BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 4 2005, pages 492–501
doi:10.1093/bioinformatics/bti022

An experimental evaluation of a loop versus a reference design for two-channel microarrays

V. Vinciotti[1,*,†], R. Khanin[2], D. D'Alimonte[3], X. Liu[1], N. Cattini[4], G. Hotchkiss[4], G. Bucca[4], O. de Jesus[5], J. Rasaiyaah[5], C. P. Smith[4], P. Kellam[5] and E. Wit[2,†]

[1]Department of Information Systems and Computing, Brunel University, Uxbridge UB8 3PH, UK, [2]Department of Statistics, University of Glasgow, Glasgow G11 8QW, UK, [3]Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, UK, [4]School of Biomedical and Molecular Sciences, University of Surrey, Guildford GU2 7XH, UK and [5]Virus Genomics and Bioinformatics Group, Department of Infection, University College London, London W1T 4JF, UK

Received on April 15, 2004; revised on June 28, 2004; accepted on September 7, 2004
Advance Access publication September 16, 2004

ABSTRACT

Motivation: Despite theoretical arguments that so-called 'loop designs' for two-channel DNA microarray experiments are more efficient, biologists continue to use 'reference designs'. We describe two sets of microarray experiments with RNA from two different biological systems (TPA-stimulated mammalian cells and *Streptomyces coelicolor*). In each case, both a loop and a reference design were used with the same RNA preparations with the aim of studying their relative efficiency.

Results: The results of these experiments show that (1) the loop design attains a much higher precision than the reference design, (2) multiplicative spot effects are a large source of variability, and if they are not accounted for in the mathematical model, for example, by taking log-ratios or including spot effects, then the model will perform poorly. The first result is reinforced by a simulation study. Practical recommendations are given on how simple loop designs can be extended to more realistic experimental designs and how standard statistical methods allow the experimentalist to use and interpret the results from loop designs in practice.

Availability: The data and R code are available at http://exgen.ma.umist.ac.uk

Contact: veronica.vinciotti@brunel.ac.uk

1 INTRODUCTION

A common aim of many microarray studies is to detect the genes in a biological system that are differentially expressed across a number of conditions of interest. In a typical two-channel DNA microarray experiment, mRNAs from biological samples under two different conditions are labelled with a green (Cy3) and a red (Cy5) dye, respectively, and then hybridized onto an array of complementary probes. After hybridization, a measure of red and green intensities for each spot provides an indication of the amount of mRNAs produced by the corresponding gene under the two conditions. A higher intensity for one condition over the other for one spot indicates that the corresponding gene was particularly active under that condition.

As DNA microarray experiments are becoming larger, involving larger number of samples and conditions, it is important to design experiments in the most efficient way in order to obtain precise estimates of the biologically important parameters. Wit and McClure (2004) provide a comprehensive overview of the various issues that need to be addressed when designing microarray experiments. The objective is to design the experiment in such a way as to minimize the effect of unwanted variation, while increasing the precision of the estimates of the parameters of interest, the changes in gene expression from one condition to another.

In this paper, we focus mainly on the problem of how to assign samples efficiently to microarrays, given a number of conditions we wish to compare and a fixed number of available arrays. The most commonly used design within the biological community is the so-called reference design. In this design, each condition of interest is compared with samples taken from some standard reference. As the reference is common to all the arrays, this design allows an indirect comparison between the conditions of interest. The main criticism raised to this approach is that 50% of the hybridization resources are used to produce a control or common reference signal of no intrinsic interest to the biologists. This reference signal is in effect processed out of the final analysis following normalization. In contrast,

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first and last authors should be regarded as joint First Authors.

a loop design compares two conditions via a chain of other conditions, thereby removing the need for a reference sample.

The aim of this study is to compare empirically these two commonly used two-channel microarray designs, the loop and the reference design. Most theoretical papers on microarray design argue that the loop design of microarray experiments is more efficient than the reference design (Landgrebe *et al*., 2004; Churchill, 2002; Glonek and Solomon, 2004; Kerr and Churchill, 2001; Khanin and Wit, 2004; Yang and Speed, 2002). Despite these theoretical advantages and some occasional examples of loop-type designs in practice (Townsend and Hartl, 2002; Townsend *et al*., 2003), there is a tendency to continue using the reference design, as evidenced by recently included studies in the Stanford microarray database (Chang *et al*., 2004; Lapointe *et al*., 2004; Pathan *et al*., 2004). A second aim of the paper is to show how elementary matrix algebra can make such loop designs more accessible to biologists. These technical details as well as a numerical example are described in Section 3.

In the present study, two sets of microarray experiments were conducted. Two entirely different biological systems (one eukaryotic and one prokaryotic) were examined, both comprising three sampling points in a time-series experiment: (1) The human B-cell lymphoma cell line Ramos, where ester tetradecanoyl phorbol acetate (TPA) was used to stimulate protein kinase C activity and (2) the mycelial growth of *Streptomyces coelicolor* bacterium cultivated on agar plates. Throughout the paper, we refer to these studies as the B-cell and *Streptomyces* studies, respectively. In each study, both a loop and a reference design were performed using the same RNA preparations to allow direct comparison of the output of the two experimental designs. This is the first time the two types of designs have been evaluated side-by-side experimentally. Both experiments are described in more detail in Section 2.

In Section 4, the results of the two designs for each of the studies are presented in two ways. That is, by comparing the standard errors of the parameter estimates for both studies as well as by plotting the fraction of differentially expressed genes for different cut-offs. This latter method is related to the theoretical receiver operating characteristic (ROC) curve, which for completeness is described in the same section by means of a simulation study. In Section 5, we discuss several issues that are closely related to the comparison of loop versus reference designs. First, we discuss the impact of considering only the individual channel data and not the log-ratios. Second, as the two studies involve only special cases of loop and reference designs, some hints are given on how these designs can be extended to situations with a larger number of conditions. Finally, we consider the practical issue of array failures and question the supposed robustness of reference designs towards these.

2 MICROARRAY EXPERIMENTS

On a two-channel microarray, it is possible to compare directly two conditions. The need for a more complicated design arrangement becomes necessary when there are at least three conditions, as it is impossible to compare all conditions on the same array. In this case, one can compare the efficiency of a loop design versus a reference design.

In this section, we describe the two experiments that we conducted to compare the two designs. Each of them considers three time-points in the development of two very different organisms, a human B-cell lymphoma cell line and a *S.coelicolor* bacterium.

2.1 *Streptomyces coelicor*

Streptomyces coelicolor is a complex Gram-positive bacterium which undergoes developmental changes, producing spore chains from branching mycelium and secondary metabolites such as antibiotics in the late stages of its development. The three RNA samples in this study are taken from a wild-type strain grown on cellophane-coated agar plates and harvested at time-points representing early, mid and late stages of the development.

Figure 1 summarizes the *Streptomyces* microarray experiment. Each hybridization pair was carried out in triplicate, with the dyes swapped on one of the array plates. The experiments associated with the two designs used the same number of slides to allow for a fair comparison. Genomic DNA (gDNA) from *S.coelicolor* was used as the reference sample in the reference design. The microarray batch used (SCp14) contained 7337 probes, representing 7337 *S.coelicolor* genes. To facilitate direct comparison of the loop and reference designs, the same labelled preparations of cDNA were divided equally between the loop and reference arrays. Details of the microarrays used and the protocols for RNA isolation, cDNA labelling and microarray hybridization are given at http://www.surrey.ac.uk/SBMS/Fgenomics/Microarrays.

2.2 Human B-cell lymphoma cell line

The B-cell lymphoma line Ramos was induced with TPA, a chemical that stimulates the activity of the protein kinase C. This protein is an upstream mediator of the herpesvirus-8-induced ERK signalling pathway. Samples were taken at 0, 2 and 4 h after the induction.

In the B-cell study six microarrays are available for both the loop and reference design experiments. As a result, a similar design to the one in the *Streptomyces* study represented in Figure 1 is used, except that arrays a_1, a_4 and a_7 are omitted. Each hybridization pair was carried out as a duplicate dye-swap. Total RNA was extracted at each of the three time-points and hybridized to Human Gen2 cDNA microarrays (http://www.hgmp.mrc.ac.uk). The microarray contains approximately 5400 probes corresponding to 3360 known human clones, 768 from the Mammalian Gene Collection and several others. For the reference design,

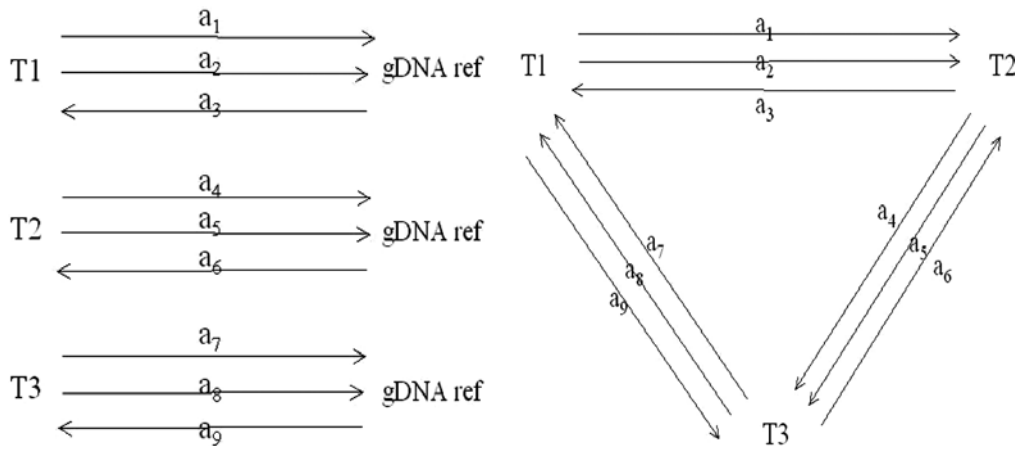

Fig. 1. Reference and loop designs used on the *Streptomyces* study. A line, indicated by a_i, represents a direct hybridization between the two samples on array i. The arrow goes from the Cy3 to the Cy5 channel.

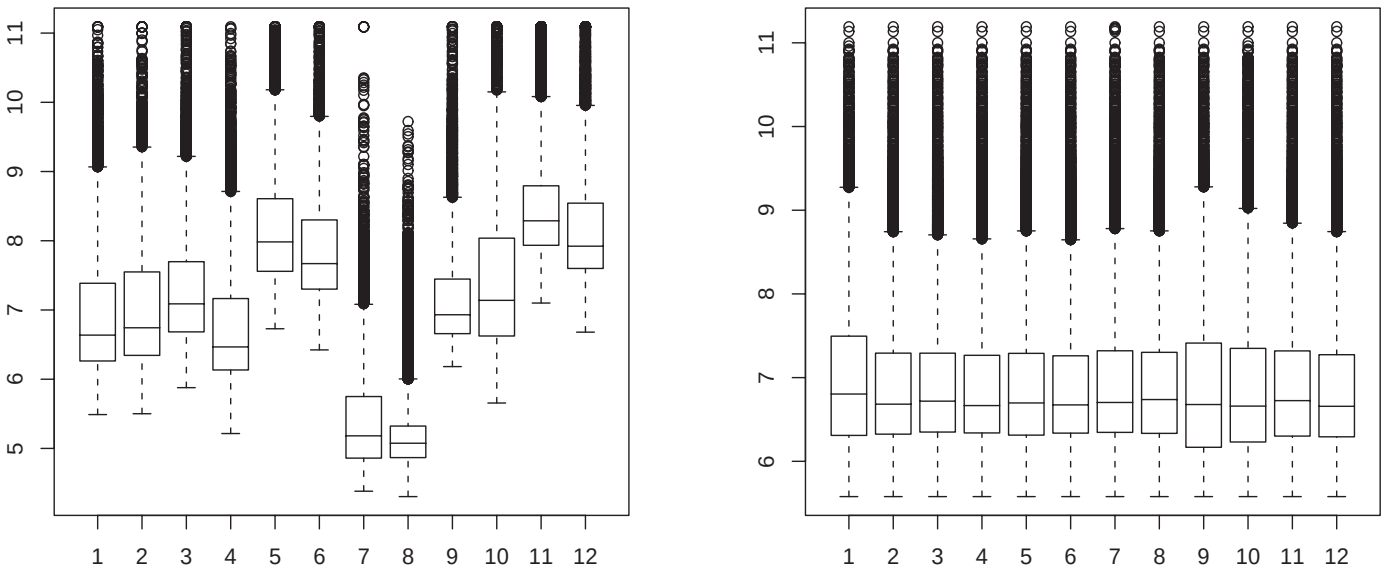

Fig. 2. Effect of across-array normalization in the B-cell reference design study across the 12 channels of its 6 arrays.

a common reference total RNA pool from several cell lines was used.

3 METHODS

3.1 Data normalization

Prior to the analysis of the data, normalization procedures were performed to remove artefacts from the data that are due to non-specific effects. The method used is described in detail in Wit and McClure (2004) and is available in the R-library `smida`. Essentially, we correct for various artefacts, such as spatial, background, dye and across-array effects. The normalization procedures are applied in a sequential manner, starting with local corrections and proceeding towards more global corrections like across-array normalization. Figure 2 shows the effects of across-array normalization in the B-cell study.

3.2 Parameter estimation

We assume that the amount of transcribed RNA is approximately proportional to its spot intensity, whereby the constant of proportionality may depend on the particular spot itself. By defining gene expression as the normalized log-intensity of a spot associated with a particular gene, the difference between the gene expressions of the two conditions in one spot is equal to the log-difference of the transcribed mRNA as the constant of proportionality cancels out. In microarray experiments, the parameters of interest are the changes in gene expression from one condition to another.

Here, we present a general methodology to process gene expression data that utilizes all the available information to produce estimates of the parameters of interest. Such a method is indispensable when a loop design is used, since log-ratios on each slide are not directly comparable, as they can be log-ratios of many combinations of conditions. In the reference design, two time-points can be compared via their associated log-ratios.

For each gene, we denote its true expression value at condition t by θ_t. For simplicity, we avoid referring to the specific gene in the notation. An observation y_{jk} is the log-ratio of condition j and condition k, that is $\log(z_j/z_k)$, where z_t is the observed intensity at condition t. For a loop design, these conditions are time-points, whereas for a reference design one of the two conditions is the reference. Wit and McClure (2004) argue that under a wide range of circumstances the variable y_{jk} is normally distributed

$$y_{jk} \sim \mathcal{N}(\mu_{jk}, \sigma^2), \quad \mu_{jk} = \theta_j - \theta_k. \tag{1}$$

Here, μ_{jk} is the true expression difference between conditions j and k. One central assumption is that the variance does not depend on the conditions involved, although we allow for the real possibility that the expression variance is gene-dependent or even design dependent. For each gene a vector of n observations $y = (y_{a_1}, \ldots, y_{a_n})$, obtained on the n arrays $a_1, \ldots, a_n$, can be represented as

$$y = X\mu + \epsilon, \tag{2}$$

where X is the *design matrix* defining the relationship between the values observed in the experiment and a set of independent parameters, μ and ϵ is a vector of independent, normally distributed, zero-mean errors. For an experiment with T conditions, we arbitrarily choose the parametrization μ:

$$\mu = (\mu_{12}, \mu_{13}, \ldots, \mu_{1T}).$$

Any of the other contrasts can be obtained by the relation $\mu_{ij} = \mu_{1j} - \mu_{1i}$.

The goal is to obtain estimates of the true expression differences, $\hat{\mu}_{jk}$, separately for each gene. Given the assumptions behind the linear model, the maximum likelihood estimates for the differences μ are

$$\hat{\mu} = (X^t X)^{-1} X^t y. \tag{3}$$

From these, any other contrast can be estimated by $\hat{\mu}_{ij} = \hat{\mu}_{1j} - \hat{\mu}_{1i}$.

For the three time-point experiments considered in this paper, the parameters of interest are the differences μ_{12}, μ_{13} and μ_{23}. We will go through a simple example to show how these parameters are estimated by Equation (3) when using a reference and a loop design. Table 1 gives the log-ratios for a particular gene for the two designs from the *Streptomyces* study. The design matrix of the loop design for this problem is given by

Table 1. Log-ratios for gene SC02348 across the nine arrays of the reference and loop designs from the *Streptomyces* study

Reference			Loop		
Cy3	Cy5	y_{a_i}	Cy3	Cy5	y_{a_i}
T1	R	−0.510	T1	T2	−0.005
T1	R	−0.370	T1	T2	−0.236
R	T1	0.633	T2	T1	−0.038
T2	R	−0.424	T2	T3	0.047
T2	R	−0.250	T2	T3	0.269
R	T2	0.468	T3	T2	−0.031
T3	R	−0.374	T3	T1	−0.139
T3	R	−0.774	T3	T1	−0.283
R	T3	0.667	T1	T3	0.082

$$X_L = \begin{pmatrix} 1 & 1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 & -1 & -1 & 1 \end{pmatrix}^{\mathrm{t}} \tag{4}$$

and the estimates for the expression ratios by

$$\begin{pmatrix} \hat{\mu}_{12}^{(L)} \\ \hat{\mu}_{13}^{(L)} \end{pmatrix} = \left(X_L^t X_L\right)^{-1} X_L^t y = \begin{pmatrix} -0.028 \\ 0.128 \end{pmatrix}.$$

From these, $\hat{\mu}_{23}^{(L)} = \hat{\mu}_{13}^{(L)} - \hat{\mu}_{12}^{(L)} = 0.156$. Similarly, the design matrix of the reference design is given by

$$X_R = \begin{pmatrix} 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}^{\mathrm{t}} \tag{5}$$

and the estimates for the expression ratios by

$$\begin{pmatrix} \hat{\mu}_{12}^{(R)} \\ \hat{\mu}_{13}^{(R)} \\ \hat{\mu}_{1r}^{(R)} \end{pmatrix} = \left(X_R^t X_R\right)^{-1} X_R^t y = \begin{pmatrix} -0.124 \\ 0.101 \\ -0.504 \end{pmatrix},$$

from which $\hat{\mu}_{23}^{(R)} = 0.225$.

4 RESULTS

4.1 Variability of estimates

The two models described above yield different estimates of the gene expression parameters. If, as under our assumptions, both sets of estimates are unbiased, then the best model is the one that produces the most precise estimates, that is, the estimates with the lowest variability. Given the assumptions behind the linear model in Equation (2), it follows that the contrast estimates for the design D are given as

$$\hat{\mu}^{(D)} \sim \mathcal{N}\left(\mu, \left(X_D^t X_D\right)^{-1} \sigma_D^2\right),$$

where X_D is the design matrix for the design D. In the case of the *Streptomyces* study, the design matrix for the reference

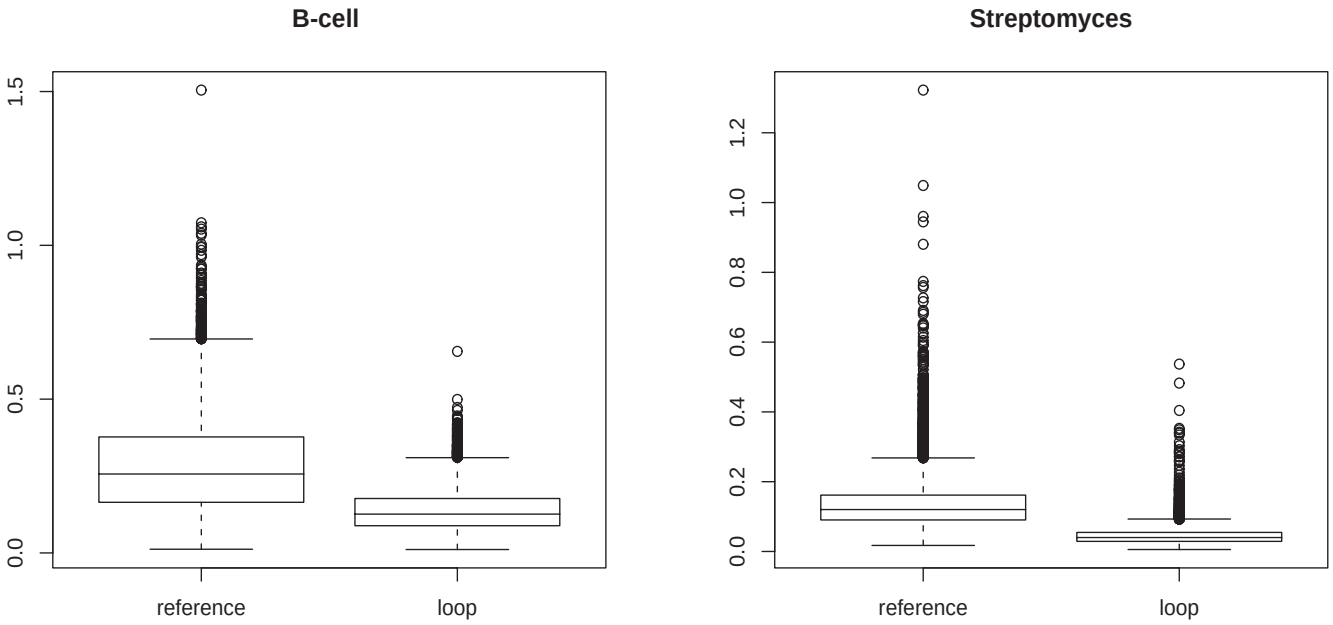

Fig. 3. Estimated standard errors $\sqrt{\hat{V}(\hat{\mu}_{12})}$ of the differential expression estimates of the first versus the second time-point for the two designs across the two biological systems.

design is given by Equation (5) and the one for the loop design by Equation (4). For the B-cell study, these matrices can be easily adapted by removing the first, fourth and seventh rows. In our formulation, the differential expression variance σ^2 is allowed to depend on the design. In theoretical comparisons, it is typically assumed equal across loop and reference design. This might unfairly favour loop designs, in the case where it is possible to use a very stable reference sample.

Let $\hat{V}\left(\hat{\mu}_{jk}^{(D)}\right)$ denote the estimated variance of the estimate $\hat{\mu}_{jk}^{(D)}$, obtained from the estimated covariance matrix $\left(X_D^t X_D\right)^{-1} s_D^2$. Figure 3 shows a box plot of the standard errors, $\sqrt{\hat{V}(\hat{\mu}_{12})}$, for the two designs of the two biological systems. It is clear from this figure that the parameter estimates obtained using the reference design have higher variability than when the loop design is used. The results for the other contrasts are similar.

The variances of the parameter estimates for all contrasts can be combined into an empirical measure of relative design efficiency. This is defined by

$$\sqrt{\frac{\sum_{\text{genes}} \sum_{\text{contrasts}} \hat{V}\left(\hat{\mu}_{jk}^{(L)}\right)}{\sum_{\text{genes}} \sum_{\text{contrasts}} \hat{V}\left(\hat{\mu}_{jk}^{(R)}\right)}}. \qquad (6)$$

This measure is in spirit similar to the so called A-optimality score, which is the sum of the variances of the parameter estimates up to a constant σ^2, which is assumed to be the same for different designs (Kerr *et al*., 2000). The theoretical relative design efficiency is defined as

$$\sqrt{\frac{\text{tr}\left(C_L \left(X_L^t X_L\right)^{-1} {C_L}^t\right)}{\text{tr}\left(C_R \left({X_R}^t X_R\right)^{-1} {C_R}^t\right)}},$$

where X_R and X_L are the design matrices for the reference and the loop design, respectively, and C_R and C_L are the matrices that transform the two designs to the same parametrization. For our experiments, these are the matrices satisfying $(\mu_{12}, \mu_{13}, \mu_{23})^{\text{t}} = C_R(\mu_{12}, \mu_{13}, \mu_{1r})^{\text{t}}$ and $(\mu_{12}, \mu_{13}, \mu_{23})^{\text{t}} = C_L(\mu_{12}, \mu_{13})^{\text{t}}$, respectively.

Table 2. Square root of the average estimated variance of the contrast estimates for the two designs across the two biological systems and a comparison of the empirical and theoretical relative design efficiencies

	B-cell	*Streptomyces*
Reference	0.572	0.279
Loop	0.274	0.091
Emp. rel. efficiency	0.479	0.326
Th. rel. efficiency	0.577	0.577

Table 2 reports the average standard error for the two designs and the empirical and theoretical design efficiencies for the two studies. It is intriguing that in both cases the empirical measure of relative efficiency is smaller than the theoretical measure. This means that the loop design in these two examples performs even better than expected theoretically.

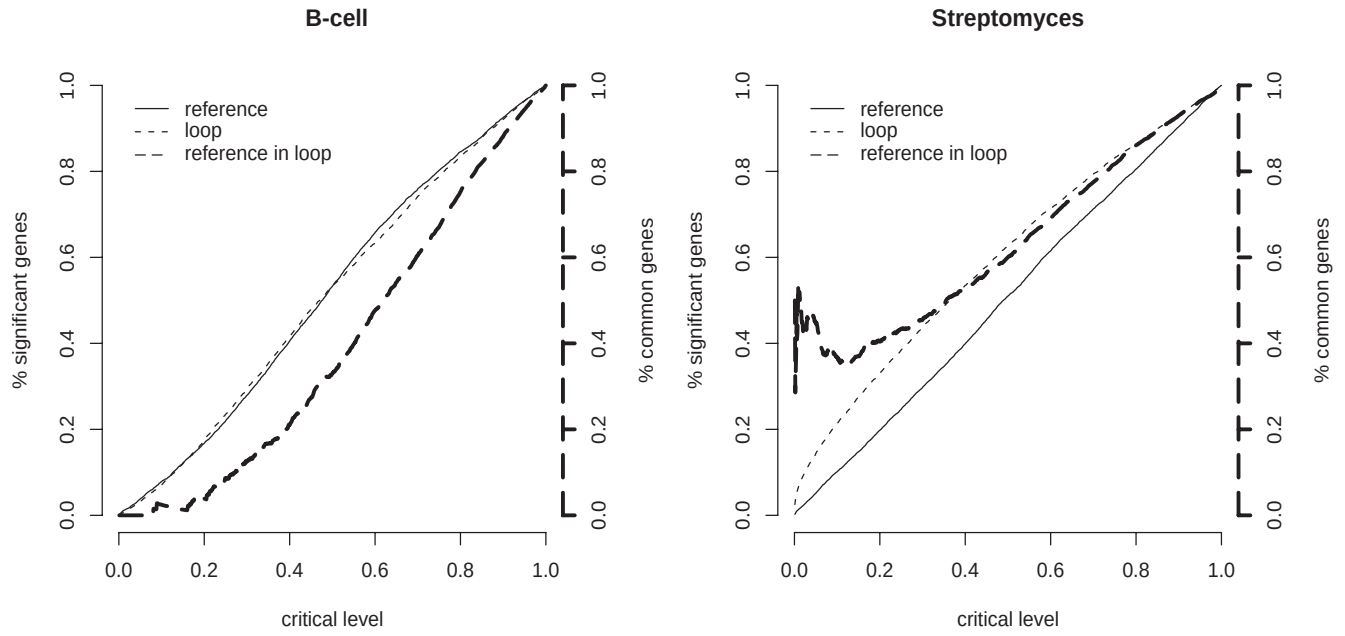

Fig. 4. Percentage of significant genes found by the two designs (solid and dotted lines) and percentage of significant genes found by both designs as a fraction of those found in the reference design (dashed line) in the B-cell and *Streptomyces* studies.

Apparently, the reference samples in both biological systems were less stable than any of the ordinary conditions.

4.2 Differentially expressed genes

For a further comparison of the two experimental designs, we have analyzed the genes for differential expression across time using the two methods described above. We use an F-test to find the genes for which $\mu = (\mu_{12}, \mu_{13})$ is significantly different from zero under either the loop or reference design. Under the assumption of normality, it follows that

$$\frac{\hat{\mu}^t C(X^t X)C^t \hat{\mu}}{s^2(p-1)} \sim F_{p-1,df} \tag{7}$$

where C is the matrix that transforms the design to the μ parametrization, s is the estimate of the standard error in the model, p is the number of conditions of interest in the design and df is the number of independent observations in the design minus the number of parameters that the design attempts to estimate. In our experiments, for both the reference and loop design it holds that $p = 3$, for the reference design $df = n - p$ and for the loop design $df = n - p + 1$, where n is the number of arrays.

Figure 4 summarizes the results obtained on the B-cell and *Streptomyces* studies. The plots show the percentage of significant genes found by the two methods for critical levels between 0 and 1. In the *Streptomyces* study, the plot shows that, for the same critical level, the percentage of genes found when using the loop design is higher than when the reference design is used. The dashed line on the same plot strengthens this result by showing a high percentage of the same genes found significantly expressed by both the loop and the reference designs in this study. These results show the advantages of using a loop design as compared with a reference design.

Interestingly, it seems that at no cut-off in the B-cell study more genes are detected than would be expected if none of the genes were differentially expressed. And consequently, there is a more or less random relationship between the number of genes detected by the reference and the loop designs. We conclude that in the B-cell study there are very few differentially expressed genes.

4.3 Simulation study

The comparative analysis conducted so far on the basis of the two pilot studies shows that the loop design is more efficient than the reference design. In this section, we complete the comparison of the two designs by conducting a simulation study.

We have simulated gene expressions for 3 conditions over 6 arrays and 100 genes, using a loop and a reference model. We have generated the data so that 30 genes were differentially expressed, with a mean expression different from zero drawn from a $\mathcal{N}(0,1)$ distribution. Furthermore, the simulation was repeated 100 times. As before, we used an F-test to detect the differentially expressed genes.

Figure 5 plots the true positive rate (proportion of active genes detected as active) versus the false positive rate (proportion of inactive genes falsely detected as active) as the critical

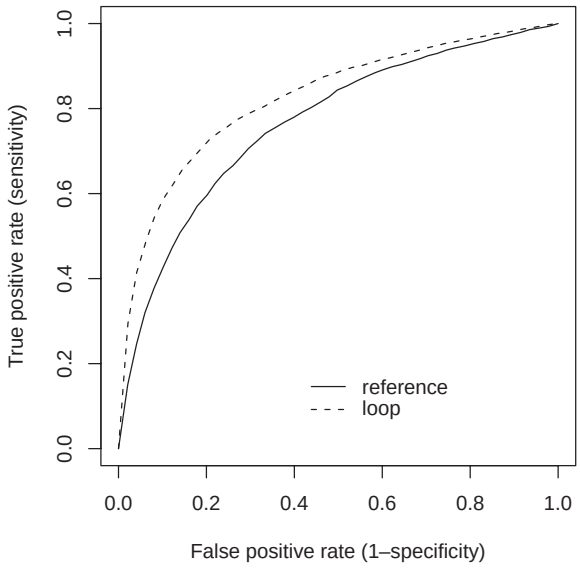

Fig. 5. ROC curve for the simulation study: 100 genes, of which 30 are differentially expressed, across 3 conditions using 6 arrays for each design.

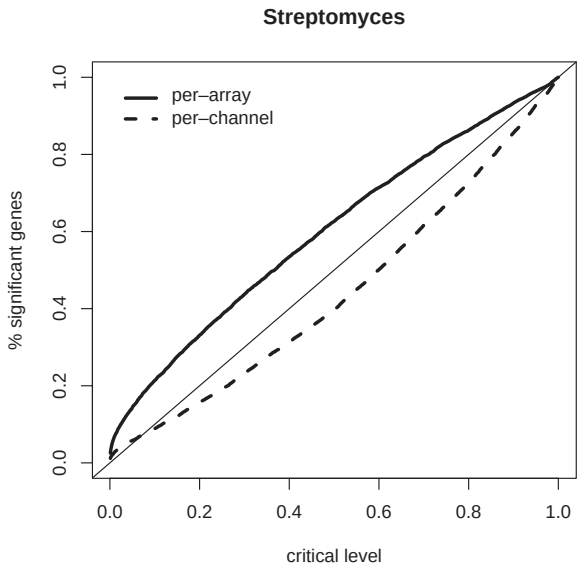

Fig. 6. Percentage of significant genes found when using the total expressions (dotted line) rather than the expression differences (solid line) in the *Streptomyces* study.

level increases from 0 to 1. The ROC curves show that the loop design detects the differentially expressed genes more accurately than the reference design. For any critical value, the loop design attains a higher true positive rate and a lower false positive rate than the reference design. This means that the loop design detects a higher proportion of differentially expressed genes while minimizing the proportion of mistakenly detected non-differentially expressed genes. That is, by designing the experiments in a more efficient way, one can obtain more precise answers to the biological questions of interest.

5 DISCUSSION

The results in this paper demonstrate that given the same number of microarrays the loop design provides more precise estimates of the parameters of interest than the reference design. The reason behind this is that in the loop design more resources are used for the measurement of the conditions of interest.

5.1 Alternative: using raw channel data

The estimation of the parameters from a loop design presented in this paper was based on the expression differences y_{jk}. This is a common starting point in microarray analysis, based on the belief that the spot intensities are only proportional to the RNA abundance. So by taking the ratios between the two channel intensities, one obtains the ratio of the RNA abundances, as the proportionality constant including a possible spot effect cancels outs.

A downside of this method is that by taking the ratio of the two channels one loses information about the gene expression variance if there are no significant spot effects. In this section, we consider the effects of working with the gene expressions, rather than the expression ratios. It implies estimating total expressions θ, rather than the expression differences μ.

For the loop design in Figure 1, ignoring explicitly possible spot effects results in the modelling equation

$$\begin{pmatrix} x_{1a_1} \\ x_{2a_1} \\ \vdots \\ x_{3a_9} \\ x_{1a_9} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{18} \end{pmatrix}, \qquad (8)$$

where x_{ta_i} denotes the total expression at time t measured on array a_i. From the estimates $\hat{\theta}_t$, one can obtain estimates of the differential expression parameters, via $\hat{\mu}_{jk} = \hat{\theta}_j - \hat{\theta}_k$.

The method described in Section 4.1 with the design matrix as in Equation (8) leads to an average estimated standard

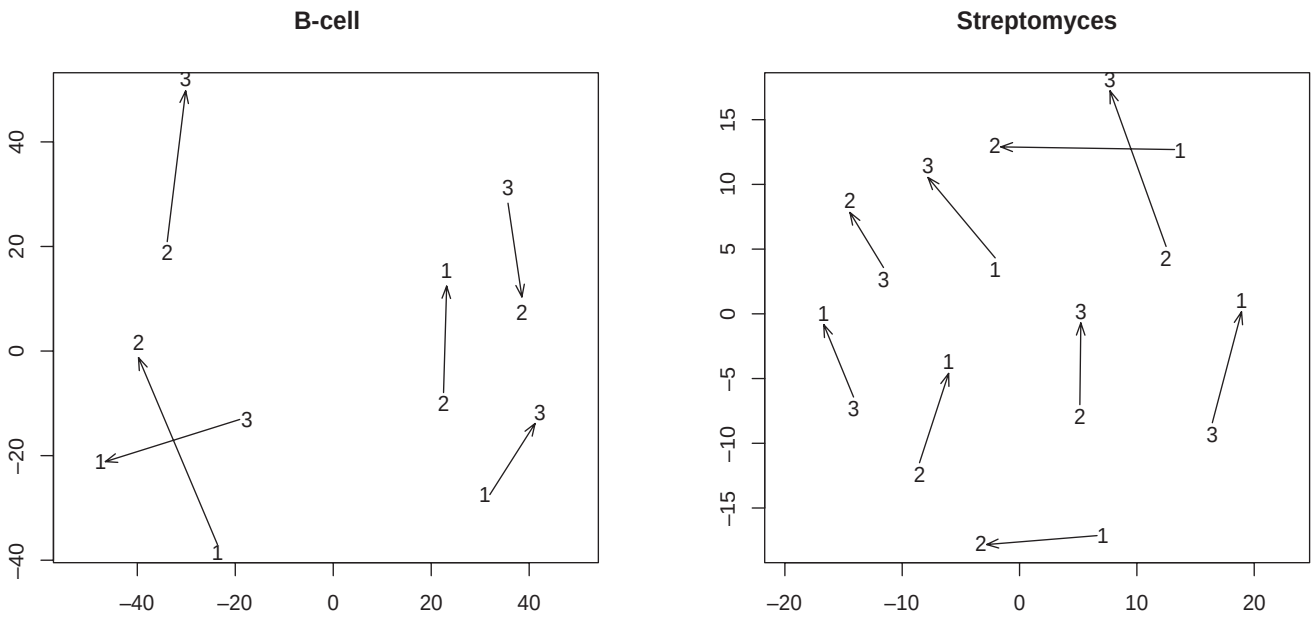

Fig. 7. Multidimensional scaling plots of the 12 and 18 channels of the B-cell and *Streptomyces* loop design study, respectively. An arrow refers to an array, with the arrowhead pointing to the condition (time 1, 2 or 3) in the Cy5 channel.

error of 0.481 for the herpesvirus study and of 0.158 for the Streptomyces study. Compared with the results in Table 2, this shows that treating each channel individually leads to a higher variance than when the log-ratios of the two channels are considered. Figure 6 strengthens this result. Here we used the F-test in Equation (7), with $p = 3$ and $df = 2n - p$, to detect the differentially expressed genes. The plot shows how, working with expressions rather than expression ratios, leads to an even worse performance than randomly labelling genes as differentially expressed.

These results suggest that the independent assumptions of the two channels are violated. It is most likely that in these microarrays the physical properties of the spots for the same gene vary from array to array. Such spot bias has the result of making different conditions applied to the same spot look more alike than the same conditions across different spots. It is important to note that the presence of spot bias makes the calculation of the standard errors completely moot, which results in a poor performance in detecting differentially expressed genes.

Multidimensional scaling plots, like the Sammon plots in Figure 7 (Sammon, 1969), can be used to check for the presence of spot effects. If the channels are truly independent, they should cluster according to conditions, whereas if there is some residual correlation of pairs of channels then they will cluster by array. In the two studies that we considered, the gene expressions across the two conditions on the same array tend to be more similar to each other than to the gene expressions for the same condition on different arrays, even after the data have been normalized.

5.2 Extension to large studies

A natural question arising from the study is how to extend the loop design for larger experiments. Assuming that all the conditions are equally important, designing an experiment that would directly compare all possible pairs of conditions would obviously require too many arrays. A more realistic design is needed.

Wit *et al*. (2004) have developed an optimization algorithm that efficiently searches for the loop design which minimizes the A-optimality criterion. The search is restricted to the family of interwoven designs. These designs guarantee that each condition is measured equally often by either dye. The optimization algorithm allows one to input the number of conditions one wants to compare and the number of arrays one can afford to hybridize. For example, Figure 8 shows the best interwoven loop design for the case of 30 conditions and 90 arrays.

Despite the high number of conditions involved in large experiments, getting estimates for the parameters of interest is no more difficult than for the smaller pilot studies investigated in this paper. The main point is to define the design matrix X of the study, which, as part of the simple linear model in Equation (2), describes the assumption that a measured value of gene expression is equal to its true value plus some normal random error. The estimates $\hat{\mu}$ returned by the maximum likelihood estimation in Equation (3) can be used to draw statistical conclusions on which genes have changed their expression across conditions. The better the statistical design, the more precise and reliable the biological answers will be.

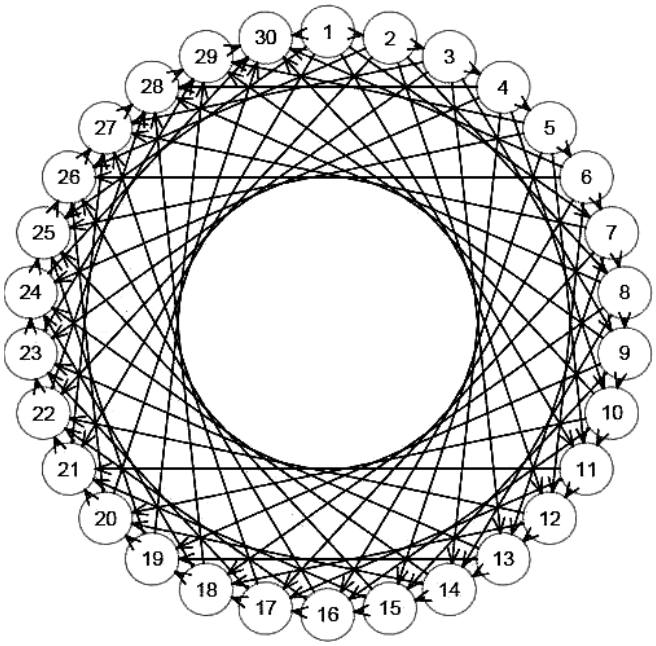

Fig. 8. A-optimal design for an experiment with 30 time-points and 90 arrays in the class of interwoven loop designs.

5.3 Practical issue: the array that did not work

An issue often raised by biologists is how a loop design would cope with the common situation when one array gets missed or damaged. The relative symmetry and simplicity of a reference design seems to become more attractive in this kind of scenario. In our experience, this argument in favour of the reference design often does not hold. For example, compare the very simple n-array loop design (one array for each contrast) to the n-array reference design. If one array fails in either design, then in the loop design all the contrasts are still estimable, whereas in the reference design all the contrasts that involve the condition in the failed array are not estimable anymore.

Multiple interwoven loops will make a loop design even more robust. For example, each condition in the design described in Figure 8 is measured six times. Random failure of even 20% (18 slides) of all slides is still unlikely to result in any contrast becoming unidentifiable. In contrast, in a reference design each condition is measured by only three slides and therefore this probability is much higher. Future work will look at precise mathematical formulations of this issue to obtain more general conclusions on the robustness of loop designs.

It is true that array failure will typically lead to imbalance in the design, but this is true for both loop and reference designs. Despite this imbalance, least squares estimates of the contrasts are still available, by eliminating those rows from the design matrix X that correspond to the missing arrays and then using Equation (3).

Although we do not recommend using a reference design, we do not advise against using a reference in the design. In fact, since all the parameters in the model are relative expressions, there are advantages in comparing all the conditions of interest to one stable condition: the parameters μ_{jk} will be more interpretable when one of the two conditions is subject to very little structural change. Moreover, if part of the experiment has to be repeated or extended, the availability of a stable intermediate makes current and future results comparable. Genomic DNA for bacterial microarray studies might be particularly suited for this purpose, as unlike RNA it is not subject to any expression changes though certainly subject to noise. The gDNA samples can be incorporated into the loop design just like any other condition.

5.4 Modelling the dye effect

As mentioned in Section 3.1, the data from the two biological systems have been normalized before being processed further. In principle, no dye effect is present in the data we used. For non-normalized data, it might be of interest to model the dye effect explicitly,

$$\begin{aligned} y_{jk} &= \theta_j - \theta_k + \delta + \epsilon, \\ &= \mu_{jk} + \delta + \epsilon \end{aligned}$$

where δ is the gene-specific dye effect. The advantage of doing it this way is that dye normalization can be done in the same framework we have presented in this paper by merely adding a column of ones to the design matrix X.

6 CONCLUSION

In this paper, we have performed a comparative study between the two commonly used designs of two-channel microarray experiments, the loop and the reference design. We have shown that the loop design is more efficient than the reference design, based on two pilot studies on two very different organisms (human B-cell lymphoma cell line *coelicolor* bacterium) where both designs were considered.

Comparisons between the designs are based on the average estimated variance of the differential expression estimates. This empirical criterion for the comparison of the two designs is related to A-optimality. In both studies the loop design resulted in a smaller average standard error. As a consequence, more genes were detected as differentially expressed by the loop design in the *S.coelicor* study than by the reference design.

These conclusions were supported by a simulation study, where we simulated gene-expression data using a reference and a loop design under the assumption of a known number of differentially expressed genes. Again, the loop design proved superior to the reference design by detecting a greater number of truly differentially expressed genes, whilst reducing the number of false detections. This confirms the assertion that

by using a loop design one can get more precise answers to the biological questions of interest.

Within this comparative study, a simple linear model was proposed to extract the information from any microarray design. From this model, we obtain estimates for all the contrasts. This will make it possible for non-experts to use and interpret loop designs in practice. Further practical recommendations were given on how the simple loop design can be extended to more realistic designs for the case of large experiments, how a dye effect can be accommodated in such designs, as well as on how to decide whether or not channel data can be analysed without transforming them to log-ratios.

ACKNOWLEDGEMENTS

We dedicate the paper to the memory of Orlando de Jesus. We would like to thank Ben Routley for making the data and R code available on the Web and to all the other members of the consortium for stimulating discussions on this work. We would also like to thank the two anonymous reviewers for their constructive comments, which improved the paper greatly. E.W. would like to thank the Dipartimento di Scienze Statistiche 'Paolo Fortunati' of the University of Bologna for its hospitality from January until July 2004. The work on this project was funded by the BBSRC/EPSRC consortium grant 'DNA microarray data analysis and modelling: an integrated approach.'

REFERENCES

Chang,H., Sneddon,J., Alizadeh,A., Sood,R., West,R., Montgomery,K., Chi,J., Mv,M.R., Botstein,D. and Brown,P. (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biol.* **2**, 206–214.

Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.*, **32**, 490–495.

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

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

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Khanin,R. and Wit,E.C. (2004) Design of large time-course microarray experiments with two-channels. *Technical Report 04-6*. Department of Statistics, University of Glasgow, Glasgow, UK.

Landgrebe,J., Bretz,F. and Brunner,E. (2004) Efficient design and analysis of two colour factorial microarray experiments. *Computational Statistics and Data Analysis* (in press).

Lapointe,J., Li,C., Higgins,J., van de Rijn,M., Bair,E., Montgomery,K., Ferrari,M., Egevad,L., Rayford,W., Bergerheim,U., Ekman,P., DeMarzo,A. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.

Pathan,N., Hemingway,C., Alizadeh,A., Stephens,A., Boldrick,J., Oragui,E., McCabe,C., Welch,S., Whitney,A., O'Gara,P., Nadel,S., Relman,P. *et al.* (2004) Role of interleukin 6 in myocardial dysfunction of meningococcal septic shock. *Lancet*, **363**, 203–209.

Sammon,J.W. (1969) A non-linear mapping for data structure analysis. *IEEE Trans. Comput.*, **18**, 401–409.

Townsend,J.P., Cavalieri,D. and Hartl,D.L. (2003) Population genetic variation in genome-wide gene expression. *Mol. Biol. Evol.*, **20**, 955–963.

Townsend,J.P. and Hartl,D.L. (2002) Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol.*, **3**, research0071.1–0071.16.

Wit,E.C. and McClure,J.D. (2004) *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley & Sons, Chichester; Hoboken, NJ.

Wit,E.C., Nobile,A. and Khanin,R. (2004) Simulated annealing for near-optimal dual-channel microarray designs. *Technical Report 04-7*, Department of Statistics, University of Glasgow, Glasgow, UK.

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