinez,J., Walker,N.J. and Churchill,G.A. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, **12**, 203–217.

Klebes,A., Biehs,B., Cifuentes,F. and Kornberg,T.B. (2002) Expression profiling of *Drosophila* imaginal discs. *Genome Biol.*, **3**, research0038.1–0038.16.

Lee,M.T., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Lin,Y., Furukawa,Y., Tsunoda,T., Yue,C., Yang,K. and Nakamura,Y. (2002) Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. *Oncogene*, **21**, 4120–4128.

Lossos,I.S., Alizadeh,A.A., Diehn,M., Warnke,R., Thorstenson,Y., Oefner,P.J., Brown,P.O., Botstein,D. and Levy,R. (2002) Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. *Proc. Natl Acad. Sci. USA*, **99**, 8886–8891.

Manduchi,E., Scearce,L.M., Brestelli,J.E., Grant,G.R., Kaestner,K.H. and Stoeckert,Jr,C.H. (2002) Comparison of different labeling methods for two-channel high-density microarray experiments. *Physiol Genomics*, **10**, 169–179.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Sadooghi-Alvandi,M. (1986) The choice of subsample size in two-stage sampling. *J. Am. Stat. Assoc.*, **81**, 555–558.

Scheffé,Henry (1999) *The Analysis of Variance*. Wiley, New York.

Stears,R.L., Getts,R.C. and Gullans,S.R. (2000) A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol Genomics*, **3**, 93–99.

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 Res.*, **29**, 2549–2557.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **9**, 5116–5121.

Wang,X., Ghosh,S. and Guo,S. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, e75.

Wilson,A.S., Hobbs,B.G., Speed,T.P. and Rakoczy,P.E. (2002) The microarray: potential applications for ophthalmic research. *Mol. Vis.*, **8**, 259–270.

Wolfinger,R.D., Gibson,G., Wolfinger,E.D., Bennett,L., Hamadeh,H., Bushel,P., Afshari,C. and Paules,R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

Yang,Y.H., Dudoit,S., Luu,P. and Speed,T.P. (2001) Normalization for cDNA microarray data. *Technical Report, 589*. http://stat-www.berkeley.edu/tech-reports/index.html

Yu,J., Othman,M.I., Farjo,R., Zareparsi,S., MacNee,S.P., Yoshida,S. and Swaroop,A. (2002) Evaluation and optimization of procedures for target labeling and hybridization of cDNA microarrays. *Mol. Vis.*, **8**, 130–137.

Zhou,Y., Gwadry,F.G., Reinhold,W.C., Miller,L.D., Smith,L.H., Scherf,U., Liu,E.T., Kohn,K.W., Pommier,Y. and Weinstein,J.N. (2002) Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: microarray analysis of dose- and time-dependent effects. *Cancer Res.*, **62**, 1688–1695.