

# Experimental design for gene expression microarrays

M. KATHLEEN KERR, GARY A. CHURCHILL\*

*The Jackson Laboratory, Bar Harbor, ME, USA*

*Email: garyc@jax.org*

## SUMMARY

We examine experimental design issues arising with gene expression microarray technology. Microarray experiments have multiple sources of variation, and experimental plans should ensure that effects of interest are not confounded with ancillary effects. A commonly used design is shown to violate this principle and to be generally inefficient. We explore the connection between microarray designs and classical block design and use a family of ANOVA models as a guide to choosing a design. We combine principles of good design and A-optimality to give a general set of recommendations for design with microarrays. These recommendations are illustrated in detail for one kind of experimental objective, where we also give the results of a computer search for good designs.

*Keywords:* A-optimality; Confounding; Connected design; Even graph; Incomplete block design; Robust design.

## 1. INTRODUCTION

Geneticists are very interested in comparing the relative quantities of mRNA sequences in cell populations. Spotted cDNA microarrays (Brown and Botstein, 1999) are emerging as a powerful and cost-effective tool for quantifying gene transcription for thousands of genes at a time. In the first step of the technique, samples of DNA clones with known sequence content are spotted and immobilized onto a glass slide or other substrate, the 'microarray'. Next, pools of purified mRNA from cell populations under study are reverse-transcribed into cDNA and labelled with one of two fluorescent dyes, 'red' and 'green'. Two pools of differentially labelled cDNA are combined and applied to a microarray. Strands of cDNA in the pool hybridize to complementary sequences on the array and any unhybridized cDNA is washed off. Although hybridization efficiency can vary from clone to clone, the efficiency for any particular clone should not be affected by the type of the dye label. The 'red' and 'green' signals from a spot indicate the relative abundance of the corresponding mRNA in the two cell populations.

Some of the first experiments with microarrays were time-series studies. DeRisi *et al.* (1997) studied gene expression patterns in yeast during metabolic shift from fermentation to respiration. Chu *et al.* (1998) conducted a similar study of yeast during sporulation. The approach of this research was to cluster genes according to their patterns of expression over timepoints of a biological process. The general idea is that when a gene of unknown function ends up in a cluster of genes with known function, one has a valuable clue as to the function of the unknown gene. Clustering ideas have similarly been used to classify tissue samples according to their global patterns of gene expression. For example, Perou *et al.* (1999) used gene expression patterns to classify human breast cancers. Ross *et al.* (2000) studied gene expression variation in 60 cancer cell lines and found associations between gene expression patterns as well as other properties such as growth rate. Alizadeh *et al.* (2000) used this approach to identify clinically relevant subtypes of B cell lymphoma.

\*To whom correspondence should be addressed.

These experiments are just the beginning of the projected use of microarray technology. For example, Alon *et al.* (1999) used a related technology to make paired comparisons of cancerous tissue samples versus normal surrounding tissues. Microarray experiments will soon become multi-factorial in nature. For example, a researcher may want to study tissue samples from male and female mice from different strain backgrounds raised on different diets. It is easy to imagine a rich variety of experimental scenarios and substantial effort will be required to develop tools for higher-order analyses of microarray data. Following the precedent of the leading experiments, there have been many new ideas proposed about the best way to cluster genes (Ben-Dor *et al.* (1999), Eisen *et al.* (1998), Heyer *et al.* (1999), Lazzeroni and Owen (2000), and Tamayo *et al.* (1999) to name a few). Yet we believe that some fundamental questions still lack satisfactory answers.

The sources of variation in microarray data are yet to be completely understood. To the extent that sources of variation are known, however, they should be considered in the design and analysis of microarray experiments. The structure of microarray data, the types of analyses that are possible, and the quality of the results are determined by the experimental design. We believe there has been a lack of healthy scepticism about the 'right' way to design a microarray experiment, and that this is an area that deserves careful consideration and study.

Different cDNAs are known to incorporate dye with differential efficiency and hybridize with their target spots on arrays at different rates. Further, with spotted arrays it is not known how much DNA is immobilized on the array in any particular spot. Therefore, as scientists have recognized, a single fluorescent intensity measurement from a spot contains little useful information because of the unknown characteristics of the spot and the unknown interpretation of a unit of fluorescence for any particular gene. This realization undoubtedly motivated the two-dye system and the practice of calculating the ratio of the pair of readings from a spot. There is meaningful information in the relative red and green intensities from a spot.

Now consider an experiment from the archives of statistics. If an agriculturalist wants to measure the yields of strains of corn, s/he would realize that different plots of land vary in soil fertility, amount of rainfall and sunlight, etc., so the only meaningful direct yield comparisons are for strains grown on the same plot of land. These twentieth century agricultural experiments share an important characteristic with twenty-first century microarray experiments: the meaningful interpretation of the data is in terms of relative comparisons. We believe there are valuable lessons to be learned from the several generations of scientists and statisticians who studied experimental designs for agriculture. In this work, we explore some of the connections between classical experimental design and microarray technology.

The cell populations under study are the factor of interest in a microarray experiment, but they are not the only sources of variation. The design of microarray experiments—how the samples are paired onto arrays—should take this into account. Section 2 identifies the experimental design factors involved with this technology. To illustrate basic design ideas, a commonly used setup for microarray experiments is studied in Section 3 and an alternative is proposed. We introduce a family of ANOVA models, explore more examples, and give general design recommendations in Section 4. In Section 5 we consider general A-optimality and generalize classical results from experimental design to microarrays. In Section 6 we discuss a search for good designs for small ( $\leq 10$ ) numbers of samples when one wants efficiency with respect to general A-optimality but also requires certain model-robustness properties. Section 7 concludes with a discussion of open questions for microarray experimental design.

## 2. SOURCES OF VARIATION IN MICROARRAY EXPERIMENTS

The simplest microarray experiment looks for changes in gene expression across a single factor of interest. This factor might be the timepoints of a biological process, or different types of tissue, or drug treatments. We generically call the categories of a factor of interest *varieties*. Fluorescent intensities

clearly also depend on the cDNA sequence spotted on the arrays. We call the spotted sequences ‘genes’ whether they are actually genes, ESTs, or DNA from another source. Further, microarray technology makes use of two different dyes and an entire experiment uses multiple arrays. Therefore, we identify four basic experimental factors: varieties, genes, dyes, and arrays.

With these four factors there are  $2^4 = 16$  possible experimental effects. Explicitly, there is the mean or baseline effect, four factor main effects for arrays ( $A$ ), dyes ( $D$ ), varieties ( $V$ ), and genes ( $G$ ), six two-factor interactions, four three-factor interactions, and one four-factor interaction. The first step in choosing a good design is to identify which effects might possibly contribute to variation in the data.

Array main effects measure overall variation in fluorescent signal from array to array. These effects arise if, for example, arrays are probed under inconsistent conditions that increase or reduce hybridization efficiencies of labelled cDNA. Dye main effects measure differences in the two dye fluorescent labels. For example, one dye may be consistently ‘brighter’ than the other. Gene main effects occur when certain genes emit a higher or lower fluorescent signal overall, compared to other genes. These effects arise because some genes have generally higher or lower levels of expression than others, and also because of differential hybridization efficiency and differential labelling efficiency for different sequences. Variety main effects occur when the varieties of the factor of interest have higher or lower overall expression levels for the genes spotted on the arrays. It is reasonable to suspect that all four of these main effects will contribute to variation in microarray data.

For a particular tissue sample, red- and green-labelled cDNA is produced in separate runs of the reverse-transcription process. Differences in the runs can produce pools of cDNA of varying concentrations or quality. This results in experimental dye  $\times$  variety ( $DV$ ) interactions. Array  $\times$  gene interactions ( $AG$ ) occur because spots for a given gene on the different arrays vary in the amount of cDNA available for hybridization. Thus  $AG$  effects are the ‘spot’ effects. By considering  $AG$  effects, we take the approach of treating each spot as a unique entity. Alternative strategies might try to capture spot characteristics by, for example, modelling spot density or properties of pin groups. Here we wish to work in the most general setting and not to assume spot characteristics can be successfully modelled. We attempt to make the most out of the two-dye system and the fact that there are two readings per spot. Dye  $\times$  gene effects ( $DG$ ) arise if there are differences in the dyes that are gene specific. For example, if the overall efficiency of incorporation of red dye is higher than that of green dye except for a small subset of genes for which the reverse is true, this would be captured in the  $DG$  effects. Although we did not anticipate such effects, we have seen one extreme case in practice. In this case, spots for a particular sequence on two arrays were consistently green despite the fact that we had reversed the labelling on two identical cDNA samples. In other words, we applied sample 1 labelled red and sample 2 labelled green to one array and sample 1 labelled green and sample 2 labelled red to a second array, and the spots for a particular sequence emitted higher fluorescent intensity for the green dye on both arrays. At this time we have no explanation for this phenomenon. However, had we run only one array the gene would have appeared as an interesting and significant result rather than an experimental anomaly.

Variety  $\times$  gene interactions ( $VG$ ) reflect differences in expression for particular variety and gene combinations that are not explained by the average effects of those varieties and genes. These are the effects of interest. Identifying genes whose expression changes in different varieties means identifying non-zero differences in  $VG$  effects.

Of the remaining interactions, three do not involve  $G$ : array  $\times$  dye ( $AD$ ), array  $\times$  variety ( $AV$ ), and array  $\times$  dye  $\times$  variety ( $ADV$ ). It is difficult to relate any of these to the process underlying microarrays and to suppose a reason why such interactions would come into play. On the other hand, let us assume we account for all factor main effects. Then since  $AD$ ,  $AV$ , and  $ADV$  are not gene specific, including or excluding any of them in the analysis of microarray data does not change the estimates of the effects of interest ( $VG$ ), so the question is academic. The rest of the interactions are three- and four-way effects involving  $G$ : array  $\times$  dye  $\times$  gene ( $ADG$ ), array  $\times$  variety  $\times$  gene ( $AVG$ ), dye  $\times$  variety  $\times$  gene ( $DVG$ ),

and array  $\times$  dye  $\times$  variety  $\times$  gene (*ADVG*). The presence of such interactions would mean there is gene-specific variation attributable to a particular array and dye, a particular array and variety, a particular dye and variety, or a particular array, dye, and variety combination. Again, these high-order interactions are difficult to relate to the physical and chemical processes that make up this technology. Therefore, in the remainder of this paper we concentrate on effects with understandable interpretations. A secondary justification for this is that, ultimately, the value of this technology will be judged by whether reproducible results can be obtained. At a fundamental level, it is not clear that results can be reproduced if every reading depends on high-order interactions between genes and other factors in the experiment.

### 3. MICROARRAY DESIGNS

Each microarray in an experiment is probed with two differently labelled cDNA samples. Informally, we say each array ‘contains’ two samples. Arrays are the experimental *blocks* with block size two. When there are more than two varieties of the factor of interest, not every variety can appear on every array. Therefore, the experimental design is an *incomplete block design*. In this paper we primarily consider *binary* designs, meaning a variety appears 0 or 1 times on any array.

We use directed graphs to describe designs. Nodes represent the varieties and edges represent arrays. The direction of edges gives information about dyes. We arbitrarily but permanently assign one dye to the tail of directed edges and the other dye to the head. An edge from variety A to variety B represents an array containing variety A labelled with the ‘tail’ dye and variety B labelled with the ‘head’ dye. Graphical design illustrations have two advantages. First, they quickly and clearly communicate the setup of a design. Second, they allow one to easily evaluate certain design properties. As a bonus, they sometimes suggest design names.

We open our discussion of design for microarrays by looking at a design commonly used in practice. We then propose an alternative design that uses the same number of arrays. Comparing these designs illustrates some of the issues that arise with microarray experiments.

#### 3.1 The ‘reference’ design

Figure 1(a) shows a commonly used design for studying  $v$  varieties of a factor of interest. We have dubbed this the ‘reference’ design because an additional  $(v + 1)^{\text{st}}$  variety is introduced to serve as a reference. Practitioners of this strategy use one dye to label the reference variety and the other dye to label the varieties of interest. This means that variety effects are completely confounded with dye effects. Consequently, the effects of interest,  $VG$ , are completely confounded with  $DG$  effects. In order to use this design, one must assume there are no gene-specific dye effects. Note that the most information is collected on the reference variety, although this is not a variety of interest.

The reference design for  $v$  varieties of interest and  $n$  genes produces  $2vn$  observations. The mean and the array, variety, and gene main effects account for  $2v + (n - 1)$  degrees of freedom. The effects of interest,  $VG$ , account for  $v(n - 1)$  degrees of freedom. If  $AG$  effects must be accounted for, they comprise the final  $(v - 1)(n - 1)$  degrees of freedom. No degrees of freedom remain to estimate error.

It is easy to see how a scientist would arrive at the reference design. Biologists recognize there is variation in the amount of cDNA from spot to spot and so fluorescent intensities are only meaningful in a relative sense. This variation is ‘controlled’ by always having the same reference in each spot. Using classical experimental design ideas, however, it is possible to obtain better results with the same number of arrays.

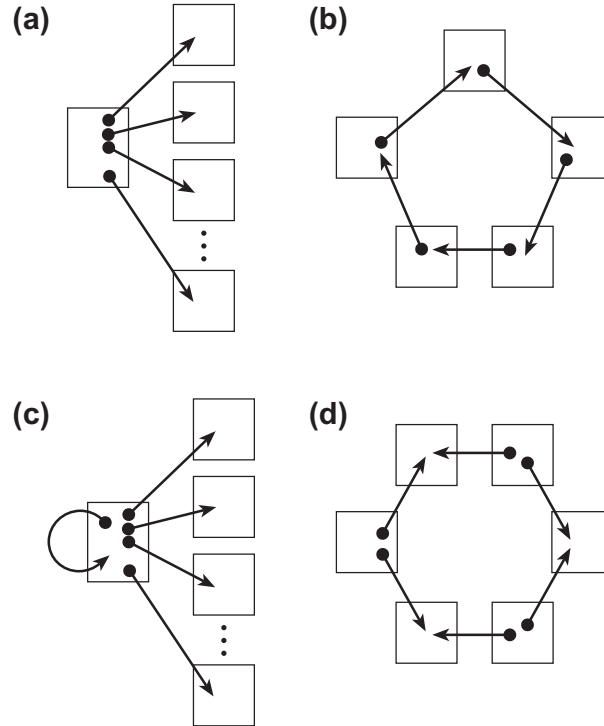


Fig. 1. (a) Reference design; (b) loop design for five varieties; (c) augmented reference design; (d) modified loop design for six varieties.

### 3.2 The 'loop' design

As an alternative to the reference design, we propose the *loop* design, shown in Figure 1(b) for  $v = 5$  varieties. Using the same number of arrays as the reference design, the loop collects twice as much data on the varieties of interest. Further, notice that varieties are balanced with respect to dyes because each variety is labelled once with the red and green dyes. This balance means that dye effects are unconfounded with variety effects, so  $VG$  effects are unconfounded with  $DG$  effects. Thus any anomalous behaviour of genes with respect to dyes, as described in Section 2, will not bias the estimates of the effects of interest.

If one estimates all factor main effects and, in addition, the  $VG$  and  $AG$  interactions, then  $n - 1$  degrees of freedom remain. These degrees of freedom provide information to estimate error variation. Therefore, using this design provides a basis for statistical inference. This puts the loop design in an arena where the reference design cannot compete.

A practical drawback of the loop design is that each sample must be labelled with both the red and green dyes, effectively doubling the number of labelling reactions. Because microarray technology is new and not yet fully understood, our opinion is that this extra effort is worthwhile. Balancing varieties with respect to dyes produces data in which  $DG$  effects can be detected. If one is unable to complete the extra work in dye labelling, then one must be willing to accept the assumption that there are no gene-specific dye effects. In that case, however, it is still possible to get the other benefits of the loop design. Figure 1(d) shows a loop for six varieties in which there is only one dye labelling for each variety. When the number

of varieties is even this strategy actually has one fewer labelling than the reference design because it does away with the reference sample. While this design does not have the orthogonality of variety and dye effects, it retains the other advantages of the loop design, namely collecting more data on the varieties of interest and providing degrees of freedom for estimating error.

#### 4. EVALUATING MICROARRAY DESIGNS

##### 4.1 Models and assumptions

We assume that there exists a transformation of microarray data on which the effects are additive, such as the log scale (Kerr *et al.*, 2000). Using this scale, let  $y_{ijk}$  be the fluorescent intensity from array  $i$  and dye  $j$  representing variety  $k$  and gene  $g$ . We further assume that the same set of genes is spotted on each array in an experiment. This assumption means that a full replication of genes is present for every array, dye, and variety combination in any design. Therefore, gene effects are orthogonal to all effects of these factors. This effectively divides effects into two groups: ‘global’ effects, which only involve  $A$ ,  $D$ , and  $V$ , and gene-specific effects, which involve  $G$ . Because  $G$  is orthogonal to all of  $A$ ,  $D$ , and  $V$ , gene-specific effects are orthogonal to global effects. Note the effects of interest,  $VG$ , are gene specific.

An ANOVA model can thus be considered as having global and gene-specific components. Our models use just the  $A$ ,  $D$ , and  $V$  main effects in the global component. One might also want to consider  $DV$  interactions to account for variation in the dye labelling reaction. However, these are generally confounded with the main effects of  $A$ ,  $D$ ,  $V$  and so are indirectly accounted for. This is a case when confounding is advantageous. Because of confounding, we account for effects that are not of interest without using additional degrees of freedom (Cochran and Cox, 1992).

A simple ANOVA model includes only the factor main effects and the effects of interest,  $VG$ :

$$y_{ijk} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + \epsilon_{ijk}. \quad (4.1)$$

A more plausible model accounts for spot-to-spot variation by including  $AG$  effects:

$$y_{ijk} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + \epsilon_{ijk}. \quad (4.2)$$

Another possibility is to further account for genes interacting with dyes:

$$y_{ijk} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \epsilon_{ijk}. \quad (4.3)$$

We assume that there is independent, additive error  $\epsilon_{ijk} \sim F$ , where  $F$  is a distribution with mean 0 and variance  $\sigma^2$ . However, it is an open question whether some genes are inherently more ‘noisy’ than others. A more general assumption is  $\epsilon_{ijk} \sim F_g$ , with  $F_g$  having mean 0 and variance  $\sigma_g^2$ . With the more general assumption, weighted least squares should produce lower-variance estimates. With gene-based heteroscedasticity, questions arise about choosing the set of genes to study. However, for our design problem—arranging varieties onto arrays—not much changes. The quantities of interest are comparisons of varieties for fixed genes (e.g.,  $(VG)_{1g} - (VG)_{2g}$ ) and for any given gene the relative merits of different designs does not depend on  $\sigma_g^2$ . Therefore, for simplicity we continue with the assumption of a common variance  $\sigma^2$ , but note this generalization.

Finally, models (4.1)–(4.3) are set up for the situation in which each gene is spotted only once per array. Multiple spotting is clearly an option, so that a data value is  $y_{ijk_s}$  for the  $s$ th replicate of gene  $g$  from array  $i$  using dye  $j$  representing variety  $k$ . Statisticians generally advocate replication whenever possible, and we believe microarrays should be no exception. When genes are replicated, the  $(AG)_{ig}$  effects in (4.2) and (4.3) should be replaced with spot effects  $(AG)_{igs}$ . As one would expect, estimation

precision increases with more data. If genes are spotted  $m$  times per array, the variance of estimates such as  $\widehat{(VG)}_{1g} - \widehat{(VG)}_{2g}$  for gene-by-gene variety comparisons decreases by a factor of  $1/m$ . Thus, for a given level of replication, the relative efficiency of designs is the same. Therefore, we proceed with the case of no replication for simplicity.

In order to fit models such as (4.1)–(4.3) a design should be connected, i.e. the graphical representation of the design is a connected graph. This is a necessary condition for the  $A$  and  $V$  effects, and the  $AG$  and  $VG$  effects, to be jointly estimable.

#### 4.2 Contrasts of interest

Microarrays are useful for studying the relative expression of genes across samples. The effects of interest are the  $VG$  interactions. Specifically, the contrasts of interest are  $(VG)_{k_1g} - (VG)_{k_2g}$  for fixed genes  $g$  and pairs of varieties  $k_1 \neq k_2$ . The interesting variety pairs will depend on the objectives of the experiment.

Consider model (4.1) and a design in which variety  $k$  appears on  $r_k$  arrays. Under the assumptions that the same set of genes is spotted on every array,  $VG$  effects are orthogonal to all other effects in (4.1). The least-squares estimate of  $(VG)_{k_1g} - (VG)_{k_2g}$  is then  $y_{\cdot k_1 g} - y_{\cdot k_2 g} - (y_{\cdot k_1 \cdot} - y_{\cdot k_2 \cdot})$ . When there are  $n$  genes in the experiment, we have

$$\text{var}(y_{\cdot k_1 g} - y_{\cdot k_2 g} - (y_{\cdot k_1 \cdot} - y_{\cdot k_2 \cdot})) = \frac{n-1}{n} \left( \frac{1}{r_{k_1}} + \frac{1}{r_{k_2}} \right) \sigma^2.$$

The variance is solely a function of the  $r_k$ ,  $n$ , and  $\sigma^2$ ; it is unimportant how varieties are paired onto arrays.

In our analyses of microarray data we have found this model to be inadequate because of spot-to-spot variation on arrays. Models (4.2) and (4.3) account for this with  $AG$  effects. For these models, the functional form of the least-squares estimates  $\widehat{(VG)}_{k_1g} - \widehat{(VG)}_{k_2g}$  depends explicitly on the design because  $VG$  effects are partially confounded with  $AG$  effects. Estimators for model (4.2) are given in Appendix A for the reference design and Appendix B for the loop design. The partial confounding of  $AG$  and  $VG$  is an unavoidable consequence of the fact that when there are more than two varieties, not every variety can appear on every array.

By the nature of the technology, arrays are automatically balanced with respect to dyes, so  $DG$  effects are orthogonal to  $AG$  effects. Whether  $VG$  effects are orthogonal to  $DG$  effects depends on the design. A design balances varieties with respect to dyes when each variety is labelled with the red and green dyes equally often. Clearly, a necessary condition for balance is that every variety appears in the design an even number of times. The graphical representation of such designs have the property that the degree of every node is even. We call such designs *even*. Evenness is also a sufficient condition for balance to be possible. This fact follows from Euler's theorem that every even graph has a circuit that traverses every edge exactly once. By directing the edges in an even graph as an Eulerian circuit, every node ends up with the same number of 'heads' and 'tails' and thus varieties are balanced with respect to dyes. Properly directed even designs prevent confounding between  $VG$  and  $DG$  effects. If  $VG$  and  $DG$  effects are orthogonal, the problem of choosing a good design considering model (4.3) reduces to the problem considering model (4.2).

#### 4.3 Example: relative efficiency of loop and reference designs

The loop and reference designs use the same number of arrays to compare  $v$  varieties of a factor of interest (we assume the reference variety in the reference design is not a variety of interest). We first note that the  $D_j$  terms must be removed from any model for reference design data because this design completely confounds dye effects with variety effects. If model (4.1) suffices, then the variance of each pairwise

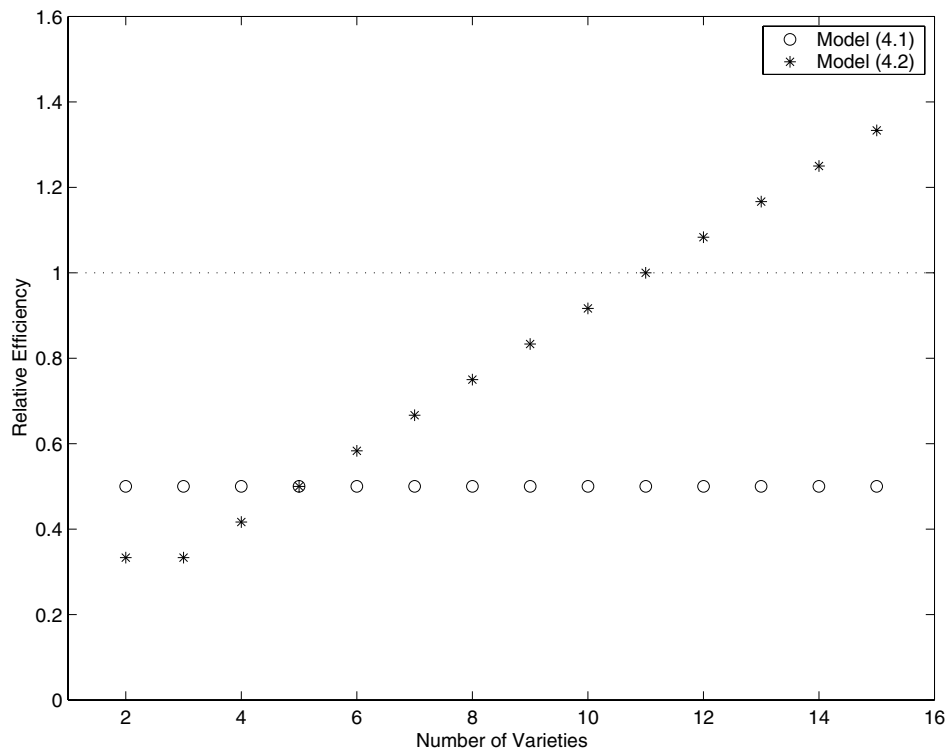


Fig. 2. Relative efficiency of reference and loop designs. The average variance over pairwise comparisons for the loop design is shown as a ratio relative to the average variance for the reference design. For model (4.1), the loop design is always twice as good because varieties are sampled twice as much. For model (4.2), the loop design has smaller average variance for the contrasts of interest than the reference design for  $v \leq 10$  varieties. The designs cannot be compared for model (4.3) because  $DG$  effects are not estimable with the reference design.

contrast  $(\widehat{VG})_{k_1g} - (\widehat{VG})_{k_2g}$ , where  $k_1$  and  $k_2$  are varieties of interest, is  $2\frac{n-1}{n}\sigma^2$  for the reference design and  $\frac{n-1}{n}\sigma^2$  for the loop design. In other words, the standard deviation for a contrast of interest is  $\sqrt{2}$  larger for the reference design.

If the larger model (4.2) is used, then the variance of each  $(\widehat{VG})_{k_1g} - (\widehat{VG})_{k_2g}$  using the reference design is  $4\frac{n-1}{n}\sigma^2$ . For the loop design the variance depends on  $v$  and the relative positions of  $k_1$  and  $k_2$  in the loop. Comparisons for varieties nearby in the loop are estimated more precisely than varieties that are far apart. One way to compare the designs is to average the variance over all pairs of varieties for the loop design. Figure 2 gives the results. We see that for  $v < 10$  the loop design does better than the reference design. Intuitively, loops are inefficient for large  $v$  because some pairs of varieties are too far apart.

Finally, we note that one cannot fit model (4.3) to the reference design because  $DG$  effects are completely confounded with  $VG$  effects. To use this design one is forced to assume there are no  $DG$  interactions. If they exist, they cannot be accounted for and  $VG$  estimates will be biased.

#### 4.4 Example: multiple treatments versus a control

Clearly it is not always the case that all variety comparisons are equally interesting—it depends on the experimental objectives. In experiments with a treatment/control structure it will be interesting to

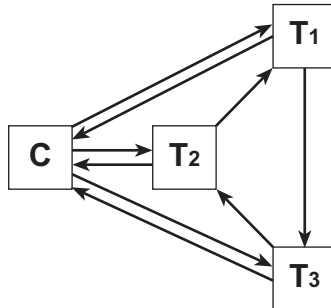


Fig. 3. A design for three treatments and a control.

compare treatment varieties with the control but not so important to have precise comparisons between the treatments. As a simple example, consider a study in which there is a control and three treatments. One idea is to use a design as in Figure 3. This is an example of a class of designs defined by Bechhofer and Tamhane (1981) called *balanced test-treatment incomplete block designs*. In some cases these designs are known to be optimal for making treatment-control comparisons (Hedayat and Majumdar, 1984). These results apply to microarrays when we consider model (4.2) because spots are incomplete blocks of size two (this connection will be described in more detail in Section 5). The design in Figure 3 has the additional property that varieties are balanced with respect to dyes, so  $VG$  effects are orthogonal to  $DG$  effects. This means the design is equally efficient for model (4.3) as for model (4.2).

#### 4.5 Example: time-course experiments

Some of the first experiments that demonstrated the power and potential of microarrays were time-course gene expression studies (DeRisi *et al.*, 1997; Chu *et al.*, 1998). A common setup for these experiments is similar to the reference design. Every timepoint is compared with time 0, including an array that contains only time 0 labelled with each dye. Figure 1(c) illustrates this design, which we call the *augmented reference design*. The additional ‘self comparison’ array in the augmented reference design means that  $DG$  effects are only partially confounded with  $VG$  effects, so it is possible to estimate model (4.3). Time-course studies have proven to be an important and useful class of experiments, and we believe that the design of these experiments deserves further consideration.

A reference or augmented reference design seems like a natural setup when there is a treatment/control structure to the varieties, but since this is not the case with time series we question whether alternative designs might be more appropriate. After all, these experiments seek expression profiles for genes over time, and it is not comparisons between every time and time 0 that are of any special interest. One alternative is to use a loop design and, for timepoints  $0, 1, \dots, v$ , to make the  $v$  comparisons of every timepoint to the previous rather than to time 0. This set of contrasts between adjacent timepoints should contain as much information about the expression pattern of a gene over time, but these contrasts can be estimated more precisely. Figure 4 shows the variance of the  $(VG)_{kg} - (VG)_{k-1,g}$  estimates from the loop design divided by the variance of the  $(VG)_{kg} - (VG)_{0g}$  estimates from the augmented reference design. We see that, for each of models (4.1)–(4.3), the estimates from the loop design have smaller variance. The inefficiency of the augmented reference design is most dramatic for (4.3) because this model accounts for  $DG$  effects. The partial confounding between  $DG$  and  $VG$  effects is almost complete with the augmented reference design, so correcting for  $DG$  effects dramatically decreases the efficiency of the design. We do

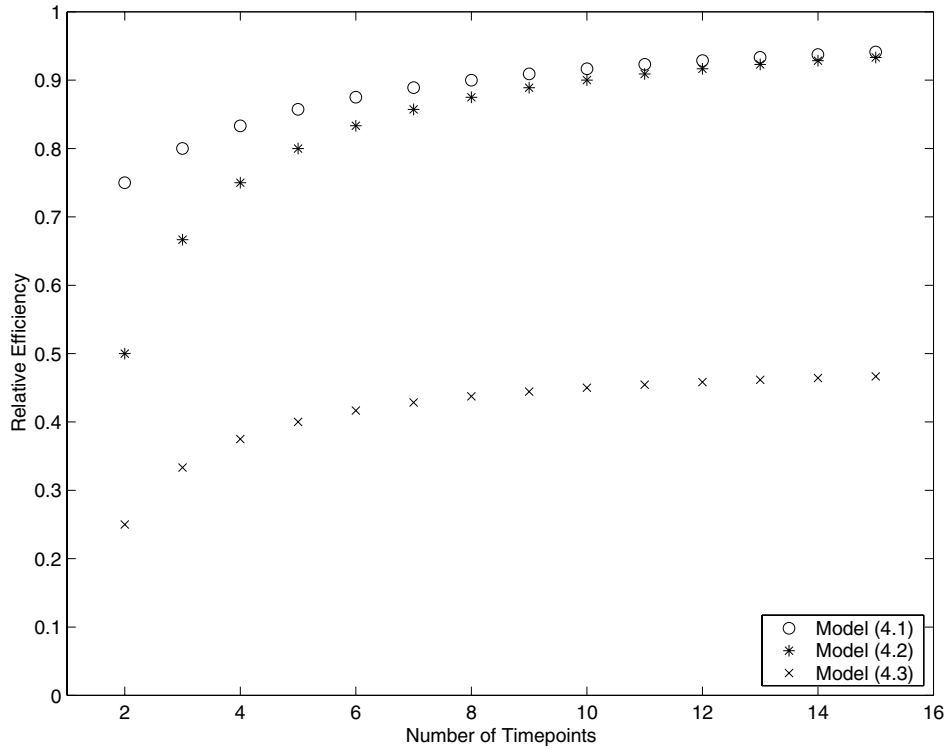


Fig. 4. Relative efficiency of two strategies for time-course experiments. The ratio of  $\text{var}(\widehat{VG}_{kg} - \widehat{VG}_{k-1,g})$  from the loop design over  $\text{var}(\widehat{VG}_{kg} - \widehat{VG}_{0g})$  for the augmented reference design, considered for models (4.1)–(4.3). For all three models, profiling genes using the loop design strategy leads to more precise results.

not claim the loop design is the best solution for time-course experiments. Rather, we only point out that alternatives exist and deserve consideration.

#### 4.6 Recommendations

We advocate choosing designs that are robust, in that they result in precise estimates of the quantities of interest regardless of the final model. Our general recommendations for microarray designs are:

- (1) Choose an even design so that varieties can be balanced with respect to dyes. This ensures that the effects of interest are not biased by genes interacting with dyes. This is especially important when genes are not replicated on arrays, because there may not be degrees of freedom to explicitly account for  $DG$  effects.
- (2) Among even designs, look for a design that is efficient for comparing gene expression across varieties while accounting for spot-to-spot variation. Section 5 gives some details about designs that have this property when all pairs of elementary contrasts are of equal interest. The basic principle is intuitive: varieties to be compared should be ‘close together’ in the design.
- (3) Most importantly, keep in mind the fundamental principles of good design: balance and replication. Balance ensures that the effects of interest are not confounded with other sources of variation.

Replication improves the precision of estimates and provides degrees of freedom for error estimation (Fisher, 1951).

### 5. GENERAL A-OPTIMALITY

To provide more detail about evaluating designs for one kind of experimental objective, in this section we suppose comparisons between all pairs of varieties are of equal interest. A reasonable criterion for evaluating designs in this case is A-optimality. This criterion favours designs that minimize

$$\frac{1}{\binom{v}{2}} \sum_{k_1 \neq k_2} \text{var}((\widehat{VG})_{k_1g} - (\widehat{VG})_{k_2g}), \tag{5.1}$$

the average variance of a contrast of interest.

Recall that with model (4.1), the important parameters of the design are  $r_1, \dots, r_k$ , the replication of varieties in the experiment. To minimize (5.1), one should replicate varieties to minimize

$$\sum_{k_1 \neq k_2} \left( \frac{1}{r_{k_1}} + \frac{1}{r_{k_2}} \right) \propto \sum_k \frac{1}{r_k}. \tag{5.2}$$

If  $b$  arrays are allotted to an experiment to study  $v$  varieties, (5.2) is minimized by choosing  $r_k = 2b/v$ , i.e. equal sampling of all varieties.

As mentioned, however, we have never found model (4.1) to be adequate because it does not contain  $AG$  effects and thus does not account for spot-to-spot variation. Including  $AG$  terms, as in model (4.2), puts us in the context of incomplete block design. In classical block design, the expected yield of an observation is  $y_{ik} = \mu + B_i + V_k$ , where  $B_i$  is the effect of block  $i$  and  $V_k$  is the effect of variety  $k$ . A well known result gives an alternative method to obtain the sum of the variances of elementary variety contrasts  $\hat{V}_{k_1} - \hat{V}_{k_2}$  (see (Raghavarao, 1971)). For  $v$  varieties appearing  $r_1, r_2, \dots, r_v$  times in a design with  $b$  blocks, let  $N$  be the  $v \times b$  incidence matrix of the design, where  $n_{ki}$  is the number of times the  $k$ th variety appears in the  $i$ th block. The  $v \times v$  matrix  $NN^t$  is known as the concurrence matrix of the design because its  $k_1, k_2$  off-diagonal entries are the number of times varieties  $k_1 \neq k_2$  occur together in the same block. The so-called  $C$ -matrix, sometimes called the information matrix of the design, is

$$C = \text{diag}[r_1, r_2, \dots, r_v] - \frac{1}{t} NN^t, \tag{5.3}$$

where  $t$  is the block size. The matrix  $C$  is always singular. Let  $0 = \mu_1 \leq \mu_2 \leq \dots \leq \mu_v$  be the eigenvalues of  $C$ . The variance of  $\hat{V}_{k_1} - \hat{V}_{k_2}$ , averaged over all pairs of varieties  $k_1 \neq k_2$ , is  $\frac{2\sigma^2}{v-1} \sum_{i=2}^v \frac{1}{\mu_i}$ . This average variance is finite if and only if  $\mu_2 > 0$ , which is true if and only if the design is connected.

This result applies to variety effects partially confounded with block effects. We are interested in  $VG$  interactions when they are partially confounded with  $AG$  interactions. Since  $VG$  interactions are orthogonal to all other effects in (4.2) aside from  $AG$  effects, the result generalizes directly (see Appendix C). Forming the  $C$ -matrix in exactly the same way and getting its eigenvalues  $\mu$ , the A-optimality criterion (5.1) becomes  $\frac{n-1}{n} \frac{2\sigma^2}{v-1} \sum_{i=2}^v \frac{1}{\mu_i}$ .

### 6. A SEARCH FOR GOOD DESIGNS

The most problematic aspect of our recommendations is of course item (2). This recommendation says to choose a good incomplete block design for block size two. This is a non-trivial problem that has been

studied extensively. In this section we give some efficient incomplete block designs when all pairwise variety differences are of equal interest.

The literature on incomplete block designs generally operates under the strategy of defining families of designs and attempting to show that designs in the family are optimal, or at least highly efficient. This is an important direction of research, but it is of limited use for the practical purposes of designing microarray experiments. For example, one of the strongest results is that a strongly regular graph design with singular concurrence matrix is A-optimal (Cheng and Bailey, 1991). A strongly regular graph design (Bose, 1963) is both a regular graph design (John and Mitchell, 1977) and a partially balanced incomplete block design with two-associate classes (PBIB(2); Bose and Nair (1939)). In Clatworthy's compilation of PBIB(2) designs (1973), for block size two this result covers only the semi-regular group-divisible designs for  $v = 2m$  varieties and  $v^2/4$  blocks (corresponding to bipartite graphs with two groups of vertices of size  $m$ ). These designs were already shown to be optimal by Cheng (1978). Moreover, these designs are even only if  $m$  is even, i.e. the number of varieties  $v$  is a multiple of 4.

Experimenters approach the design question from a different direction. They are less interested in knowing families of optimal designs than in learning the answer to a practical question: for the number of arrays budgeted for an experiment, which design should I use? Along the same lines, experimenters with some flexibility in the number of arrays to allot might ask how much could be gained by adding a few more arrays.

Every binary design for  $v$  varieties with  $b \leq \binom{v}{2}$  blocks of size two corresponds to a simple graph on  $v$  nodes. We can take advantage of this fact to do a search on all possible designs when  $v$  is not too large. We used the computer program 'makeg' (McKay, 1991) to generate full sets of non-isomorphic connected designs. For  $v \leq 10$  we searched over the set of all possible connected designs and recorded A-optimal designs. Because we have special interest in even designs, we also recorded the best even designs in our search. The result is definitive lists of optimal designs for  $v \leq 10$  and  $v \leq b \leq \binom{v}{2}$ .

Figure 5 contains plots of the A-optimality criterion (5.1) (ignoring the factor  $(n-1)\sigma^2/n$ ) for the best designs and the best even designs for  $v = 6, \dots, 10$  and  $v \leq b \leq \binom{v}{2}$ . The reference sample strategy uses  $v$  arrays and always yields an average variance of 4 regardless of  $v$ . This is noted with an 'R' on the plots. The only even design for  $v$  varieties and  $v$  arrays is the loop design, so trivially the loop is the best even design of its size. The loop is A-optimal when  $v \leq 8$ .

We note that there is a relatively small improvement in precision for using  $v+1$  arrays over  $v$  arrays, but a fairly substantial improvement in using  $v+2$  arrays over  $v+1$  arrays. From a practical point of view, we would encourage investigators planning to use  $v$  arrays for  $v$  varieties to budget the extra two arrays. This is particularly true for large  $v$  because loops become very inefficient and the relative cost of extra arrays is small. Figure 6 shows the best even designs for studying  $v$  varieties with  $v+2$  arrays. Experimenters able to use more arrays could use the plots to gauge what a sensible number might be. A catalogue of the resulting designs is given at <http://www.jax.org/research/churchill>.

There are 11 716 571 non-isomorphic connected graphs on 10 nodes. The corresponding counts for eleven and twelve nodes are 1006 700 565 and 164 059 830 476. Obviously a naive search of all possible designs becomes computationally infeasible for larger  $v$ . Hopefully, families of efficient designs will continue to be discovered. Another line of research could work out algorithms for constructing designs. Bagchi and Cheng (1993) give a method for constructing efficient, although not necessarily optimal, designs with block size two for any  $v$  by combining smaller, efficient designs. Their method produces designs in which each variety appears  $r$  times,  $\frac{1}{2}v < r < v-1$ . Because microarrays are expensive, algorithms that produce smaller designs would be of great interest.

In the meantime, for larger  $v$  we can only recommend the sensible heuristic of mimicking the patterns of optimal designs for smaller values of  $v$ . When the number of arrays is not much greater than  $v$ , these patterns are easy to see in the graphical representation of designs. One example is the designs with  $v+2$

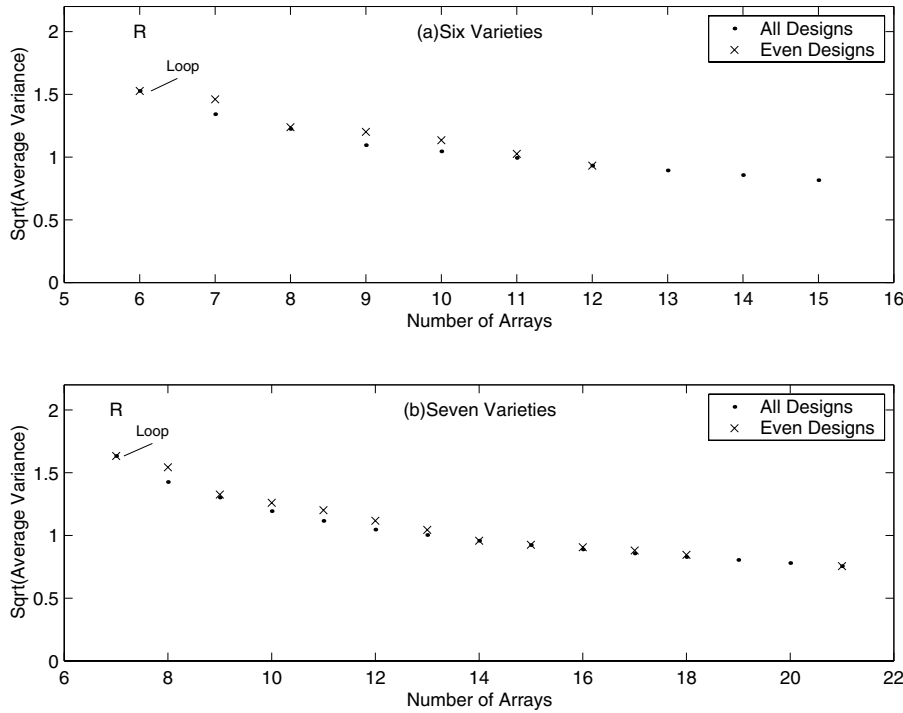


Fig. 5. A-optimality scores for A-optimal designs. Each plot for  $v = 6, 7, 8, 9, 10$  varieties gives the smallest attainable average variance for the contrasts of interest for a design with  $b$  arrays,  $v \leq b \leq \binom{v}{2}$ , assuming model (4.2). The factor  $\frac{n-1}{n}\sigma^2$  is removed from each average variance and then plotted on the square-root scale. The best A-optimality score over all designs is given with a ‘.’ and the best over even designs is given with an ‘X.’ The absence of an ‘X’ means there is no even design for that  $v$  and  $b$ . The reference design always results in an average variance of  $4\frac{n-1}{n}\sigma^2$ , and this is denoted with an ‘R’ at 2 on each plot over  $b = v$ . Loop designs are the only even design for  $v = b$ , and they are A-optimal for  $v \leq 8$ . These are indicated on the graphs. Regular bipartite designs are always A-optimal and they are even when  $v$  is a multiple of 4. This design is indicated on the plot for  $v = 8$ .

arrays for  $v$  varieties above. Another example is designs using  $2v$  arrays for  $v$  varieties, shown in Figure 7. These designs can be described as ‘interwoven’ loops. Notice that for  $v = 7, 9$  the A-optimal design is not a cyclic design (John *et al.*, 1972). It is reasonable to suspect that designs of this type are A-optimal for larger  $v$ . In general, when constructing designs of any size one should keep in mind the basic principles of good design: balance among the factors, approximately equal sampling of varieties, and minimizing the distance between pairs of varieties.

## 7. DISCUSSION

We have had success studying microarray data with models such as (4.1)–(4.3) (Kerr *et al.*, 2000), so we have used linear models as a starting point for studying microarray experimental design. A simple linear model seems to be an obvious place to start for data that depend on many multi-level factors. We stress, however, this is the beginning and not the end of the story. Other modeling assumptions are similarly open to scrutiny. The terms we include in our models and the design factors we consider are

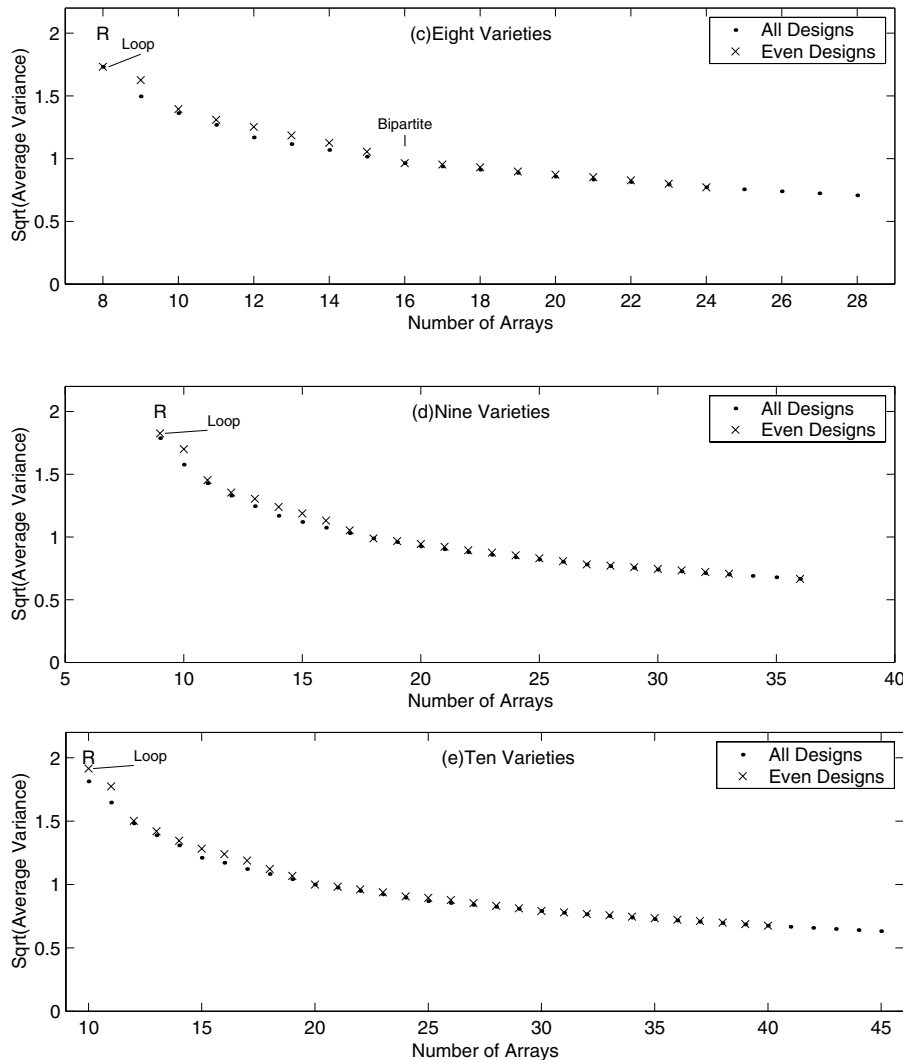


Fig. 5. cont.

certainly debatable. Further, we based our results on the simple assumption of independent, additive error with constant variance  $\sigma^2$ . Again, we have found this to be a reasonable assumption in our experience with data, but do not consider the question to be settled. As discussed in Section 4, a more general assumption of a gene-dependent  $\sigma_g^2$  variance can be incorporated into our framework.

In our exploration of design, we have treated all effects as fixed. We made this assumption for simplicity and convenience and not out of a conviction that it is the correct assumption. Indeed, we are inclined to agree that it may be more appropriate to model certain factors as random effects (Robinson, 1991). The consequences of random or mixed models for experimental design are not entirely clear. Many of our results generalize the developed theory for incomplete block design, but the vast majority of the literature on incomplete block design assumes fixed effects models. Some work has been done to

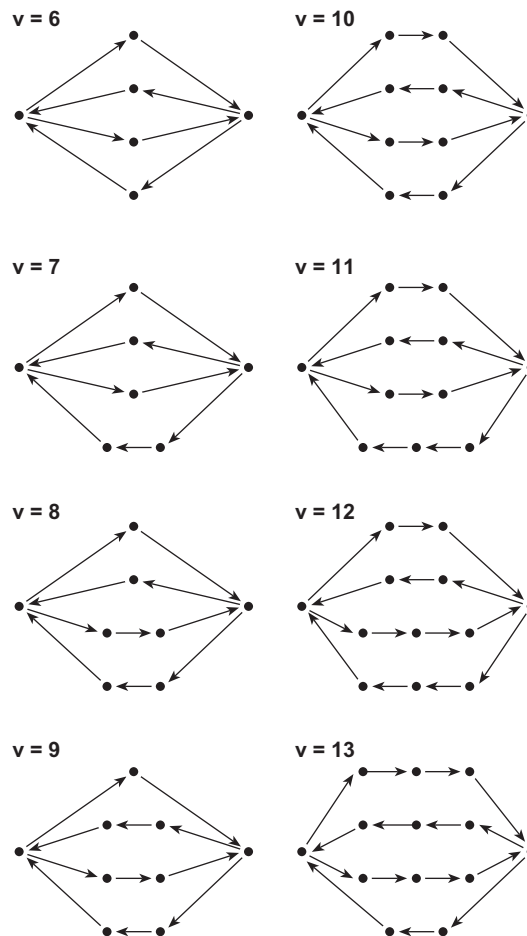


Fig. 6. A-optimal even designs with  $v + 2$  arrays. The A-optimal designs among the set of even incomplete block designs for studying  $v$  varieties with  $v + 2$  arrays are shown for  $v = 6, \dots, 13$ . Nodes ( $\cdot$ ) represent varieties and edges represent arrays. Orientation of edges should be assigned to maintain balance of varieties with respect to dyes.

study incomplete block design when block effects are taken to be random effects. (In our context, this corresponds to ‘spot’ or  $AG$  effects taken to be random.) This research has confirmed the optimality of some families of designs under mixed models (Bhattacharya and Shah, 1984; Mukhopadhyay, 1984) when one is restricted to a general family of designs such as binary designs. Such results confirm our intuition that design efficiency should not depend heavily on whether effects are fixed or random, but further research is needed.

One of our main purposes in this paper was to connect microarray experimental design with classical results. However, we realize the usefulness of classical results can only extend so far. Higher-order analytical tools, such as cluster analysis, are often applied to gene expression data. Several groups have studied different clustering algorithms (Ben-Dor *et al.*, 1999; Eisen *et al.*, 1998; Heyer *et al.*, 1999; Lazzeroni and Owen, 2000; Tamayo *et al.*, 1999), seeking methods best suited to microarray data. We

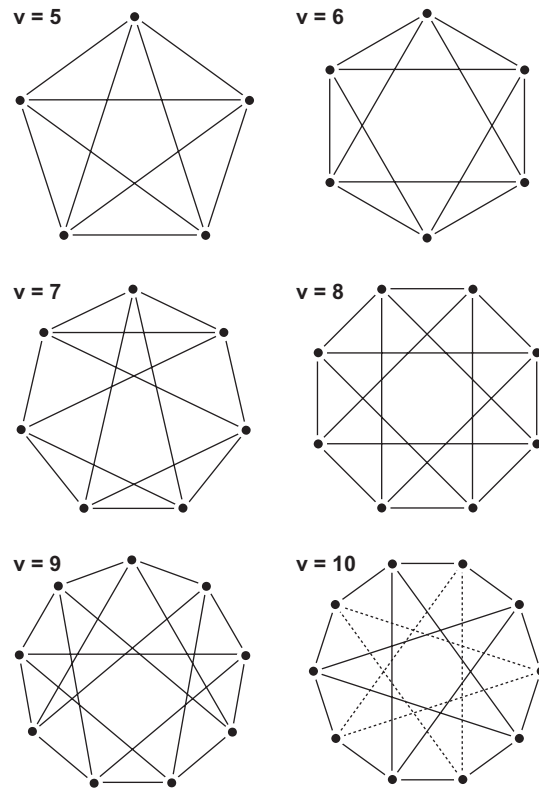


Fig. 7. A-optimal designs with  $2v$  arrays. The A-optimal incomplete block designs with  $2v$  arrays are seen to be even for  $5 \leq v \leq 10$ . For  $v = 5$  the design is the balanced incomplete block design because  $2v = \binom{v}{2}$ . For  $v = 8$  the design is the group-divisible PBIB(2), or bipartite, design. The general pattern of these designs is ‘interwoven’ loops. For  $v = 7, 9$  the A-optimal design is not cyclic.

believe a question of equal importance is how to assess one’s confidence in the output of any clustering algorithm, given that estimates of relative expression are just that—estimates. Clearly, more reliable clusters should follow from more precise estimates, and the foundation of efficient estimation is design. We would like to see an exploration of design issues for experiments where clustering or other higher-order analyses are planned.

Finally, we have used the A-optimality criterion to evaluate designs in this paper, but also consider other design properties, such as balance. Balance and replication are general principles of good design, and should make the designs we recommend robust to our modeling assumptions. In an experiment where all variety pairs are equally interesting to compare, A-optimal even designs are efficient and ensure these robust properties. For experiments in which the varieties have a design structure of their own, it will be important to choose an arrangement of samples on arrays that does not confound the comparisons of interest with other design factors and yields good precision for the estimates of interest. Exactly how such experiments should be designed remains an area of open investigation. We look forward to many new and creative developments in this area of statistical design.

APPENDIX A

*Least-squares estimators for reference design*

Denote the reference variety in the reference design as  $k = 0$ . The varieties of interest are  $k = 1, \dots, v$ . For  $v$  varieties of interest there are  $v$  arrays in the design. The design can be summarized as  $\{(k, 0, g), (k, k, g) : k = 1, \dots, v; g = 1, \dots, n\}$ .

There can be no dye effects in the model when working with the reference design because ‘dye’ is completely confounded with ‘variety’. The revised model (4.2) is

$$y_{ikg} = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg},$$

where the  $V_k$  nominally measure variety effects but actually measure a combination of variety and dye effects. A convenient set of linear constraints is  $\sum A_i = vV_0 + V_1 + \dots + V_{n_k} = \sum G_g = \sum_g (AG)_{ig} = \sum_g (VG)_{kg} = v(VG)_{0g} + (VG)_{1g} + \dots + (VG)_{vg} = 0$ . With these constraints the least-squares estimators of the  $VG$  effects are

$$\widehat{(VG)}_{kg} = 2(y_{kkk} - y_{kk\cdot} - y_{k\cdot g} + y_{k\cdot\cdot}) + y_{\cdot 0g} - y_{\cdot 0\cdot} - y_{\cdot\cdot g} + \bar{y}.$$

APPENDIX B

*Least-squares estimators for loop design*

The loop designs uses  $v$  arrays to study  $v$  varieties of interest. The varieties of interest are  $k = 1, \dots, v$  and the arrays are  $i = 1, \dots, v$ . Without loss of generality, say array  $i$  contains variety  $i$  and  $i + 1$  for  $i = 1, \dots, v - 1$  and array  $v$  holds varieties  $v$  and  $1$ . In other words, variety  $1$  is on arrays  $v$  and  $1$  and variety  $k$  is on arrays  $k - 1$  and  $k$  for  $k = 2, \dots, v$ . To simplify things, all subscripting with  $i$  and  $k$  is understood to be modulo  $v$ .

To estimate (4.2) we use the constraints  $\sum A_i = \sum D_j = \sum V_k = \sum G_g = \sum_g (AG)_{ig} = \sum_g (VG)_{kg} = \sum_i (AG)_{ig} = \sum_k (VG)_{kg} = 0$ . Call  $v_k = v_{kg} = y_{\cdot kg} - y_{\cdot\cdot k} - y_{\cdot\cdot\cdot g} + y_{\cdot\cdot\cdot\cdot}$  and  $\alpha_i = \alpha_{ig} = y_{i\cdot\cdot g} - y_{i\cdot\cdot\cdot} - y_{\cdot\cdot\cdot g} + y_{\cdot\cdot\cdot\cdot}$ . Call  $\gamma_k = \gamma_{kg} = v_k - \frac{1}{2}(\alpha_{k-1} + \alpha_k)$ . The functional form for the estimator for  $(VG)_{kg}$  depends on  $v$  and its parity.

Define sequences  $a_l = l(l + 1)$  and  $b_l = l^2$ . Let  $m = [v/2]$ . If  $v$  is odd,  $\frac{v}{2}\widehat{(VG)}_{kg} = a_m\gamma_k + \sum_{i=1}^{m-1} a_{m-i}(\gamma_{k-i} + \gamma_{k+i})$ . If  $v$  is even,  $\frac{v}{2}\widehat{(VG)}_{kg} = b_m\gamma_k + \sum_{i=1}^{m-1} b_{m-i}(\gamma_{k-i} + \gamma_{k+i})$ . For example, if  $v = 7$  then  $\frac{7}{2}\widehat{(VG)}_{kg} = (2, 6, 12, 6, 2) \cdot (\gamma_{k-2}, \gamma_{k-1}, \gamma_k, \gamma_{k+1}, \gamma_{k+2})$ . If  $v = 8$  then  $\frac{8}{2}\widehat{(VG)}_{kg} = (1, 4, 9, 16, 9, 4, 1) \cdot (\gamma_{k-3}, \gamma_{k-2}, \gamma_{k-1}, \gamma_k, \gamma_{k+1}, \gamma_{k+2}, \gamma_{k+3})$ .

APPENDIX C

*Average variance of pairwise VG effect contrasts*

In classical block design, the reduced normal equations for estimating the variety effects  $\tau = (V_1, \dots, V_v)'$  are given by  $C\tau = Q$ , where  $C$  is the information matrix of the design as at (5.3) with eigenvalues  $0 = \mu_1 \leq \mu_2 \leq \dots \leq \mu_v$ ,  $Q$  is the vector of ‘variety totals adjusted for blocks’, and  $\text{var}(Q) = \sigma^2 C$ . Fitting model (4.2) using the linear constraints  $\sum A_i = \sum D_j = \sum r_k V_k = \sum G_g = \sum_g (VG)_{kg} = \sum_k r_k (VG)_{kg} = 0$ , let  $\tau^* = \tau_g^* = ((VG)_{1g}, \dots, (VG)_{vg})'$ . Let  $Q_k^* = Q_{kg}^* = r_k(y_{\cdot kg} - y_{\cdot\cdot k}) - \sum_{i \ni k} (y_{i\cdot g} - y_{i\cdot\cdot})$ , where  $r_k$  is the number of times variety  $k$  appears in the design and  $i \ni k$  means the arrays  $i$  containing variety  $k$ . The  $Q_k^*$  can be considered ‘VG totals adjusted for

AG totals' and we have  $C\tau^* = Q^*$  and  $\text{var}(Q^*) = \sigma^2 \frac{n-1}{n} C$ , where  $n$  is the number of genes. Thus the proof that the average variance of a pairwise contrast  $\tau_{k_1}^* - \tau_{k_2}^*$  is  $\frac{n-1}{n} \frac{2\sigma^2}{v-1} \sum_{i=2}^v \frac{1}{\mu_i}$  follows precisely as the proof of the classical result that the average variance of a pairwise contrast  $\tau_{k_1} - \tau_{k_2}$  is  $\frac{2\sigma^2}{v-1} \sum_{i=2}^v \frac{1}{\mu_i}$  (see Raghavarao (1971)).

## REFERENCES

- ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., MA, C., LOSSOS, I. S., ROSENWALD, A., BOLDRICK, J. C., SABET, H., TRAN, T., YU, X., POWELL, J. I., YANG, L., MARTI, G. E., MOORE, T., HUDSON, J., LU, L., LEWISH, D. B., TIBSHIRANI, R., SHERLOCK, G., CHAN, W. C., GREINER, T. C., WEISENBURGER, D. D., ARMITAGE, J. O., WARNKE, R., LEVY, R., WILSON, W., GREVER, M. R., BYRD, J. C., BOTSTEIN, D., BROWN, P. O. AND STAUDT, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.
- ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D. AND LEVINE, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* **96**, 6745–6750.
- BAGCHI, S. AND CHENG, C.-S. (1993). Some optimal designs of block size two. *Journal of Statistical Planning and Inference* **37**, 245–253.
- BEN-DOR, A., SHAMIR, R. AND YAKHINI, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology* **6**, 281–297.
- BECHHOFFER, R. E AND TAMHANE, A. C. (1981). Incomplete block designs for comparing treatments with a control: general theory. *Technometrics* **23**, 45–57.
- BHATTACHARYA, C. G. AND SHAH, K. R. (1984). On the optimality of block designs under a mixed effects model. *Utilitas Mathematica* **26**, 339–345.
- BOSE, R. C. (1963). Strongly regular graphs, partial geometries and partially balanced designs. *Pacific Journal of Mathematics* **13**, 389–419.
- BOSE, R. C. AND NAIR, K. R. (1939). Partially balanced incomplete block designs. *Sankhya* **4**, 337–372.
- BROWN, P. O. AND BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**(1 Suppl), 33–37.
- CHENG, C.-S. (1978). Optimality of certain asymmetrical experimental designs. *Annals of Statistics* **6**, 1239–1261.
- CHENG, C.-S. AND BAILEY, R. A. (1991). Optimality of some two-associate-class partially balanced incomplete-block designs. *Annals of Statistics* **19**, 1667–1671.
- CHU, S., DERISI, J., EISEN, M., MULHOLLAND, J., BOSTEIN, D., BROWN, P. O. AND HERSHKOWITZ, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705.
- CLATWORTHY, W. H. (1973). Tables of two-associate-class partially balanced designs. *Applied Mathematics Series 63*, Washington DC: National Bureau of Standards.
- COCHRAN, W. G. AND COX, G. M. (1992). *Experimental Designs*. New York: Wiley.
- DERISI, J. L., IYER, V. R. AND BROWN, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.
- EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. AND BOTSTEIN, D. (1998). *Proceedings of the National Academy of Sciences* **25**, 14863–14868.
- FISHER, R. A. (1951). *The Design of Experiments*, 6th edn. London: Oliver and Boyd.

- HEDAYAT, A. S. AND MAJUMDAR, D. (1984). A-optimal incomplete block designs for control-test treatment comparisons. *Technometrics* **26**, 363–370.
- HEYER, L. J., KRUGLYAK, S. AND YOOSEPH, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Research* **9**, 1106–1115.
- JOHN, J. A. AND MITCHELL, T. J. (1977). Optimal incomplete block designs. *Journal of the Royal Statistical Society, Series B* **39**, 39–43.
- JOHN, J. A., WOLOCK, F. W. AND DAVID, H. A. (1972). Cyclic designs. *Applied Math Series 62*, Washington DC: National Bureau of Standards.
- KERR, M. K., MARTIN, M. AND CHURCHILL, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.
- LAZZERONI, L. AND OWEN, A. (2000). Plaid models for gene expression data. *Technical Report No. 211 in the Stanford University Biostatistics Series*.
- MUKHOPADHYAY, S. (1984).  $\Psi_f$  optimality of MBGDD of type I under mixed effects model within the restricted class of binary designs. *Sankhya: The Indian Journal of Statistics, Series B*, **46**, pp. 113–117.
- MCKAY, B. (1991). “Nauty” and “makeg” C-programs available at <http://cs.anu.edu.au/people/bdm/nauty/>.
- PEROU, C. M., JEFFREY, S. S., VAN DE RIJN, M., REES, C. A., EISEN, M. B., ROSS, D. T., PERGAMENSCHIKOV, A., WILLIAMS, C. F., ZHU, S. X., LEE, J. C.F., LASHKARI, D., SHALON, D., BROWN, P. O. AND BOTSTEIN, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences* **16**, 9212–9217.
- RAGHAVARAO, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. New York: Wiley.
- ROBINSON, G. K. (1999). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15–51.
- ROSS, D. T., SCHERF, U., EISEN, B. M., PEROU, C. M., REES, C., SPELLMAN, P., SPELLMAN, V., JEFFREY, S. S., VAN DERIJN, M., WALTHAM, M., PERGAMENSCHIKOV, A., LEE, J. C. F., LASHKARI, D., SHALON, D., MYERS, T. G., WEINSTEIN, J. N., BOTSTEIN, D. AND BROWN, P. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**, 227–235.
- TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREWAN, S., SMITROVSKY, E., LANDER, E. AND GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences* **96**, 2907–2912.

[Received May 3, 2000; revised August 10, 2000; accepted for publication August 31, 2000]