



UW Biostatistics Working Paper Series

6-13-2003

Design Considerations for Efficient and Effective Microarray Studies

M. Kathleen Kerr

University of Washington, katiek@u.washington.edu

Suggested Citation

Kerr, M. Kathleen, "Design Considerations for Efficient and Effective Microarray Studies" (June 2003). *UW Biostatistics Working Paper Series*. Working Paper 210.

<http://biostats.bepress.com/uwbiostat/paper210>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

1. Introduction

Gene expression microarrays are a tool of modern genomics that have contributed to the current excitement in molecular biology and genetics (Schena et al. (1995)). The impact of microarrays has been similarly profound for the field of statistics. Microarray data have stimulated research in diverse areas such as transposable data (Lazzeroni and Owen (2002)), multiple testing and false discovery rates (Storey (2002)), and classification tools (Dudoit et al. (2002)). However, not every aspect of microarrays requires revolutionary statistical approaches. To the contrary, many aspects of microarray studies call for careful attention to fundamental statistical principles (Kerr and Churchill (2001b)).

Microarrays have been used in some fascinating research and the technology is tantalizing in the possibilities it presents. In light of this excitement, some perspective is warranted. Basically, microarrays are a measurement tool, albeit a high-tech and high-throughput one. There are many unknown quantities in a microarray hybridization, such as the sizes and densities of the probe spots, and the hybridization and labeling efficiencies of different sequences. However, regardless of these variations, the basic principle is the following: For a given sequence spotted on the array, if one sample contains more of the corresponding transcript, then the signal intensity for the dye used to label that sample should be higher than for the other dye. There is a further assumption of proportionality. That is, if the red-labeled sample has twice as much of a transcript as the green-labeled sample, then the red signal should be twice as high as the green signal. (Readers are referred to the literature for an introduction to microarray technology – Nguyen et al.

(2002) provide an excellent summary for quantitative scientists.)

A microarray assay is more complicated in reality, largely because the behavior of the two fluorescent dyes is more complex. Endeavors in data normalization (Cui et al. (2003), Tseng et al. (2001), Yang et al. (2002)) are essentially efforts to return to the basic idea of how a microarray assay is supposed to work. With this simplified perspective, a microarray is really just a comparison between two RNAs and a microarray design is a block design with block size two. Although there is a mature literature in statistical block design, some aspects of microarray design present new and interesting questions in this area.

This paper discusses design in the context of two-color spotted microarrays. However, most of the principles in Sections 2, 3, and 4 also apply to single-channel platforms such as Affymetrix (Lipshutz et al. (1999)). This paper has two primary goals. The first goal is to give practical guidelines for microarray experimental design that have been developed through experience designing these studies in collaboration with biologists. The second goal is to identify areas where microarrays present new design problems and where additional research is needed. Randomization, replication, and blocking are generally considered to be the three fundamental principles of statistical design. These are discussed with respect to microarrays in Sections 2, 3, and 5. Section 4 discusses the considerations in pooling RNA samples.

2. Randomization

One of the fundamental principles of good design is randomization, yet it is seldom mentioned with microarrays. In fact, microarray experiments are often multi-stage experiments (McIntyre (1955)), and thus require multi-

ple levels of randomization. If there is a “treatment phase,” then individuals should be randomly assigned to treatment groups as in any other experiment. The microarray assays comprise the “measurement phase” of the experiment. Experience shows microarray data are extremely prone to influence by technical artifacts, so it is important to randomize as much as practical to protect against unanticipated biases. For example, arrays should be chosen randomly for each planned hybridization from the batch of arrays to be used in case there is systematic variation in the order in which the arrays were printed.

3. Replication

Replication is another fundamental principle of design and may be the most widely appreciated. Every scientist who conducts a sample-size calculation is recognizing the importance of replication. Microarray users now acknowledge that “replication” means different things in the microarray context (Kerr and Churchill (2001b), Nguyen et al. (2002), Yang and Speed (2002)). “Replication” might refer to:

- (A) Spotting genes multiple times per array;
- (B) Hybridizing multiple arrays to the same RNA samples;
- (C) Using multiple individuals of a certain variety or type.

Replication types (A) and (B) are sometimes referred to as *technical* replication. These are fundamentally different from type (C). Only (C) represents replication in the classical statistical sense — random sampling of individuals from a population in order to make inferences about that population. For example, when a treatment is applied to a mouse model of a disease, the

scientific question is how the treatment affects diseased mice in general, not how it affects the particular mice in the study. We study individual mice in order to make inference about the population.

Technical replication does not address biological variability. Instead, it addresses the measurement error of the assay. Technical replicates reduce the uncertainty about gene expression in the particular RNAs in a study. This is extremely useful in situations where RNAs are of interest individually. For example, if microarrays are ever used in medical diagnosis, technical replicates could be useful to improve the precision of measurements in a particular patient and thereby improve diagnostic accuracy. However, technical replicates can never substitute for biological replicates to assess biological variability, which is essential, for example, to infer that the mean expression of a gene differs in two populations. As another example, developing classification tools based on gene expression data (Dudoit et al. (2002)) requires knowledge of biological variability. Experimental variability is important to understand and control, but is unrelated to the biology under investigation. Confusing technical replication with true replication is not new to microarrays. “Too often researchers use duplicate or split samples to generate two observations and call them replicates, when, in reality, they are actually subsamples or repeated measures” (Milliken and Johnson (1994, p. 49)).

An elementary calculation in a simplistic setting is instructive. Suppose that we want to compare the means of two populations. Let X_i have mean μ_x and variance τ^2 and let Y_i have mean μ_y and variance τ^2 . We sample n individuals from each population. Instead of observing X_i or Y_i we observe $X_{ij} = X_i + \epsilon_{ij}$ and $Y_{ij} = Y_i + \epsilon'_{ij}$ due to measurement error, $i = 1, \dots, n$, $j =$

$1, \dots, r$; The random variables $\epsilon_{ij}, \epsilon'_{ij}$ are *i.i.d.* with mean 0 and variance σ^2 ; they are also independent of the X_i and Y_i . To control measurement error, we may take $r > 1$ measurements on each sampled individual and compute the estimates $X_{i.}$ and $Y_{i.}$ for the i^{th} individual from each sample. Then the difference in population means $\mu_x - \mu_y$ is estimated by $X_{..} - Y_{..}$. Now, suppose measurements are expensive, and we are limited to a fixed total of $N = 2nr$ measurements. How do the relative sizes of population variance τ^2 and error variance σ^2 determine the optimal allocation of N to sampling individuals versus repeating measurements on individuals? One might guess that if measurement error is very large ($\sigma^2 \gg \tau^2$), then it is advantageous to re-measure sampled individuals to control measurement error rather than sampling additional individuals. This turns out not to be the case. The variance of our estimate $X_{..} - Y_{..}$ is

$$\frac{2}{n}\tau^2 + \frac{2}{rn}\sigma^2 = \frac{2}{n}\tau^2 + \frac{4}{N}\sigma^2, \quad (1)$$

which shows that when the total number of measurements N is fixed, it is always preferable to sample new individuals (increase n) to reduce (1) rather than expend resources on repeated measurements of the same individuals. That is, if the cost of sampling individuals is negligible compared to the cost of taking a measurement, which may be the case with microarrays, it is always advantageous to forego repeated measurements and sample as many individuals as possible. Intuitively, repeated measurements provide new information about only the error variance σ^2 whereas additional individuals provide independent information about the total variance $\sigma^2 + \tau^2$.

Of course, even with expensive microarrays the cost of sampling additional individuals may be much greater than the cost of the assays. In other

words, N is not fixed but there is a limit on the total cost C based on the cost of sampling individuals C_s and the cost of taking measurements C_m , $C = 2nC_s + 2rnC_m$. If the error variance σ^2 is substantial then it is efficient to take repeated measurements by increasing r when additional measurements are much cheaper than sampling individuals ($C_m \ll C_s$). These simple calculations do not apply directly to two-color microarrays because of the blocking structure in the design, as discussed in Section 5. However, the general lesson is instructive: true replication beats technical replication for gains in precision when estimating population parameters.

In some pilot studies investigators may wish to assume that a single individual is representative of the population. If a population is relatively homogeneous (e.g. genetically identical mice) and an investigator wishes to use microarrays to identify genes with the largest effects for further study, this may be reasonable. However, the investigator should be aware that he or she is making a critical assumption and, moreover, the assumption cannot be evaluated with the data. In general, microarray studies should assess or account for biological variability with replicates in order to produce rigorous scientific results about the populations of interest. The difference between “assessing” and “accounting for” this variation is discussed in Section 4.

4. Pooling

Investigators sometimes propose to pool RNA samples from individuals (e.g., Jin et al. (2001)). Pooling may be wholly or partly motivated by the fact that an insufficient quantity of RNA can be obtained from a single individual to hybridize to an array. If this is the case, then either pooling RNAs or using an RNA amplification procedure are the only courses of action if a microarray

study is to be performed.

Sometimes pooling is proposed because biological variation is recognized and pooling is meant to “control” this variation. Often this intention is misguided, because biological variability is crucial to understand, not eliminate. Physically averaging together replicate RNAs reduces biological variability, but one also loses the ability to measure or assess that variability. Consider the following kinds of microarray studies.

1. A study to identify genes where the mean expression level is different in two populations.
2. A study to find a classification scheme for known disease classes based on gene expression measurements.
3. A study to discover unknown sub-classes of a disease.

For a study of type 1, a necessary component of making inference about population means is the population variances. But the data will lack the information to estimate the population variance if the RNA within each population is pooled. A statistical analysis can use measurement error to make inference about $\bar{X} - \bar{Y}$, the difference in sample means, but this is not the quantity of interest.

Kendziorski, Zhang, Lan and Attie (2003) propose an “in between” strategy, whereby multiple, independent pools would be used for each population. Such a strategy may allow a study to increase the number of individuals without increasing cost. Using multiple pools for each population retains some ability to estimate population variability. This kind of strategy is also known

as “composite sampling” (Boswell et al. (1996)). For a population with variability τ^2 , the variability of a pool of m individuals is τ^2/m . This approach sacrifices potential benefits that might come from learning more about two populations than just their difference in means, such as non-symmetries and multi-modalities. Moreover, there are other ways to define “differentially expressed” besides a difference in means (Pepe et al. (2003)). Still, the intermediate scheme of Kendzierski et al. (2003) may be a practical strategy for many investigations.

For studies like 2 and 3 above, pooling is generally inappropriate. Classification tools make predictions about individuals, and therefore require data at the individual level. If one hopes to discover new disease sub-classes, then clearly one cannot collapse data from individuals who are potentially from different sub-classes. Without distributional assumptions, pooling limits analysis to those such as 1 above.

5. Experimental layout

For any given gene a microarray makes a quantitative comparison between two RNAs. This makes a microarray design a block design with block size 2 (Kerr and Churchill (2001a)). Consequently, the experimental layout (how samples are paired onto arrays) is a crucial determinant of design efficiency. The layout also determines the confounding structure of the design. In this paper microarray designs appear as directed graphs (Kerr and Churchill (2001a), Yang and Speed (2002)), which quickly communicate the structure of a design and emphasize the comparative nature of a microarray assay. The nodes of the graph are the RNAs and the directed edges are the microarrays. Let the tail of an edge represent one dye and the head of an edge represent

the other dye. Thus an edge from node A to node B means an array is to be hybridized with RNA A labeled with dye 1 and RNA B labeled with dye 2.

5.1 *Efficiency*

As discussed, microarrays are a comparative measurement tool. When two differently-labeled RNAs are hybridized to the same array, the sample that contains more of a given transcript should produce proportionately higher signal (properly normalized) in the corresponding spots. There are various equivalent ways to formulate this as ANOVA or other linear models (Kerr (2003)). The outstanding issue in applying any of these models is the sources of random variation they include.

Kerr and Churchill (2001a) evaluated the efficiency of various microarray designs, including the popular “reference” design. In the reference design, a (usually) extraneous “reference sample” is used along with the RNAs of interest. Every sample of interest is compared in a hybridization to this sample. The design is intuitive: every RNA of interest can be compared indirectly because each is compared directly to the reference. For example, if the expression level of a gene is twice as high in RNA 1 compared to the reference, and six times as high in RNA 2 compared to the reference, then the expression is three times higher in RNA 2 compared to RNA 1. The same, simple logic can be applied to more sophisticated designs.

Alternative designs can have substantial advantages over the reference design (Churchill and Oliver (2001), Jin et al. (2001), Kerr and Churchill (2001a), Yang and Speed (2002)). Kerr and Churchill (2001a) considered microarray designs in a similar framework as in classical block design. Specifically, the viewpoint was that the individual RNAs were of interest in them-

selves, not as a sample from a population of interest. As discussed in Section 3 and 4, this is appropriate for some settings. However, if the RNAs are sampled from populations to identify genes that are differentially expressed between the populations, then the individual RNAs are no longer the primary objects of interest. Instead, these RNAs are studied for the purpose of making inferences about the populations from which they were sampled. Classical block design does not usually consider this situation.

{ Figure 1 about here. }

Suppose we want to compare two treatments (such as a treatment and a control) on a population. We obtain $2n$ individuals and randomly divide them equally among the treatment groups. Let us consider the merits of 3 design strategies. One option is to use a reference design (Figure 1a), comparing the RNA of each individual against some reference RNA. A second option is to use a loop design, alternating individuals from each group within the loop (Figure 1b). A final option is to pair the samples, comparing a sample of each type in separate dye-swap assays (Figure 1c).

In models with a single source of error (e.g. Kerr and Churchill (2001a)), the relative efficiency for comparing the means of the two groups is always $\frac{1}{4}$ favoring the loop and multi-dye-swap designs over the reference design. That is, the variance of the difference in means between the two groups is four times larger with the reference design. However, such models are not appropriate here because the assayed RNAs are samples. Instead, to consider design efficiencies our model has two sources of random effects:

$$y_{ikl} = \mu + A_i + V_k + \nu_{kl} + \epsilon_{ikl}, \quad (2)$$

where y_{ikl} is the normalized log fluorescence of the l^{th} replicate from variety

k measured on the i^{th} array. In (2), μ is the overall signal from the gene, A_i is the array effect of the i^{th} array (effectively, the “spot” effect), and V_k is the mean signal from “variety” $k = 1, 2$, i.e. the unknown population means. These are all fixed effects. The random effects are the ν_{kl} , which represent the variation within varieties $k = 1, 2$, and the measurement error ϵ_{ikl} . We have such a model for every gene. For design considerations, the only consequential difference between this model and the global ANOVA models of Kerr and Churchill (2001a) is the addition of the ν_{kl} (Kerr (2003)). While (2) does not contain dye effects, such effects are not estimable for the reference design (Figure 1a) due to confounding and do not affect the efficiency of the other two designs (Figures 1b and 1c) because their layouts are balanced with respect to dyes.

Say $\text{var}(\epsilon_{ikl}) = \sigma^2$ is the error variance and $\text{var}(\nu_{kl}) = \tau^2$ is the population variability. The ν_{kl} term represents the random variation of the l^{th} individual from the mean of population k . These random effects induce correlations among repeated measurements on individuals. The parameters of interest are the population means V_k .

An outstanding issue in microarray analysis is whether to consider the array effects, A_i , as fixed or random. Unlike the question of whether to use a global or gene-specific model, this decision can have a substantial effect on the results (Kerr (2003)). There is a compelling argument for treating these effects as random, as in Jin et al. (2001) and Wolfinger et al. (2001). However, incorporating this assumption into an analysis involves questionable distributional assumptions. Here, these effects are initially considered fixed, but this issue is re-visited at the end of this subsection.

Consider first the reference strategy (Figure 1a). The least-squares estimate of $V_1 - V_2$ is the usual estimate: for each array take the log ratio of treated vs. reference sample, and subtract the averaged log ratios for variety 2 from the averaged log ratios for variety 1. For a sample of size n from each population, the variance of $\hat{V}_1 - \hat{V}_2$ is

$$\frac{1}{n}(4\sigma^2 + 2\tau^2). \quad (3)$$

Consider next the loop strategy in Figure 1b. This is a balanced design, in that every array contains both variety 1 and variety 2. Each individual RNA is measured twice, although these repeated measurements are correlated. The variance of the least-squares estimate $\hat{V}_1 - \hat{V}_2$ is

$$\frac{1}{n}(\sigma^2 + 2\tau^2). \quad (4)$$

Note (4) is always smaller than (3). However, the gain in precision for using the loop design over the reference design depends on the relative sizes of σ^2 and τ^2 . Across the genes on the array, there will be a range of variances τ^2 . For genes with large biological variation such that $\tau^2 \gg \sigma^2$, the efficiency advantage for using the loop design is diminished. Kerr and Churchill (2001a) showed that loops become inefficient for large numbers of RNAs when the different RNAs are treated as different varieties. Loops retain their efficiency in the current context because the number of varieties is fixed at 2, regardless of the total number of RNAs.

We improve our precision with the alternating loop strategy, but at the cost of adopting a more complicated design and analysis. Is this added complication necessary? Consider the multiple dye-swapping strategy (Figure 1c). For a single dye-swap, the variance of the estimate of $V_1 - V_2$ is $\sigma^2 + 2\tau^2$.

With n individuals for each variety there are n dye-swaps, and the variance of the combined estimate is then

$$\frac{1}{n}(\sigma^2 + 2\tau^2), \quad (5)$$

the same as for the loop design. Although it yields the same precision as the alternating-loop strategy, the multi-dye-swap is much more robust (see Section 5.2), less complicated to execute and model, and likely offers greater opportunity for non-parametric analyses.

We return to the question of random spot effects. The variances at (4) and (5) do not change when spot effects are treated as random because of the balance in the design – each array has one sample from each variety. The same does not hold for the reference design. Letting $\text{var}(A_i) = \alpha^2$ in the model (2) with random A_i , the variance at (3) becomes

$$\frac{2}{n} \left[\frac{\sigma^4 + 2\sigma^2\alpha^2 + \tau^2(\sigma^2 + \alpha^2)}{\alpha^2 + \sigma^2} \right] \quad (6)$$

As $\alpha^2 \rightarrow \infty$, (6) converges to (3) since $\hat{V}_1 - \hat{V}_2$ estimated with random spot effects converge to the estimate with fixed spot effects. As $\alpha^2 \rightarrow 0$, (6) converges to $\frac{1}{n}(2\sigma^2 + 2\tau^2)$, which remains larger than (4) and (5). This makes sense, because when there is no spot-to-spot variation, it is clearly a waste to expend resources on a reference sample that is not of interest.

The general lesson here is that experimental layout can be an important factor in reducing error due to technical variability. But if population variability is large there is a smaller advantage in using an alternative design than the reference design. On the other hand, using a well-designed experiment along with the pooling strategy of Kendzioriski et al. (2003) (described in Section 4) may be an effective strategy.

5.2 *Robustness*

Robustness is an additional design consideration. This paper takes robustness in microarray design to mean the relative efficiency of the effective design if there are missing data due to failed arrays or bad spots in the intended design. The loop design (Figure 1b) is not robust. If a loop design is planned, but data are not obtained from some array, the resulting design has greatly reduced efficiency. In contrast, the reference design (Figure 1a) is robust. A reference design remains a reference design if an array is lost, and the loss only affects comparisons with the sample on that array.

{ Figure 2 about here. }

Loop designs should be avoided in experimental situations where array hybridizations sometimes fail and cannot be replaced. However, variations on loops are very robust designs. Compare the “double reference” and “double loop” designs in Figure 2. These two designs each use $2n$ arrays to study n samples. For 5 samples, if we consider a pairwise comparison between any pair of samples and only model measurement error (i.e., remove the ν term from (2)) the relative efficiency of these designs is 40% in favor of the double loop. In addition to its advantages in estimation precision, the double loop design is more robust. This is because there are many more connections between every pair of samples. In contrast, the loss of a single array in the double reference design means half the data on a particular RNA are lost. For five RNAs the double loop is superior in both efficiency and robustness.

5.3 *Dye-bias*

Some researchers have observed gene-specific dye-biases in their microarray data. Biologists sometimes report the phenomenon as genes for which

“the ratios don’t flip.” This bias is easily seen in simple “dye-swap” experiments, in which two samples are hybridized onto two arrays, and the dye-labels are switched in the two hybridizations. After normalizing for other sources of variation, including overall differences in the dyes, one expects the Green/Red ratio from Array 1 to be about the same as the Red/Green ratio from Array 2 for any particular gene. For a subset of genes, instead one finds that Green/Red from Array 1 is about the same as Green/Red from Array 2. It has been argued that this phenomenon is unimportant because it is a negligible part of the total variation in the data (Tseng et al. (2001)). However, ultimately inferences will be made for individual genes, and for individual genes this can be a substantial source of bias (Kerr et al. (2002)). However, dye-bias can be handled by extending the model at (2):

$$y_{ijkl} = \mu + A_i + D_j + V_k + \nu_{kl} + \epsilon_{ikl}. \quad (7)$$

The additional terms D_j are the “dye effect” for dye $j = 1, 2$.

The source of this dye-bias is currently under investigation. Fortunately, it is straightforward to protect against being misled by dye bias. A simple solution is to dye-swap every assay. An example of such a design is the “double reference” design in Figure 2(a). While this strategy is effective, it can be expensive. In fact, any design that is “even” handles the potential problem (Kerr and Churchill (2001a)). An even design is one where every sample is labeled with both dyes, and each differently-labeled sub-sample is used equally often in the experimental layout. This balance makes dye effects and gene expression effects orthogonal. The designs in Figures 1b, 1c, 2a, 2b, 3a, and 3b are all even. In contrast, in a confounded design such as the reference design (Figure 1a), dye biases cannot be corrected or detected.

This may not be a problem when the reference sample is not of interest, since the dye-bias will be the same in all RNAs of interest. On the other hand, if the reference sample is of scientific interest then, in a simple reference design, one cannot know whether results are biased.

5.4 *Practical considerations*

In addition to efficiency and robustness, there may be other practical considerations in choosing an experimental layout. These include

- (a) Simplicity
- (b) Extendability
- (c) Useful sub-designs

Simplicity (a) may be important for a large study in which many technicians will perform the assays. If a study is somewhat open-ended, a design that is easily extendable (b) may be preferred, so that additional samples can be added to the design in a sensible way. “Double reference” (Figure 2a) and “symmetric reference” (Figure 3) designs are natural to extend. Finally, it may be desirable for a design to contain useful sub-designs (c). If a question of interest applies to a subset of samples, a good sub-design will allow the question to be studied by analyzing only a subset of the data.

{ Figure 3 about here. }

6. **Summary**

The two key design issues in a microarray study are usually (1) which RNA samples will be assayed, and (2) which experimental layout will be used. The scientific question of interest should drive the choices in each case.

Selecting the RNA samples involves ensuring a study involves appropriate replication. If a study does not use biological replicates (type (C) in Section 3) or if all biological replicates are pooled, biological variability cannot be assessed and often the desired inferences cannot be made. Certain overarching goals of the study, such as finding a classification scheme, may not be possible without replication. Technical replication ((A) and (B) in Section 3), can be useful for increasing the precision of estimates. But technical replication is always less effective and can never substitute for true replication.

A spotted microarray is effectively a comparison between two samples. Because of this, the way samples are paired onto arrays can have a large impact on how effectively one can make the comparisons of interest at the end of the study (Kerr and Churchill (2001a)). In the author's experience, the number of microarrays budgeted for an experiment is at most twice the number of RNAs. For studies that include biological replicates for the purpose of inferring differences in the mean expression between groups, the advantage of using alternative designs to the reference design will be minor when biological variation is large. A simple design strategy such as the double reference (Figure 2a) is then very practical. In other kinds of studies, the individual RNAs will be of interest, such as pilot studies with no biological replicates. Generally, there is less advantage in efficiency in using a complicated design as the number of RNAs increases. Conversely, for smaller studies the advantage of a good layout is greater; Kerr and Churchill (2001a) give some guidance here. The general rule of thumb is that samples to be compared should be "close" in the design.

The information content of a dataset is determined by the design of the

experiment that produced it, regardless of the particular data values. Once a dataset is collected, its information content cannot be increased by any amount of ingenuity expended by a data analyst (Fisher (1971)). Good design is crucial to all scientific experimentation, and microarrays are no exception.

ACKNOWLEDGEMENTS

This work was supported by Career Development Funds from the University of Washington Department of Biostatistics. The author thanks Steve Self at the Fred Hutchinson Cancer Research Center, Biometrics co-Editor Brian Cullis, and three anonymous reviewers for comments that improved this paper substantially.

REFERENCES

- Boswell, M. T., Gore, S. D., Lovison, G. and Patil, G. P. (1996). Annotated bibliography of composite sampling part A: 1936-92. *Environmental and Ecological Statistics* **3**, 1–50.
- Churchill, G. A. and Oliver, B. (2001). Sex, flies and microarrays. *Nature Genetics* **29**, 355–356.
- Cui, X., Kerr, M. K. and Churchill, G. A. (2003). Data transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology*, **2**, Article 4. <http://www.bepress.com/sagmb/vol2/iss1/art4>.
- Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimina-

- tion methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.
- Fisher, R. A. (1971). *The Design of Experiments, 8th edition*. Hafner Press, New York, NY, USA.
- Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics* **29**, 389–395.
- Kendzioriski, C. M., Zhang, Y., Lan, H. and Attie, A. D. (2003). The efficiency of pooling mRNA in microarray experiments. *Biostatistics* **4**, 465–477.
- Kerr, M. K. (2003). Linear models for microarray data analysis: Hidden similarities and differences. *Journal of Computational Biology*, to appear.
- Kerr, M. K. and Churchill, G. A. (2001a). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.
- Kerr, M. K. and Churchill, G. A. (2001b). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123–128.
- Kerr, M. K., Leiter, E. H., Picard, L. and Churchill, G. A. (2002). Sources of variation in microarray experiments. In Zhang, W. and Smulevich, I., editors, *Computational and Statistical Approaches to Genetics*, pages 41–51. Kluwer Academic Publishers, Norwell, MA, USA.
- Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica* **12**, 61–86.
- Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics (Supplement)* **21**, 20–24.

- McIntyre, G. A. (1955). Design and analysis of two-phase experiments. *Biometrics* **11**, 324–334.
- Milliken, G. A. and Johnson, D. E. (1994). *Analysis of Messy Data, Volume I. Designed Experiments*. Wadsworth, Inc., Belmont, CA, USA.
- Nguyen, D. V., Arpat, A. B., Wang, N. and Carroll, R. J. (2002). DNA microarray experiments: Biological and technological aspects. *Biometrics* **58**, 701–717.
- Pepe, M. S., Longton, G., Anderson, G. and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133–142.
- Schena, M., Shalon, D., Davis, R. and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.
- Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.
- Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. 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.
- Wolfinger, R. D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust com-

posite method addressing single and multiple slide systematic variation.

Nucleic Acids Research **30**, e15.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews* **3**, 579–588.



Figure Captions

Figure 1 Designs for n individuals sampled from two populations using $2n$ arrays. Here $n = 4$. See Kerr and Churchill (2001a), Yang and Speed (2002, Box 2), or Section 5 for information about this representation of designs. The nodes of the graphs represent different individuals and the circles and triangles distinguish the two populations. The edges of the graphs represent the microarrays. (a) “Reference” design; the rectangle represents the reference RNA; (b) “Alternating loop” design; (c) Multiple dye-swap design.

Figure 2 Designs for n individuals sampled from two populations using $4n$ arrays. The circles and triangles distinguish the two populations. (a) “Double reference” design; (b) “Double loop” design.

Figure 3 “Symmetric reference” designs are more robust than loop designs and more efficient than reference designs. The design extend naturally to include additional RNAs. These designs use $n + 2$ microarrays for n RNA samples. The design is shown for (a) 10 samples and (b) 13 samples.



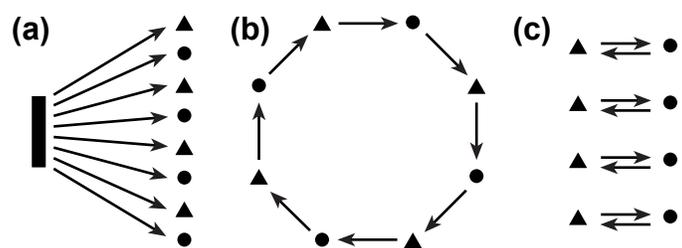


Figure 1

Kerr
Page 23

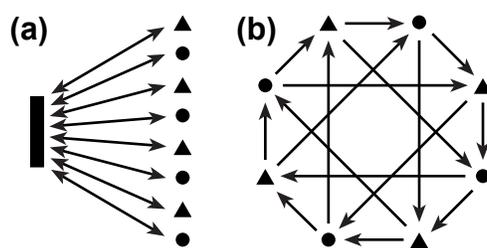


Figure 2
Kerr
Page 24

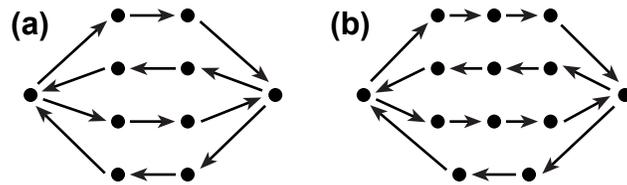


Figure 3
Kerr
Page 25