Extending the loop design for two-channel microarray experiments

NAOMI S. ALTMAN* AND JUN HUA

Department of Statistics, Thomas Building, The Pennsylvania State University, University Park, PA 16801-2111, USA

(*Received 25 July 2006 and in revised form 16 November 2006*)

Summary

The loop design of Kerr and Churchill is a clever application of incomplete blocks of size 2 to two-channel microarray experiments. In this paper, we extend the loop design to include more replicates, biological and technical replication, multi-factor experiments, and blocking. Loop and extended loop designs are shown to be more efficient than the reference design for any given number of arrays. We also show that adding new treatments to a loop design requires the same number of additional arrays as adding treatments to a reference design, with a greater gain in power. Given the flexibility of extended loop designs and their power, we propose that these should be the designs of choice for most experiments using two-channel microarrays.

1. Introduction

A microarray is a platform for simultaneously measuring the expression of thousands of genes. Since they were first devised in 1995, microarrays have become one of the tools of choice for genome-wide studies of expression. Because of the noisy nature of the data, and the high dimension of the response, development of statistical tools for processing microarray data is a growth industry.

However, it was not until the seminal papers of Kerr & Churchill (2001 *a*, *b*) that serious effort was put into the design of microarray experiments. In particular, Kerr & Churchill (2001 *a*) put the design of microarray experiments in the context of classical experiment design. Subsequently, Dobbin *et al.* (2003), Landgrebe *et al.* (2004), Glonek & Solomon (2004), Wit *et al.* (2005) and Latif (2005) have applied the principles of optimal design to two-channel microarrays. This paper is in the same spirit. The many sources of both biological and technical variation in the experiments imply that considerable gains in efficiency may be possible with good experimental design.

The basic principle of measuring gene expression on a microarray is simple. The array is a substrate on which are printed 'spots' or probes. When a gene expresses in a tissue, it produces mRNA which can be processed into single-stranded cDNA. Each spot on the array consists of many copies of single-stranded DNA which is complementary to a selected cDNA. A gene may be represented by one or more spots – if more than one, these may be identical or may represent different sections (oligos) of the same gene. Gene expression is detected by extracting the mRNA from the tissue, processing into cDNA, labelling with a fluorescent dye, and then allowing the sample to hybridize with the complementary spots on the array. The measurement of gene expression is a quantification of the intensity of fluorescence.

A commonly used technology is the two-channel array. This technology allows two samples, labelled with two dyes, generally Cy3 (green) and Cy5 (red), to be hybridized to the same array. This reduces some of the noise in the system, since it eliminates sources of variability due to differences among arrays. However, the chemistry of sample labelling can introduce gene-specific dye biases if a cDNA has greater affinity for one of the two dyes. Because of the need to guard against dye bias, we consider only designs in which each treatment has an even number of samples, with equal numbers labelled with green and red (dye-balanced designs). However, we make explicit that technical replication is not required for dye-swap

* Corresponding author. Telephone: +1 (814) 965 3791. Fax: +1 (814) 863 7114. e-mail: naomi@stat.psu.edu

designs, and allow the dye-swap pairs to come from different biological replicates. Thus, unlike Kerr & Churchill (2001*a*) all our samples may be biological replicates, so that designs without dye-swap do not save any labelling reactions, and our reference designs include dye-swaps of the reference and treatment samples. Dobbin *et al.* (2003) discuss the dye balance issue in more detail.

In this paper, we will use the term 'treatment' where Kerr & Churchill (2001*a*) used the term 'variety'. This terminology is used in the statistical sense of the combination of factor levels which apply to the sample. For example, in a time course experiment observing the effects of various doses of a chemical agent on liver and brain tissue in two strains of mice, the treatment would be the combination of genotype, tissue type, dose and time point.

We are also careful to distinguish between biological and technical replication. Kerr & Churchill (2001*a*) considered only designs in which each treatment was represented by a single sample of mRNA, which was hybridized to two microarrays. This type of design is suitable for making inferences about the samples. For biological inference, a measure of biological variation is required, and this requires multiple biologically independent samples. Most of this paper discusses designs in which there is only biological replication – that is each treatment is represented by several biologically independent samples of mRNA, each of which is hybridized to a single microarray. In Section 6, we will briefly take up the question of experiments having both biological and technical replication in which the biological samples are split and hybridized to two microarrays.

A two-channel microarray can be thought of as a block of size 2. If only two experimental treatments are to be compared, the principles of experimental design suggest that the experiment should be treated as a randomized complete block design, i.e. one sample from each condition should be hybridized to each array, with biological replicate dye-swaps. However, if there are three or more treatments, this approach cannot be utilized, as each array can have only two samples. The choice of which samples are paired on the arrays and how they are labelled is called the hybridization design.

A popular hybridization design for two-channel microarray studies is the reference design, in which one channel on each array is used for a reference sample. The same reference is used for all arrays, and the difference or ratio of expression between the treatment sample and the reference sample is usually used as the unit of analysis. The reference design is briefly discussed in Section 3 to provide a basis of comparison with the loop design, and because it is a very popular design for two-channel arrays.
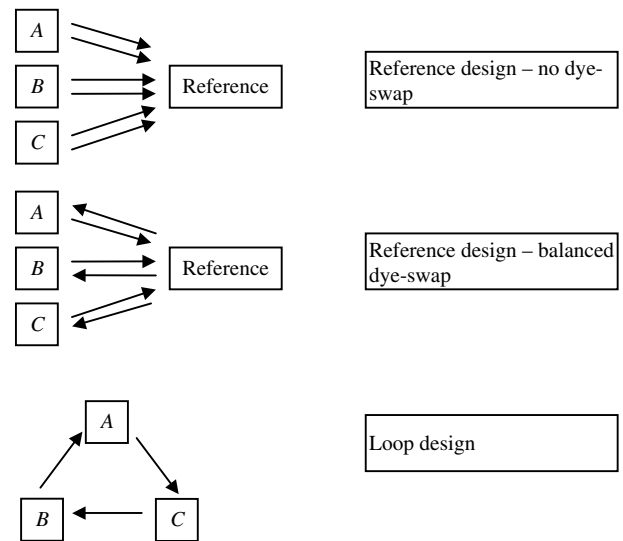

Fig. 1. Some basic designs for two-channel microarrays with three treatments and two samples per treatment. Each boxed letter represents two biologically independent samples.

Kerr & Churchill (2001*a*, *b*), introduced the loop design, which is a type of partially balanced incomplete block design. A loop design includes exactly two replicates of each treatment, one labelled with green and the other with red. The design is most readily visualized by a diagram such as Fig. 1, in which each arrow designates an array. The tail of the arrow is attached to the sample labelled with green and the head to the sample labelled with red. The loop design does not use reference samples, as a result of which only half the arrays are required to attain the same sample size as the reference design. The loop design is balanced for dye effects (i.e. equal numbers of replicates of each treatment are labelled with each dye). Kerr & Churchill assumed that two replicates would be sufficient and that dye-swaps would be achieved by splitting samples. Again, we emphasize that in this paper, each sample is biologically distinct.

In this paper, we model the data for each channel ('single-channel analysis'), instead of taking differences or ratios. This allows us to compute the variance of any contrast, and to determine the effect of various types of replication on the variance. We feel it is most natural to consider the array effect to be a source of unexplained variation, and thus treat it as a random effect in our analyses. Variance computations allow us to design the hybridizations to minimize the variance of selected contrasts, a basic concept in optimal experiment design. This allows us to explore optimal allocation of additional replicates, the arrangement of factorial treatments around the loop, the use of technical and biological replicates, and the use of randomized complete block designs for the samples.
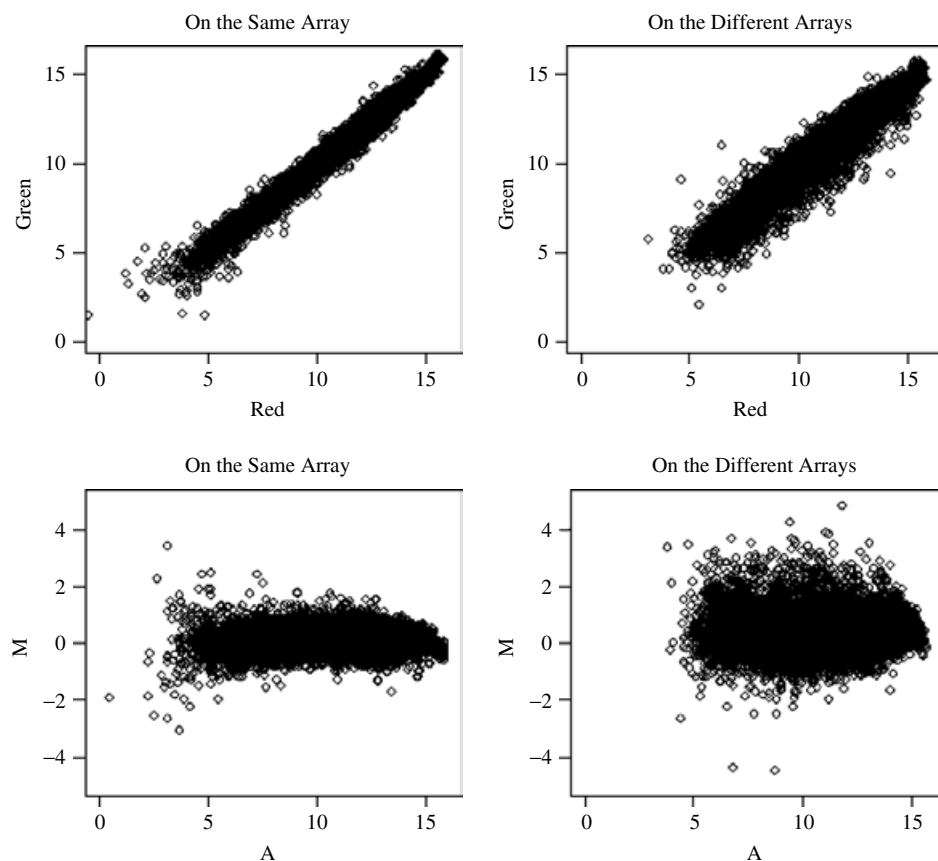
Fig. 2. Log$_2$(Expression) from a *Drosophila* experiment. The Green and Red samples are different biological samples that were split after labelling into two aliquots. A is the average Log$_2$(Expression) while M is the difference in expression. In the two left-hand panels, the samples were hybridized to one array. In the two right-hand panels, the samples were hybridized to different arrays.

Two commonly cited 'problems' with loop designs are the loss of information for some spots on some arrays, leading to unbalanced designs, and the problem of adding treatments to the experiment. We note that with single-channel analysis, information loss leads to an unbalanced design, but is otherwise not problematic unless all the data are lost for a given gene and treatment combination. Reference designs also suffer from imbalance if there is information loss. We also suggest a simple method of 'patching' a new treatment into an existing loop.

2. To Difference or not to difference?

The objective of two-channel microarray experiments is comparative analysis. As a result, the two channels on the array are often differenced (after converting to logarithms base 2), and the difference is treated as the observation of interest. That is, denoting the log$_2$ expression as Y_{ica} for gene i in channel c (red or green) on array a, analysis is generally done on $M_{ia} = Y_{iRa} - Y_{iGa}$. This is particularly convenient, because normalization is often done by normalizing M_{ia} against the average of the two channels, $A_{ia} = (Y_{iRa} + Y_{iGa})/2$.

In this paper, we use the individual channels as the observations of interest. Thus, each array provides two observations for each gene. As needed, we convert the normalized differences, $\tilde{M}_{ia}$ back to channel-level data via the transformation:

$$\tilde{Y}_{iRa} = (2A_{ia} + \tilde{M}_{ia})/2$$
$$\tilde{Y}_{iGa} = (2A_{ia} + \tilde{M}_{ia})/2.$$

Often A is normalized across arrays. See Smyth (2005) for single-channel analysis of two-channel microarrays.

To see the efficacy of this approach, we look at some data from a *Drosophila* array (B. McIver, personal communication). In this experiment, two biological samples were labelled and split. Both samples were then hybridized to the same array, as well as to other arrays with different samples. The normalized data are displayed in Fig. 2. The two left-hand panels show the normalized log$_2$(expression) values for the two samples on the same array. The two right-hand panels show the values for the different arrays. While the two right-hand panels are much noisier (Var(M) = 0·453 as opposed to Var(M) = 0·126

in the two left-hand panels) there is still considerable information in the individual channels.

In the next section we show that a partial difference of the form $\tilde{Y}_{iRa} - w\tilde{Y}_{iGa}$ is the optimal unit of analysis for a reference design. The optimal value of w depends on the correlation between samples hybridized to the same spot (or intraclass correlation). This quantity is unknown, but can be estimated from a mixed-effect ANOVA. In unpublished experiments using Agilent two-colour arrays, using the LIMMA (Smyth, 2004) software to assess the mean intraclass correlation, we found that the correlation between the two channels was about 0·2–0·7, indicating an information loss of 17–65 % by differencing, instead of using the single-channel analysis.

The same computations can be used to determine the relative efficiency of different experimental designs and to determine optimal designs for microarray experiments. While the optimal weights vary from gene to gene, the ranking of the relative efficiency of different designs does not depend on the value of the weights, and hence the optimal design is the same for all genes.

We start with the simplest case, with one hybridization per biological replicate and one spot per gene on each array. We will use a simplified mixed model for the normalized expression data by channel:

$$\tilde{Y}_{ijca} = \mu_i + \tau_{ij} + \gamma_{ic} + \alpha_{ia} + \varepsilon_{ijca} \qquad (1)$$

where

μ_i is the mean normalized expression for gene i over the entire experiment

τ_{ij} is the mean effect of treatment j on gene i

γ_{ic} is the mean effect of dye c on gene i (dye by gene interaction)

α_{ia} is the random effect of array a for gene i with variance σ_{ia}^2

ε_{ijca} is the random error with variance $\sigma_{i\varepsilon}^2$.

The variance of an observation is $\sigma_{ia}^2 + \sigma_{i\varepsilon}^2$. The covariance between the channels at the same spot is σ_{ia}^2. A key quantity is the intraclass correlation, $\rho_i = \frac{\sigma_{ia}^2}{\sigma_{ia}^2 + \sigma_{i\varepsilon}^2}$, which in the experience of the authors is often above 0·5. The variance of the difference between two expression levels of the same gene on different arrays includes a component due to the two arrays and a component due to the two errors, giving $2(\sigma_{i\varepsilon}^2 + \sigma_{ia}^2)$, while the variance of difference on a single array does not include array variance, giving $2\sigma_{i\varepsilon}^2$. Thus the relative efficiency of comparisons between samples on the same array, compared with comparisons of samples on different arrays is $2\sigma_{i\varepsilon}^2/[2(\sigma_{i\varepsilon}^2 + \sigma_{ia}^2)] = 1 - \rho_i$. As discussed in more detail in the remainder of the paper, good designs take advantage of this variance reduction by pairing treatments of most interest on the same arrays. Good analyses take advantage of this

variance reduction by weighting more efficient comparisons more highly in computations.

On some arrays, genes may be represented by multiple spots. On cDNA arrays, these are duplicates of the same cDNA. On oligo arrays, these may be duplicates of the same oligos, or other oligos from the same gene. The information from multiple spots is often combined to form a gene summary for each gene on each array, such as a simple average or methods resistant to outliers, such as the median. (Note that using a linear model with spot effects is equivalent to averaging.) Alternatively, some investigators prefer to analyse each spot, rather than each gene. If each spot is analysed, then model (1) applies directly, but the subscript i refers to spots, rather than genes. If a gene summary is used, model (1) still applies, but α_{ia} and ε_{ijca} refer to effects for the gene summaries, rather than for individual spots.

Model (1) assumes that independent biological samples are hybridized to each array. However, it is still common to have technical replication in which a single mRNA sample is split and hybridized to multiple arrays. For these cases, the model needs to be expanded to

$$\tilde{Y}_{ijlcab} = \mu_i + \tau_{ij} + \gamma_{ic} + \alpha_{ia} + \beta_{ijb} + \varepsilon_{ijlcab} \qquad (2)$$

where μ_i, τ_{ij} and γ_{ic}, α_{ia} and ε_{ijlcab} have the same meanings as in model (1) and β_{ijb} is the random effect of biological sample b within treatment j for gene i with variance $\sigma_{i\beta}^2$. Sections 2 to 5 of this paper concern model (1). We return to model (2) in Section 6.

The focus is on a gene-by-gene analysis of the efficiency of pairwise comparisons between treatments. Hence we do not make distribution assumptions except for the independence of the random effects in models (1) and (2), and the usual constraints on the fixed effects required for identifiability. The optimal weights and the variance of pairwise comparisons depend on the variances and covariances of the data. When the variance components are known, general least squares can be used to estimate the pairwise comparisons using the optimal weights, and also compute the variance of the estimates. For the purposes of planning the hybridization design for a fixed number of arrays and biological samples, estimates of the components are not required.

For data analysis, estimates of the variance components can be estimated by restricted maximum likelihood (REML) or maximum likelihood (ML), but these methods are very inaccurate with the small sample sizes often used in microarray studies. Various approaches using all the spots on the array have been suggested to improve estimation. Smyth *et al.* (2005) use empirical Bayes estimation to improve the estimate of $\sigma_{i\varepsilon}^2$ and assume that $\rho_i = \rho$, a single value that can be computed from all the genes. Cui *et al.* (2005)

use shrinkage estimation to improve the estimation of all the variance components.

3. Reference design

In a reference design, one channel of each array is a reference sample, which is the same biological material on every array. With T treatments and k replicates per treatment, we use $k \times T$ arrays.

Usually the analysis is done on the difference between the treatment and reference samples on each array. For convenience of exposition we assume that the differences have been computed as $M = $ treatment – reference, so we can denote the normalized difference as $\tilde{M}_{ia}$ without keeping track of $+/-$ signs for dye-swaps.

To compute the variance of estimated comparisons, we use the formula

$$\mathrm{Var}(Y - wX) = \mathrm{Var}(Y) + w^2\mathrm{Var}(X) - 2w\mathrm{Cov}(Y, X) \tag{3}$$

where w is a fixed number and Y and X are the random observations. We will consider optimizing our choice of w to minimize variance. Through calculus or algebra, we can see the optimal choice of w is $\mathrm{Cov}(Y, X)/\mathrm{Var}(X)$ and the minimum variance achievable is $\mathrm{Var}(Y) - \mathrm{Cov}^2(Y, X)/\mathrm{Var}(X)$. Using model (1), the variance of the log(expression) in a single spot in a single channel is $\sigma_{ia}^2 + \sigma_{i\varepsilon}^2$. The covariance between the channels at the same spot is σ_{ia}^2. Ignoring the effect of normalization and using R to indicate the reference sample, we have for gene i on array a, $\tilde{M}_{iAa} = \tilde{Y}_{iA1a} - \tilde{Y}_{iR2a}$ with variance

$$\mathrm{Var}(\tilde{Y}_{iA1a}) + \mathrm{Var}(\tilde{Y}_{iR2a}) - 2\mathrm{Cov}(\tilde{Y}_{iA1a}, \tilde{Y}_{iR2a})$$
$$= 2(\sigma_{ia}^2 + \sigma_{i\varepsilon}^2) - 2\sigma_{ia}^2 = 2\sigma_{i\varepsilon}^2.$$

Then if treatment A is on array 1 and treatment B is on array 2, the contrast $A - B$ is estimated by $\tilde{M}_{iA1} - \tilde{M}_{iB2}$ and $\tilde{M}_{iA1}$ and $\tilde{M}_{iB2}$ are independent (and hence have covariance 0). Accordingly, using (3), we find that $\mathrm{Var}(\tilde{M}_{iA1} - \tilde{M}_{iB2}) = 4\sigma_{i\varepsilon}^2$. With k replicates, the estimated contrast would have variance $4\sigma_{i\varepsilon}^2/k$.

We now consider using the partial difference for a treatment T, $\tilde{M}_{iTa}^w = \tilde{Y}_{iTa} - w_i\tilde{Y}_{iRa}$ and estimate the contrast with $\tilde{M}_{iA1}^w - \tilde{M}_{iB2}^w = \tilde{Y}_{iA1} - w_i\tilde{Y}_{iR1} - (\tilde{Y}_{iB2} - w_i\tilde{Y}_{iR2})$. Noting that $\mathrm{Cov}(\tilde{M}_{iA1}^w, \tilde{M}_{iB2}^w) = 0$, we optimize as we did above and find that the optimal weight is $w_i = \rho_i$, with resulting variance $2\sigma_{i\varepsilon}^2(1 + \rho_i)$ and with k biological replicates the variance is $2\sigma_{i\varepsilon}^2(1 + \rho_i)/k$.

While we do not know the variance and intraclass correlation, if we use mixed-model ANOVA, the weights are approximated from the data – leading to more efficient estimates and more powerful tests of differential expression. For small sample sizes, ρ_i (or equivalently, the variances) are not well estimated,

and the improvements using the weighted approach may be smaller than indicated by statistical theory.

4. Loop designs

A loop is balanced for dye effects and has two replicates at each node. For T treatments using Tk arrays we have $2k$ replicates, compared with a reference design for which the same number of arrays yields only k replicates. In this section and Section 5, all replicates are biological replicates. In Section 6, we consider designs with both biological and technical replicates.

The computation of variance of a pairwise comparison in a loop design depends on the number of treatments and the shortest distance between the treatments on the loop. For expository purposes we demonstrate with four treatments, but the conclusions are similar regardless of the number of treatments and are displayed in Table 1 for three to seven treatments.

The three possible arrangements (ignoring loop direction) of the four-treatment loop design are depicted in Fig. 3. Note that each uppercase letter represents two independent biological samples with the same treatment.

As for the reference design, we consider the variance of a comparison. For example, suppose we wish to estimate the mean difference in expression for treatments A and B ($\tau_{iA} - \tau_{iB}$, using the notation in model 1). Using g and r to denote Cy3 and Cy5 respectively, we have one expression value for each treatment with each dye (e.g. for treatment A, we have Y_{Ag} and Y_{Ar}). To avoid confusion and so that the computation applies to all three diagrams in Fig. 3, we suppress the other subscripts, which can be determined from the diagram. Consider a linear combination of the observations:

$$w_1 Y_{Ag} + w_2 Y_{Ar} + w_3 Y_{Bg} + w_4 Y_{Br} + w_5 Y_{Cg}$$
$$+ w_6 Y_{Cr} + w_7 Y_{Dg} + w_8 Y_{Dr}.$$

This will be an unbiased estimator of $\tau_{iA} - \tau_{iB}$ if $w_1 + w_2 = 1$, $w_3 + w_4 = -1$, $w_5 + w_6 = 0$ and $w_7 + w_8 = 0$. Notice that under these constraints, due to the dye balance, the dye effect cancels. We can optimize the weights subject to these constraints to find the weighting which minimizes the variance and the variance using the optimal weights. Equivalently, this is the general linear model estimator of the contrast.

Using the optimal weighting, we find that for the designs in Fig. 3, the variance of the difference of adjacent treatments is $\sigma_{i\varepsilon}^2 + \sigma_{ia}^2\sigma_{i\varepsilon}^2/2(\sigma_{ia}^2 + \sigma_{i\varepsilon}^2) = \sigma_{i\varepsilon}^2(1 + \rho_i/2)$, while the variance of the difference of diagonally opposite treatments is $\sigma_{i\varepsilon}^2 + \sigma_{ia}^2\sigma_{i\varepsilon}^2/(\sigma_{ia}^2 + \sigma_{i\varepsilon}^2) = \sigma_{i\varepsilon}^2(1 + \rho_i)$. Both of these are smaller than the variance of the contrast from a reference design with the same

Table 1. *The variance of a comparison between two treatments which are D steps apart in a loop design with T treatments and T arrays. σ_{ie}^2 is the error variance for gene i. σ_{ia}^2 is the within-spot variance. Due to the larger sample size, the loop design always has smaller variance than the reference design with the same number of arrays, regardless of D*

Design	T	D	Variance	D	Variance	D	Variance
Reference	T	1	$2\sigma_{ie}^2 + \dfrac{2\sigma_{ia}^2\sigma_{ie}^2}{\sigma_{ia}^2+\sigma_{ie}^2}$				
Loop	3	1	$\sigma_{ie}^2 + \dfrac{\sigma_{ia}^2\sigma_{ie}^2}{3\sigma_{ia}^2+2\sigma_{ie}^2}$				
	4	1	$\sigma_{ie}^2 + \dfrac{\sigma_{ia}^2\sigma_{ie}^2}{2(\sigma_{ia}^2+\sigma_{ie}^2)}$	2	$\sigma_{ie}^2 + \dfrac{\sigma_{ia}^2\sigma_{ie}^2}{\sigma_{ia}^2+\sigma_{ie}^2}$		
	5	1	$\sigma_{ie}^2 + \dfrac{3\sigma_{ia}^4\sigma_{ie}^2+2\sigma_{ia}^2\sigma_{ie}^4}{5\sigma_{ia}^4+10\sigma_{ia}^2\sigma_{ie}^2+4\sigma_{ie}^4}$	2	$\sigma_{ie}^2 + \dfrac{7\sigma_{ia}^4\sigma_{ie}^2+4\sigma_{ia}^2\sigma_{ie}^4}{5\sigma_{ia}^4+10\sigma_{ia}^2\sigma_{ie}^2+4\sigma_{ie}^4}$		
	6	1	$\sigma_{ie}^2 + \dfrac{2\sigma_{ia}^4\sigma_{ie}^2+2\sigma_{ia}^2\sigma_{ie}^4}{3\sigma_{ia}^4+8\sigma_{ia}^2\sigma_{ie}^2+4\sigma_{ie}^4}$	2	$\sigma_{ie}^2 + \dfrac{5\sigma_{ia}^4\sigma_{ie}^2+4\sigma_{ia}^2\sigma_{ie}^4}{3\sigma_{ia}^4+8\sigma_{ia}^2\sigma_{ie}^2+4\sigma_{ie}^4}$	3	$\sigma_{ie}^2 + \dfrac{2\sigma_{ia}^2\sigma_{ie}^2}{\sigma_{ia}^2+2\sigma_{ie}^2}$
	7	1	$\sigma_{ie}^2 + \dfrac{5\sigma_{ia}^6\sigma_{ie}^2+10\sigma_{ia}^4\sigma_{ie}^4+4\sigma_{ia}^2\sigma_{ie}^6}{7\sigma_{ia}^6+28\sigma_{ia}^4\sigma_{ie}^2+28\sigma_{ia}^2\sigma_{ie}^4+8\sigma_{ie}^6}$	2	$\sigma_{ie}^2 + \dfrac{13\sigma_{ia}^6\sigma_{ie}^2+22\sigma_{ia}^4\sigma_{ie}^4+8\sigma_{ia}^2\sigma_{ie}^6}{7\sigma_{ia}^6+28\sigma_{ia}^4\sigma_{ie}^2+28\sigma_{ia}^2\sigma_{ie}^4+8\sigma_{ie}^6}$	3	$\sigma_{ie}^2 + \dfrac{17\sigma_{ia}^6\sigma_{ie}^2+24\sigma_{ia}^4\sigma_{ie}^4+8\sigma_{ia}^2\sigma_{ie}^6}{7\sigma_{ia}^6+28\sigma_{ia}^4\sigma_{ie}^2+28\sigma_{ia}^2\sigma_{ie}^4+8\sigma_{ie}^6}$

$$
\begin{array}{ccc}
A \rightarrow B & A \rightarrow B & A \rightarrow C \\
\uparrow \quad \downarrow & \uparrow \quad \downarrow & \uparrow \quad \downarrow \\
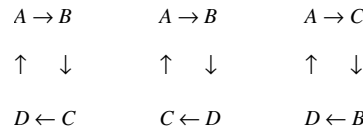D \leftarrow C & C \leftarrow D & D \leftarrow B
\end{array}
$$

Fig. 3. The three possible four-treatment loop designs.

number of treatments and arrays, primarily because there are two biological replicates per sample, rather than one. Also, with four treatments, the loop design with four arrays has sufficient replication for statistical analysis, while the reference design with four arrays has no replication.

Table 1 shows the variance of pairwise contrasts for designs with T treatments on T arrays for the reference and loop designs. For the loop design, the variance depends on the 'distance' between the treatments in the shortest route around the loop. For example, in Fig. 3, the distance between any two adjacent arrays is $D=1$ and between any two diagonally opposite arrays is $D=2$. Table 2 shows the same information as a function of the error variance when ρ_i is 0·25, 0·5 or 0·75. The variance of pairwise contrasts depends on the 'distance' between the treatments on the loop. In practical terms, this means that loop designs must be carefully planned so that the comparisons of most interest are close together on the loop. Alternatively, if we wish to compare treatments which are expected to have similar gene expression patterns, these should be placed closely together on the loop. As will be seen in Section 5, replicated loops can be designed to minimize differences in variance in pairwise contrasts.

Tables 1 and 2 show that the loop design is always more efficient than the reference design with the same number of arrays. In the parlance of Glonek & Solomon (2004), the reference design is inadmissible for pairwise comparisons of treatments. Also, the loop design with T treatments and T arrays has enough replication for statistical analysis. The reference design requires additional arrays in order to have replication, and also requires the investigator to maintain a large uniform reference sample.

5. Replicated loop designs

The basic loop design has two replicates per condition. If more replication is desired, the design can be replicated.

One way to replicate the design is to run an identical loop, or a loop with the dye-swaps of the arrays in the original loop. When there are k replicates of the loop, the entries in Tables 1 and 2 would be divided by k.

When many of the comparisons are of equal importance, a better choice of replication is to run a different loop. The choice of loop can be made to

Table 2. *Numerical variance of pairwise contrasts for one-factor loop and reference designs. We show the numerical values when* $\sigma_{i\varepsilon}^2 : \sigma_{ia}^2 = 1/3$, *1 and 3, which correspond to intraclass correlations of 0·75, 0·5 and 0·25, respectively*

Design	T	D	$\rho_i=0{\cdot}75$	$\rho_i=0{\cdot}5$	$\rho_i=0{\cdot}25$	D	$\rho_i=0{\cdot}75$	$\rho_i=0{\cdot}5$	$\rho_i=0{\cdot}25$	D	$\rho_i=0{\cdot}75$	$\rho_i=0{\cdot}5$	$\rho_i=0{\cdot}25$
Reference	T	1	$3{\cdot}5\sigma_{i\varepsilon}^2$	$3\sigma_{i\varepsilon}^2$	$2{\cdot}5\sigma_{i\varepsilon}^2$								
Loop	3	1	$1{\cdot}273\sigma_{i\varepsilon}^2$	$1{\cdot}2\sigma_{i\varepsilon}^2$	$1{\cdot}111\sigma_{i\varepsilon}^2$								
	4	1	$1{\cdot}375\sigma_{i\varepsilon}^2$	$1{\cdot}25\sigma_{i\varepsilon}^2$	$1{\cdot}125\sigma_{i\varepsilon}^2$	2	$1{\cdot}75\sigma_{i\varepsilon}^2$	$1{\cdot}5\sigma_{i\varepsilon}^2$	$1{\cdot}25\sigma_{i\varepsilon}^2$				
	5	1	$1{\cdot}418\sigma_{i\varepsilon}^2$	$1{\cdot}263\sigma_{i\varepsilon}^2$	$1{\cdot}127\sigma_{i\varepsilon}^2$	2	$1{\cdot}949\sigma_{i\varepsilon}^2$	$1{\cdot}579\sigma_{i\varepsilon}^2$	$1{\cdot}268\sigma_{i\varepsilon}^2$				
	6	1	$1{\cdot}436\sigma_{i\varepsilon}^2$	$1{\cdot}267\sigma_{i\varepsilon}^2$	$1{\cdot}27\sigma_{i\varepsilon}^2$	2	$2{\cdot}036\sigma_{i\varepsilon}^2$	$2{\cdot}036\sigma_{i\varepsilon}^2$	$1{\cdot}6\sigma_{i\varepsilon}^2$	3	$2{\cdot}2\sigma_{i\varepsilon}^2$	$1{\cdot}667\sigma_{i\varepsilon}^2$	$1{\cdot}286\sigma_{i\varepsilon}^2$
	7	1	$1{\cdot}445\sigma_{i\varepsilon}^2$	$1{\cdot}268\sigma_{i\varepsilon}^2$	$1{\cdot}127\sigma_{i\varepsilon}^2$	2	$2{\cdot}075\sigma_{i\varepsilon}^2$	$1{\cdot}60\sigma_{i\varepsilon}^2$	$1{\cdot}270\sigma_{i\varepsilon}^2$	3	$2{\cdot}311\sigma_{i\varepsilon}^2$	$1{\cdot}690\sigma_{i\varepsilon}^2$	$1{\cdot}288\sigma_{i\varepsilon}^2$

minimize the variance of selected comparisons. Some clever choices can be made.

When the number of treatments, T, is odd, the replicates are readily combined into a balanced incomplete block design in which each pair of conditions is found together on an array with complete dye balance for each condition. This is a desirable design, because all the pairwise comparisons have the same variance. This can be achieved simply by adding 'stars' to the original loop consisting of arrays with every second node, every third node up to every $(T-1)/2$ node, giving the $T(T-1)/2$ arrays needed for the balanced incomplete block. This is illustrated for $T=5$ and 7 in Fig. 4.

Unfortunately, when T is an even number, the design with each pair of conditions on an array is balanced for the conditions but not for dyes, and hence is not desirable. To completely balance the design, the dye-swap of each array is also needed, leading to $T(T-1)$ arrays being required. For example, for five conditions, the 10 arrays shown in Fig. 3 form a balanced incomplete block design, but for four conditions, 12 arrays are required. Of course, for five conditions on 10 arrays, there are four replicates per condition, whereas for the four conditions on 12 arrays, there are six replicates per condition, yielding additional power for detecting differential expression. Alternatively, if the number of treatments is even, a control treatment can be added for the purpose of balancing the design for the dye effect with fewer arrays. For example, with 10 treatments, the dye-balanced, balanced incomplete block design requires $10 \times 9 = 90$ arrays but with 11 conditions only $11 \times 10/2 = 55$ arrays are required.

Fig. 3 shows the three possibilities for loops for the four-treatment design. With 12 arrays, the variance of any treatment difference in the reference design is $2/3[\sigma_{i\varepsilon}^2 (1+\rho_i)]$. If we replicate any one loop three times, the variance of treatment differences adjacent on the loop is $1/3\sigma_{i\varepsilon}^2 (1+\rho_i/2)$ and the variance of those diagonally opposite is $1/3\sigma_{i\varepsilon}^2 (1+\rho_i)$. If we use all three possible loops as our replicates, which is the balanced incomplete block design, the variance of
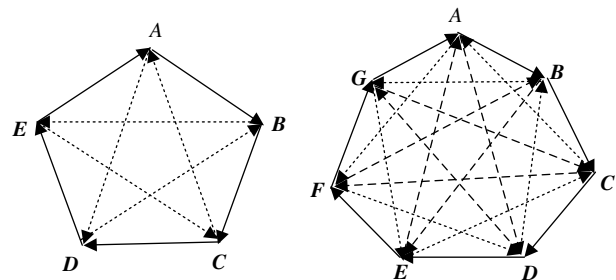

Fig. 4. Replicated loop designs with $(T-1)/2$ replicates forming balanced incomplete block designs. For the design with five treatments, each uppercase letter represents four biologically independent samples. For the design with seven treatments, each letter represents six biologically independent samples. Each type of line represents one complete loop design.

any treatment difference for the loop design is $1/3\sigma_{i\varepsilon}^2 [1{\cdot}5-3/(4\rho_i+3)]$, which is considerably smaller than the variance of the reference design with the same number of arrays and is between the adjacent and diagonal values of the design using the same number of replicates.

In general, for large T, the number of arrays required for a balanced incomplete block design is prohibitive. Designs with two or three loops can be very effective. Kerr & Churchill (2001*a*) and Wit *et al.* (2005) discuss the selection of optimal designs for two-channel microarray experiments. They call designs like those in Fig. 4 'interwoven' loop designs, and note that these are often optimal or near optimal in the sense of minimizing a criterion based on the variance of all the possible contrasts. However, Wit *et al.* also find some designs that are very far from loops which optimize overall variance by having very small variance for a few comparisons and much bigger variance for others (as in their Figure 3). This may be desirable in some experiments. However, in many experiments all the pairwise comparisons are equally important (e.g. in comparison of genotypes), there is an expected ordering (e.g. the single knock-outs closer to wild-type than the double mutants), or there is a factorial arrangement, which indicates which

Table 3. *Numerical variance of pairwise contrasts for one-factor replicated loop design with eight treatments. We show the numerical values when ρ_i is 0·25, 0·5 and 0·75, respectively. Notice that the optimal design will depend on which comparisons are of most interest. Two identical loops is most powerful for comparisons for treatments which are adjacent on the outer loop, but the interwoven design illustrated in Fig. 5b is preferable if all pairwise comparisons are equally important. There is little difference among the designs when ρ_i is small*

ρ_i	D	Two identical loops	Interwoven (5a)	Interwoven (5b)
0·75	1	$0 \cdot 724\sigma_\varepsilon^2$	$0 \cdot 760\sigma_\varepsilon^2$	$0 \cdot 781\sigma_\varepsilon^2$
	2	$1 \cdot 046\sigma_\varepsilon^2$	$0 \cdot 789\sigma_\varepsilon^2$	$0 \cdot 875\sigma_\varepsilon^2$
	3	$1 \cdot 181\sigma_\varepsilon^2$	$0 \cdot 937\sigma_\varepsilon^2$	$0 \cdot 781\sigma_\varepsilon^2$
	4	$1 \cdot 217\sigma_\varepsilon^2$	$0 \cdot 941\sigma_\varepsilon^2$	$0 \cdot 875\sigma_\varepsilon^2$
0·5	1	$0 \cdot 634\sigma_\varepsilon^2$	$0 \cdot 667\sigma_\varepsilon^2$	$0 \cdot 688\sigma_\varepsilon^2$
	2	$0 \cdot 804\sigma_\varepsilon^2$	$0 \cdot 687\sigma_\varepsilon^2$	$0 \cdot 750\sigma_\varepsilon^2$
	3	$0 \cdot 848\sigma_\varepsilon^2$	$0 \cdot 773\sigma_\varepsilon^2$	$0 \cdot 688\sigma_\varepsilon^2$
	4	$0 \cdot 857\sigma_\varepsilon^2$	$0 \cdot 774\sigma_\varepsilon^2$	$0 \cdot 750\sigma_\varepsilon^2$
0·25	1	$0 \cdot 564\sigma_\varepsilon^2$	$0 \cdot 590\sigma_\varepsilon^2$	$0 \cdot 594\sigma_\varepsilon^2$
	2	$0 \cdot 635\sigma_\varepsilon^2$	$0 \cdot 593\sigma_\varepsilon^2$	$0 \cdot 625\sigma_\varepsilon^2$
	3	$0 \cdot 644\sigma_\varepsilon^2$	$0 \cdot 630\sigma_\varepsilon^2$	$0 \cdot 594\sigma_\varepsilon^2$
	4	$0 \cdot 645\sigma_\varepsilon^2$	$0 \cdot 630\sigma_\varepsilon^2$	$0 \cdot 625\sigma_\varepsilon^2$
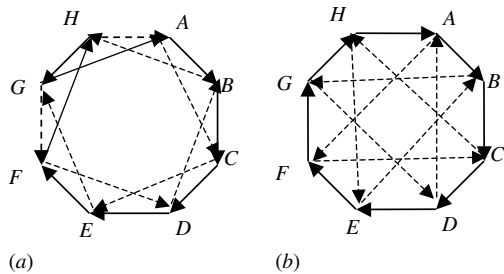
Fig. 5. Two interwoven loop designs with eight treatments. Each type of line indicates one complete loop.

(a) $AB \to aB$ (b) $AB \to ab$ (c) $AB \to ab$

$\uparrow$ $\downarrow$ $\uparrow$ $\downarrow$ $\uparrow$ $\downarrow$

$Ab \leftarrow ab$ $Ab \leftarrow aB$ $aB \leftarrow Ab$
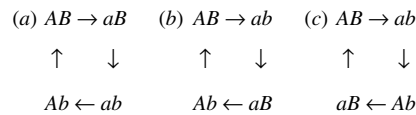
Fig. 6. A two-factor experiment with two levels of factor 1 (*A* and *a*) and two levels of factor 2 (*B* and *b*). Each pair of letters represents a treatment with two independent biological replicates, each of which has the indicated levels of the factor.

comparisons are of most interest and hence require the smallest variance. Glonek & Solomon (2004) discuss the selection of designs for two-channel microarrays when some comparisons are of particular interest.

The estimation of the variance of contrasts provided by equation (1) allows the investigator to select an appropriate design from those available. For example, in a study involving eight tissues in one loop, the investigators chose to place the most similar tissues adjacent on the loop to provide the most power to detect differences between tissues that may differ the least. However, when the opportunity arose to obtain another loop, the two interwoven loop designs depicted in Fig. 5 were examined. The variance of comparisons of treatment pairs at distance D is given in Table 3 for several values of ρ_i. The investigators

selected design 5b, as the variances of pairwise comparisons among tissues are almost equal for all D. If comparisons among tissues at $D = 1$ and $D = 2$ had been deemed to be of more importance, design 5a would be preferable.

It is also useful to notice that interwoven loop designs are actually composed of several loops. For example, the design in Fig. 5b has the loops ABCDEFGH and AFCHEBGD. The two loops in Fig. 5a are harder to see. They are: ABCDEFHG and ACEGFDBH.

6. Factorial designs

So far, this paper has considered pairwise comparisons, which are often of interest in one-factor designs. However, two-factor (and even three-factor) designs, such as genotype or drug (or both) and time are often done. In this case, the main effects and interactions are of interest.

Factorial treatment designs can readily be incorporated into loop designs. For example, a two-factor design with two levels per factor (*Aa* and *Bb*) and main effects of primary interest could be laid out as in Fig. 6.

The variances of the effects are listed in Table 4. The three contrasts are:

main effect of A: $(AB + Ab)/2 - (aB + ab)/2$
main effect of B: $(AB + aB)/2 - (Ab + ab)/2$
interaction: $(AB + ab)/2 - (Ab + aB)/2$

We say that the effect is on the diagonal of the design if the pairs of arrays which are averaged are on the diagonal of the square. Otherwise the effect is adjacent. The interaction, B main effect and A main effect are diagonal in designs 6a, 6b and 6c, respectively. From Table 4, we can see that the diagonal effect has the smallest variance. As in the previous sections, replication of the design by choice of two different loops can serve to equalize the variance of the effects if more samples and arrays are available and more precision is required.

For designs with more factors and more levels, there are many more choices of loop arrangements. Also, the main effects and interactions are defined by orthogonal sets of comparisons. Thus, optimal

Table 4. *The variance of the effects in a 2 × 2 factorial design in a single loop*

		$\rho_i = 0{\cdot}75$	$\rho_i = 0{\cdot}5$	$\rho_i = 0{\cdot}25$
Adjacent	$\dfrac{\sigma_{ia}^2 + \sigma_{i\varepsilon}^2}{2} - \dfrac{\sigma_{ia}^4}{8(\sigma_{ia}^2 + \sigma_{i\varepsilon}^2)}$	$1{\cdot}719\sigma_{i\varepsilon}^2$	$0{\cdot}928\sigma_{i\varepsilon}^2$	$0{\cdot}656\sigma_{i\varepsilon}^2$
Diagonal	$\dfrac{\sigma_{ia}^2 + \sigma_{i\varepsilon}^2}{2} - \dfrac{\sigma_{ia}^2}{4}$	$1{\cdot}25\sigma_{i\varepsilon}^2$	$0{\cdot}75\sigma_{i\varepsilon}^2$	$0{\cdot}583\sigma_{i\varepsilon}^2$

arrangements of the treatments around the loop can be determined by examining the variance of a set of orthogonal contrasts to define the effects. For example, in a two-factor design where A has two levels (Aa) and B has three levels ($Bb\beta$) the effects are defined by:

A main effect: $(AB + Ab + A\beta)/3 - (aB + ab + a\beta)/3$
B main effect: $(AB + aB)/2 - (Ab + ab)/2$
$(AB + aB)/2 + (Ab + ab)/2 - (A\beta + a\beta)$
AB interaction: $(AB - aB)/2 - (Ab - ab)/2$
$(AB - aB)/2 + (Ab - ab)/2 - (A\beta - a\beta)$

The variance of each contrast is readily computed from equation (1) and a loop (or replicated loop) can be chosen to minimize the variance of the effects of most interest. Alternatively, we may not be interested in the particular contrasts, but rather in the main effects and interactions. In this case, we may consider optimizing the general linear F-test for the effect or using one of the optimality criteria based on the variance of the contrasts comprising the effect (Shah & Sinha, 1989). The F-test and optimality criteria are invariant to the choice of orthogonal contrasts chosen to represent the effects (Hua, 2005), and hence the optimal design is unambiguously defined.

7. Incorporating biological replicates and blocks

Model (1) assumes that the only source of correlation in the study is due to the two channels on the microarray. However, often there are sources of variation that induce correlation between arrays, such as subsamples from the same individual hybridized to multiple arrays (technical replication), or blocking factors such as multiple treatments measured on the same individual (e.g. several tissues), multiple samples from the same cell line in a time course experiment, or arrays processed in batches due to constraints on personnel or equipment.

(i) *Technical replication*

Technical replication is often done in the form of dyeswap. This means that each RNA sample is split into two, and the subsamples are labelled with different dyes. For example, in the original Kerr & Churchill formulation of the loop design (2001 *a*), each loop
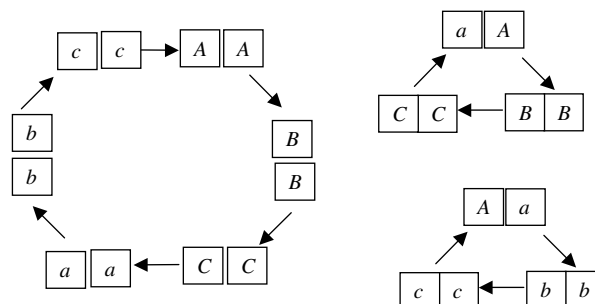

Fig. 7. Two ways of depicting a design with three treatments, two biological replicates per treatment and two technical replicates per biological replicate. Each boxed letter depicts a single sample.

includes two technical replicates for each condition and no biological replication. Such a design is adequate for making inferences for the samples at hand, but for inference about the biological population, biological replication is required. Hence, in what follows, we assume that there is both biological and technical replication, with the technical replicates done in dye-swap pairs. The discussion in this section is limited to two technical replicates per sample, as this is most common and illustrates the issues. The final recommendation, however, is to use each set of technical replicates for a loop; this is readily extended to any (even) number of technical replicates with complete dye-swap.

When we have both biological and technical replication, arrow diagrams for representing designs can be misleading. For example, Fig. 7 shows two different arrow diagrams for a design with three treatments (different letters), each of which has two biological replicates (lower and upper case letters) with dyeswap technical replicates. The two diagrams define the same set of arrays, and hence depict the same hybridization design. (Unlike earlier figures, each letter in Fig. 7 depicts a single sample.) Even with only three treatments and two biological replicates per treatment, and keeping to the principle that each array has two different treatments, there are 10 hybridization designs to consider (not counting permutations of the treatment names), and the number quickly grows with the number of treatments and biological replicates.
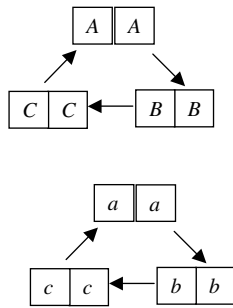
Fig. 8. A hybridization design with one biological replicate and its dye-swap technical replicate in each loop. Each letter depicts a single sample.
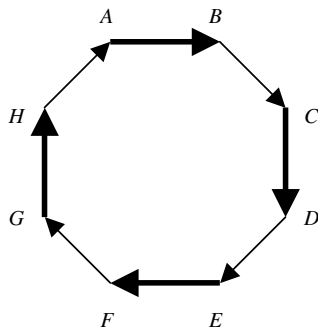

Fig. 9. An eight-treatment loop design in two complete blocks. The bold lines are one block and the thin lines are the other.
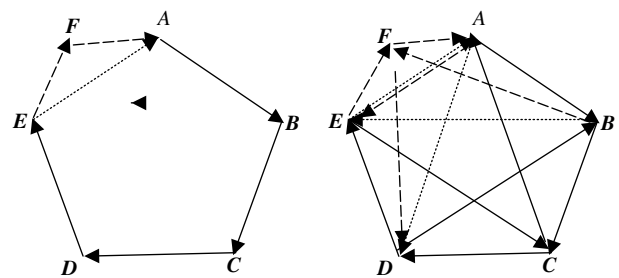

Fig. 10. Adding treatment *F* to a five-treatment loop design or woven loop design. The continuous lines are the arrays which are retained. The dashed lines are the added arrays. The dotted lines are arrays which may be removed from the analysis if a balanced design is desired.

A reference design, while requiring twice as many arrays to accommodate the same number of biological and technical replicates as 'loop-like' designs, is readily performed. Alternatively, in the spirit of Wit *et al.* (2005), it is possible to enumerate all the designs, compute the theoretical variance of the pairwise comparisons and thus select the optimal design.

As an alternative to these extremes which maintains the cost-effectiveness of the loop design, we suggest that each biological replicate be used for one loop (Fig. 8). By enumerating the designs, we can see that for two treatments the two possible designs {*AB,BA,ab,ba*} and {*AB,bA,Ba,ab*} have the same variance of pairwise comparison. The design depicted in Fig. 8 is optimal for three treatments (although there are other designs with the same variance). When we reach four or more treatments, we need to consider replicated loop designs as in Section 5. However, designs in which each loop is a single biological replicate will continue to have good properties, because they are partially balanced for the effects of technical replication and of array, which means that many random effects cancel when considering pairwise comparisons of the treatment effects.

(ii) *Complete block designs*

Experimental material may come in clusters that are more similar within cluster than between, or may be

handled in batches. For example, in tissue comparison experiments, tissues dissected from the same specimen are likely to be more similar than tissues dissected from different specimens. Similarly, samples prepared on the same day, or using the same batch of reagent, are likely to have similarities induced by handling compared with samples produced on different days or using different reagents. Groups of similar samples are called blocks.

It is very desirable to balance the design so that block effects are not confounded with the effects of interest. When feasible, it is best to have a measurement for every condition in every block – that is, a complete block design. It is very simple to incorporate complete block designs in loop designs. When *T* is even, the smallest available complete block has *T*/2 arrays. This is illustrated in Fig. 9 for *T*=8. The thick arrows indicate block 1 and the thin arrows indicate block 2. Clearly the pattern can be repeated for each loop in the design. When *T* is odd, the smallest available complete block is an entire loop.

8. Adding treatments

In a reference design, it is quite obvious how to add a treatment to the design – just include new arrays with the new treatment. Usually arrays will be added in pairs to accommodate dye-swapping.

It is equally simple to add a treatment to a loop design. It is simply spliced in as illustrated in Fig. 10, where condition *F* is added to a five-treatment loop. For a loop design, as in a reference design, two arrays are required to maintain dye balance for the new treatment. The balance of the design at the splice point is readily maintained by removing array *AE* from the experiment, or adding a dye-swap of this array. For woven loop designs, to maintain the woven pattern it is usually necessary to hybridize one or two extra arrays, as shown in Fig. 8*b*, to accommodate the increase in distance between treatments adjacent to the new treatment.

9. Missing data

Loss of an array in a loop design is problematic primarily due to the loss of balance for the dye effect. Of course, if a whole array fails, the investigator is likely to take another sample and redo the array. More commonly, some spots fail on some arrays, creating a loss of balance for some genes. The ANOVA computations automatically adjust for such imbalance, but the loss of information can severely affect statistical power. Each 'lost' spot loses information about two treatments for the affected gene. In reference designs, a 'lost' spot also causes imbalance in the dye effect, but loses information only for the single treatment on the array. Recall, however, that the reference design has only about half as much information per array, so that missing spots have about the same effect on both designs. The robustness of microarray experiments to spot-loss is discussed in Latif (2005).

10. Discussion

While the reference design can be useful when the primary objective of the microarray experiment is comparison of several treatments with a control, when the objective is comparison of treatments with each other, it is inefficient compared with designs which have two treatments on each array. Replicated loop designs are very convenient and efficient designs for handling this latter situation. In this paper, we have tried to show that these designs are more powerful than reference designs and can readily be adapted to complex experiments such as factorial designs including time course studies, incorporating additional replication, and incorporating complete blocks. Loop and interwoven loop designs are not the only possibilities – the set of comparisons of interest to the investigators should drive the design. Mixed models and the principles of optimal design can usefully be employed to improve the efficiency of microarray experiments, by allowing the experimenter to set up the hybridization design to minimize the variance of comparisons of most interest.

This paper has not discussed the analysis of the resulting data. Clearly, since the design is based on mixed model (1) or (2) the resulting data can be fitted by mixed-model ANOVA, using the normalized data from each channel. This is readily extended to more powerful Bayes and empirical Bayes methods such as LIMMA (Smyth *et al.*, 2005). Permutation methods such as SAM (Tusher *et al.*, 2001) can also be adapted for the single-channel analysis of loop and related designs, but permutations must be restricted to take into account the pairing of the observations on the arrays.

References

Cui, X., Hwang, J., Qiu, J., Blades, N. J. & Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75.

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

Glonek, G. F. V. & Solomon, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**, 89–111.

Hua, J. (2005). Extending the loop design for two-channel microarray experiments. MS thesis, Dept of Statistics, Pennsylvania State University.

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

Kerr, M. K. & Churchill, G. A. (2001*b*). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123–128.

Landgrebe, J., Bretz, F. & Brunner, E. (2004). Efficient Two-Sample Designs for Microarray Experiments with Biological Replications. *In Silico Biology* **4**, 461–470.

Latif, A. H. M. M. (2005). Efficiency and robustness issues in complex statistical designs for two-color microarray experiments. PhD thesis, University of Göttingen.

Shah, R. K. & Sinha, B. K. (1989). *Theory of Optimal Designs*. Heidelberg: Springer.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1), article 3. http://www.bepress.com/sagmb/vol3/iss1/art3

Smyth, G. K. (2005). Individual channel analysis of two-colour microarray data. Invited session IPM 11: *Computational Tools for Microarray Analysis, 55th Session of the International Statistics Institute*, Sydney. http://www.statsci.org/smyth/pubs/ISI2005-116.pdf

Smyth, G. K., Michaud, J. & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067–2075.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the USA* **98**, 5116–5121.

Wit, E., Agostino, N. & Khanin, R. (2005). Near-optimal designs for dual-channel microarray studies *Applied Statistics* **54**, 817–830.