

Design and Analysis of Two-color Microarray Experiments Using Linear Models

F. Bretz¹, J. Landgrebe², E. Brunner³

¹B&SR, Novartis Pharma AG, Basel, Switzerland

²Abteilung Biochemie II, Universität Göttingen, Göttingen, Germany

³Abteilung Medizinische Statistik, Universität Göttingen, Göttingen, Germany

Summary

Objectives: A variety of linear models have recently been proposed for the design and analysis of microarray experiments. This article gives an introduction to the most common models and describes their respective characteristics.

Methods: We focus on the application of linear models to logarithmized and normalized microarray data from two-color arrays. Linear models can be applied at different stages of evaluating microarray experiments, such as experimental design, background correction, normalization and hypothesis testing. Both one-stage and two-stage linear models including technical and possibly biological replicates are described. Issues related to selecting robust and efficient microarray designs are also discussed.

Results: Linear models provide flexible and powerful tools, which are easily implemented and interpreted. The methods are illustrated with an experiment performed in our laboratory, which demonstrates the value of using linear models for the evaluation of current microarray experiments.

Conclusions: Linear models provide a flexible approach to properly account for variability, both across and within genes. This allows the experimenter to adequately model the sources of variability, which are assumed to be of major influence on the final measurements. In addition, design considerations essential for any well-planned microarray experiments are best incorporated using linear models. Results from such experimental design investigations show that the widely used common reference design is often substantially less efficient than alternative designs and its use is therefore not recommended.

Keywords

ANOVA techniques, hypothesis testing, normalization, experimental design

Methods Inf Med 2005; 44: 423–30

1. Introduction

Microarray experiments may efficiently be designed and analyzed using linear models, in particular with analysis of variance (ANOVA) techniques. The application of such well-established techniques has several advantages. First of all, linear models provide a flexible approach to properly account for variability both across and within genes. This allows the experimenter to adequately model the sources of variability that are assumed to be of major influence on the final measurements. Linear models accommodate a wide range of experimental designs, including time course experiments and inter-treatment comparisons in factorial experiments. This flexibility allows the application of linear models at different stages of the statistical evaluation of microarray experiments, e.g., in the experimental design phase, for background correction, during the normalization process of the raw data as well as the final hypothesis testing stage. Second, well-known results from linear and nonlinear model theory can be applied to both the design and analysis of microarray data. A large background of experience and theoretical results are available so that the microarray experiment can be tailored to the experimenter's specific requests. Third, linear models are easily implemented and interpreted, once a suitable statistical model has been established. Standard software packages are available, which can be adapted to the requests and circumstances of a specific microarray experiment in many ways.

This article focuses on the description of some applications of linear models in microarray experiments. Large parts of this article are devoted to experimental designs

issues, since we feel this topic to be of particular importance in view of good statistical practice. This article is not intended to serve as a comprehensive review of existing methods. Instead, it is our aim to give the reader an introduction to various applications and we refer to the respective literature for details and further results.

The article is organized as follows. In Section 2 we review some common one- and two-stage linear models for the normalization and analysis of microarray data. In Section 3 we discuss experimental design issues related to the selection of robust and efficient designs, which lead to parameter estimates of interest with minimum variance. An illustrative example is discussed in Section 4. Concluding remarks are given in Section 5.

2. Linear Models

One common approach to evaluate the gene expression in a microarray experiment is to set up a statistical model to describe the measurement responses of a gene expression experiment. A main objective of such an approach is to estimate the relevant effects of interest after adjusting for potential systematic factorial effects. The most widely used statistical models for inter-treatment microarray analysis assume the systematic variations to arise from dye, array, treatment and gene effects, which thus need to be appropriately modeled and included in any statistical consideration. For a general introduction to the statistical analysis of microarrays we refer to the books of Speed [1], Parmigiani et al. [2] and Simon et al. [3].

Since interest in microarray experiments is often focused on fold changes in expression level, the data are typically transformed to the log-scale to allow the use of a linear (instead of a multiplicative) model. In addition, it has been observed [4] that the technical errors tend to be multiplicative (at least for signals with higher intensities). Thus it is common to model the logarithmized observed measurements Y_{ijkl} on array $i = 1, \dots, M$, for dye $j = 1, 2$ under treatment $k = 1, \dots, K$, for gene $l = 1, \dots, G$.

Kerr, Martin and Churchill [5] were among the first to apply linear models to microarray data. They proposed the fixed-effects model

$$Y_{ijkl} = \mu + \alpha_i + \delta_j + \tau_k + \gamma_l + (\alpha\gamma)_{il} + (\tau\gamma)_{kl} + \varepsilon_{ijkl}, \quad (1)$$

where μ denotes the overall mean across all factors, i.e., arrays, dyes, treatments, and genes, α_i , β_j , τ_k , and γ_l denote the overall effects of array i , dye j , treatment (or condition) k , and gene l , while $(\alpha\gamma)_{il}$ and $(\tau\gamma)_{kl}$ denote the array by gene and treatment by gene interactions, respectively. The error terms ε_{ijkl} are assumed to be identical and independent having normal distribution. The interaction term $(\alpha\gamma)_{il}$ accounts for potential spot effects while $(\tau\gamma)_{kl}$ are the effects of main interest in model (1), since they describe experimental responses associated with the specific combination of treatment k and gene l . Note that model (1) is a global model in the sense that it applies to all genes simultaneously. Other global models are also discussed in the literature [5, 6]. Here, in this introductory overview article, we restrict ourselves to the simpler model (1). Single-gene models will be introduced further below. Note also that model (1) assumes no replication of spots on the arrays, although such an experiment would easily be modeled by adding a further index.

Lee et al. [7] described a fixed-effects linear model, which provides an alternative interpretation of model (1). Since model (1) involves tens of thousands of genes, the parameter estimates are not reliable. This is why Lee et al. [7] proposed to split the global linear model (i) into a normalization model describing dye and array effects and (ii) a genewise linear model describing the

remaining effects. More specifically, the measurements are initially modeled as

$$Y_{ijkl} = \mu + \alpha_i + \delta_j + \tau_k + W_{ijkl}. \quad (2)$$

The residual estimates \hat{W}_{ijkl} are subsequently modeled (separately for each gene) as

$$\hat{W}_{ijkl} = \gamma_l + (\alpha\gamma)_{il} + (\tau\gamma)_{kl} + \varepsilon_{ijkl}. \quad (3)$$

The two-stage model is statistically equivalent to the global model under balanced conditions [6] but the parameters can be estimated with higher precision. As seen from equations (2) and (3), the two-stage model has the effect of calibrating the dye effects on a common scale across the arrays (centering of the gene expressions) while adjusting for global effects other than treatment effects (normalization of the gene expressions). Note that the output of two-channel normalization methods (see Itrich [8] for a recent review) may also be used as a response for the second stage instead of the residuals estimates \hat{W}_{ijkl} from model (2).

The interaction effects $(\tau\gamma)_{kl}$ can be used in either model (1) or models (2) and (3) to obtain the final parameters of interest, such as pairwise comparisons, factorial main or interaction effects. Let, for example, Z_{il} denote the difference of the expressions for gene l on array i under treatment k (dye $j = 1$) and k' (dye $j = 2$). The log-differences

$$Z_{il} = \hat{W}_{i1kl} - \hat{W}_{i2kl} = (\tau\gamma)_{kl} - (\tau\gamma)_{k'l} + \varepsilon_{i1kl} - \varepsilon_{i2kl}$$

can then be used for the final inter-treatment analysis. Least square estimates for $(\tau\gamma)_{kl}$ and corresponding variance estimates can be found, for example, in Kerr and Churchill [9].

Wolfinger et al. [10] considered a model similar to Lee's et al. [7] two-stage model but assumed the array effect to be random. In the following, capital Latin letters are used to denote random effects (distinguishing them from fixed effects). In the model of Wolfinger et al. [10] the first stage again consists of normalizing the observed signals to account for global systematic effects,

$$Y_{ijkl} = \mu + A_i + \tau_k + (A\tau)_{ik} + W_{ijkl}. \quad (4)$$

In contrast to Kerr et al. [5] and Lee et al. [7], the dye effect is not included. At the second stage, the estimated residuals are then used as input for the single-gene model

$$\hat{W}_{ijkl} = \gamma_l + (A\gamma)_{il} + (\tau\gamma)_{kl} + \varepsilon_{ijkl}. \quad (5)$$

With this approach, the array effects A_i , and thus the interaction terms $(A\tau)_{ik}$ and $(A\gamma)_{il}$, are assumed to be random (independently normal distributed, to be more specific) since the arrays are not of particular interest in microarray experiments and are randomly selected from a large set of available arrays.

Similarly, Landgrebe et al. [11] used a two-step genewise model. The data were normalized with the method of Yang et al. [12] (step 1) and the residuals from the non-linear regression were modeled as (step 2):

$$\hat{W}_{ijk} = \mu + A_i + \delta_j + \tau_k + \varepsilon_{ijk}. \quad (6)$$

Note that the residuals \hat{W}_{ijk} are obtained from residual log-ratios as described in Landgrebe et al. [13]. The gene index is dropped here, leading to main effects for dye (δ_j) and treatment effects (τ_k) which are interaction effects if the gene index is retained as in model (1). The dye effect (or dye by gene effect) is included in the second step because global normalization approaches can fail to fully eliminate the dye bias [14].

The set of models (1)-(6) reflects the flexibility of linear models to accommodate the relevant sources of variability. Numerous modifications of the above standard models could be considered, such as – but not restricted to – the inclusion of replicated spots as indicated above, the modeling of time course experiments [15], possibly describing the dependency patterns of repeated measurements using standard covariance structures, or the inclusion of spatial information in a suitable linear model to perform a background correction of the raw signals [16]. Another example of applying linear models is to include biological variability in the statistical model and the final analysis [13, 14, 17-19]. None of the models (1)-(6) include biological replicates (i.e., subjects) as a random factor. If multiple subjects are available, they can be considered as

a random subset from a much larger population, in which case the biological variability can fully be taken into account in the design and the analysis of microarray experiments. We refer to the original articles for more details.

3. Experimental Design

An important issue is the selection of efficient designs. Two-color microarrays offer the possibility of labeling the RNA targets of interest with two different dyes. That is, with more than two treatments, the question arises regarding how to label the treatments and allocate them across the arrays such that the parameters of interest can be estimated with minimum variance.

To compare different treatments with each other, RNA from the treatments can be processed in at least two ways. One method is to label the treatment RNA with a first dye (e.g., green) and hybridize it to different arrays using a co-hybridization to each array with a common reference target labeled with the second dye (e.g., red) (indirect comparison design). A second method is to compare the different treatments with each other using either a swap design or a loop design or a combination of both [9]. In such designs, two treatments per array are labeled alternately with the first and the second dye. In swap designs, two microarrays per non-replicated experimental unit are used: the RNA from the first treatment is labeled with the first (e.g., green) dye and the RNA from the second treatment is labeled with the second (e.g., red) dye. Both are mixed and hybridized to the first microarray. For the second array, this is done vice-versa (first RNA with second dye, second RNA with first dye). In loop designs, the first, green-labeled RNA is co-hybridized with the second, red-labeled RNA to the first array; the second, green-labeled RNA is co-hybridized with the red-labeled third RNA to the second array etc. until the loop is closed by co-hybridizing the last RNA labeled in green with the first RNA labeled in red to the last array. See Figure 1 for examples of a common reference design and a loop design with three treatments each; for

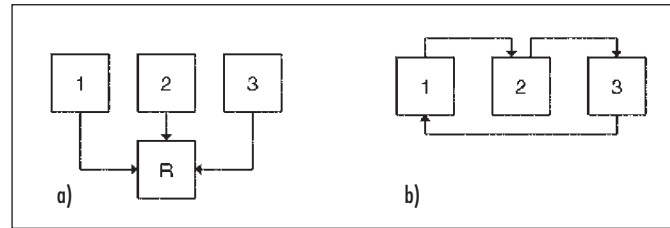


Fig. 1 Common reference (a) and loop (b) designs for comparing three treatments. The arrows indicate the arrays used for the hybridization of connected treatments. Arrowheads = labeling with first dye; arrowbase = labeling with second dye

a detailed overview, see Churchill [20] and Speed [1].

Swap and loop designs can be termed as direct comparison designs. In statistical terminology, both this type of design and the common reference design are incomplete two-factor block designs of size two. The incomplete two-factor block design situation is characterized by two factors with two levels each: the first factor is the dye effect, as depicted by Kerr et al. [5] (two levels, green and red), describing the observation that the dyes differ in brightness and labeling efficiency. The second factor is the effect of interest (treatment effect), which can have various levels. The block is the spot on one microarray to which two RNAs from different treatments are hybridized using two different dyes. The design is said to be incomplete because the block size of two is insufficient for direct comparisons if more than two treatments are to be analyzed. Moreover, the dye effect and the treatment effect are confounded on the level of the arrays.

Two measures are typically taken in all microarray experiments to handle the incompleteness. i) The dyes and the targets are assigned in the hybridization in such a way that the dye effect can be mathematically eliminated and does not affect the final parameter estimates of interest. ii) The allocation of the treatments across the arrays should be done such that all comparisons of interest can be made using a minimum number of slides while also minimizing the error variance of the comparisons. These measures ensure that the variance of the hybridization reaction, the microarray manufacturing, and the interaction of both can be eliminated by usage of the blocking principle. Note that by construction this is not

possible with one-color arrays like commercial in-situ-photolithography oligoarrays or nylon-filter macroarrays. For these types of microarrays, the above mentioned interactions cannot be statistically eliminated and contribute to the overall measurement errors.

Recent publications dealing with the problem of experimental design in microarrays pursue the following questions: How can one define the best design for a given experimental situation with a defined number of treatments and contrasts of interest, i.e. how can different possible designs be compared and evaluated? Which constraints affect the design choice? How many technical and/or biological replicates are needed?

To answer these questions, Landgrebe et al. [11] applied the linear model to the residuals of an initial normalization [8, 12, 21-24]. Because experimenters often want to address very specific questions with microarray experiments, they proposed the use of special contrast vectors or matrices that describe the biological questions. Landgrebe et al. [11] then optimized the design matrix taking into account (i) the estimability of the contrasts and (ii) technical constraints like the availability of material and financial resources to obtain the most efficient designs for the contrasts of interest by minimizing the variances of the treatments' effects estimators. Furthermore, they provide conditions for the estimability of contrasts and a thorough definition of efficiency that goes beyond the notion of admissibility introduced by Glonek and Solomon [15]. This definition allows the comparison of two designs and the assessment of how many microarrays are required for a particular experimental design to achieve

the same efficiency as a second design potentially involving more slides.

The results can be applied to search for the most efficient designs given the experimental questions. In case of single contrasts, variance factors are used to compare different designs. For factorial designs, standard optimal design theory is applied to quadratic forms. Similar ideas, though not developed to full extent, have been proposed by Kerr et al. [5], Kerr and Churchill [9] and Glonek and Solomon [15]. In the following we give a concise description of methods to select efficient designs. At the end of this section further issues related to designs including biological replicates and robustness properties of efficient designs are briefly discussed.

Based on model (6) we consider the observed differences separately for each gene

$$Z_i = Y_{i1k} - Y_{i2k} = \delta_1 - \delta_2 + \tau_k - \tau_{k'} + \phi_i, \quad (7)$$

where $\phi_i = \varepsilon_{i1k} - \varepsilon_{i2k}$. Note that for the remaining section the gene index i is dropped. Let $\mathbf{Z}' = (Z_1, \dots, Z_N)$ denote the associated vector of differences across the N arrays for a single gene. Let $\boldsymbol{\beta}' = (\delta_1, \delta_2, \tau_1, \dots, \tau_K)$ denote the parameter vector containing the fixed effects, where for convenience τ_K denotes the mean of the reference group. The relevant design information of the current microarray experiment is contained in a $N \times (K+2)$ design matrix \mathbf{X} . Assume, for example, a common reference experiment with $K-1$ treatments. Then the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & \cdots & 0 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & \cdots & 0 & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & -1 & 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}$$

describes the arrangement of the effects on the slides and thus serves as a protocol for the experimenter. The first two columns in \mathbf{X} stand for the dye effects δ_i , $i = 1, 2$, the following $K-1$ columns stand for the treatment effects $\tau_1, \dots, \tau_{K-1}$, while the last column stands for the effect τ_K of the common reference. We can therefore reformulate model (7) as

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi}, \quad (8)$$

where $\boldsymbol{\phi}' = (\phi_1, \dots, \phi_N) \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ and \mathbf{I}_N denotes the identity matrix of order N . The subsequent methods for selecting efficient experimental designs are distribution-free and thus do not rely on specific assumptions about the distribution of the residuals. The normality assumption for $\boldsymbol{\phi}$ is only used when deriving the test statistics for the data analysis. In addition, we do not assume variance homogeneity over the genes. Instead, we use a gene-wise error model where variance homogeneity is assumed over the arrays. This is reasonable since the arrays are standardized after normalization. This assumption also applies to the single gene models in Section 2.

In order to perform selected comparisons among the K treatments, we define contrast vectors \mathbf{c} of length $K+2$. Contrast vectors have the property that their elements sum to 0, i.e., $\mathbf{c}'\mathbf{1} = 0$. A contrast vector then reflects a single biological comparison of interest if we consider the linear combination $\mathbf{c}'\boldsymbol{\beta}$. For example, to compare treatment 1 with treatment 2, the contrast vector $\mathbf{c}' = (0, 0, 1, -1, 0, \dots, 0)$ is used to get the desired difference by multiplying the parameter vector $\boldsymbol{\beta}$ from left with \mathbf{c}' , so that we obtain $\mathbf{c}'\boldsymbol{\beta} = \tau_1 - \tau_2$. In the following, we refer to $\mathbf{c}'\boldsymbol{\beta}$ as the (linear) function of interest. If we are interested, for example, in testing for a possible dye effect, we set $\mathbf{c}' = (1, -1, 0, 0, 0, \dots, 0)$, which yields $\mathbf{c}'\boldsymbol{\beta} = \delta_1 - \delta_2$. In general we will be faced with a given set of experimental questions described by a set of contrast vectors, say $\mathbf{c}_1, \dots, \mathbf{c}_p$. Examples of such sets of experimental hypotheses are discussed later.

An important condition in linear model theory is the estimability of contrasts, since otherwise the linear function of interest $\mathbf{c}'\boldsymbol{\beta}$ is biased by some unknown systematic effect. A function $\mathbf{c}'\boldsymbol{\beta}$ is called estimable, if there exists a vector \mathbf{t} of length N such that $E(\mathbf{t}'\mathbf{Z}) = \mathbf{c}'\boldsymbol{\beta}$ [25, p. 114]. Assume, for example, that we compare three treatments in the loop design (see Fig. 1)

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & -1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 \end{pmatrix},$$

where for technical reasons the elements in the last column are set to 0, since no com-

mon reference group is included. Suppose further that we are interested in the difference $\tau_1 - \tau_2$. It is easily seen that $E(\mathbf{t}'\mathbf{Z}) = E((2, -1, -1, \dots)\mathbf{Z}) = \tau_1 - \tau_2$ for any \mathbf{Z} following model (8). We thus obtain the appropriate weights $\mathbf{t}' = (2, -1, -1)$ for the linear combination $2Z_1 - Z_2 - Z_3$ to unbiasedly estimate $\mathbf{c}'\boldsymbol{\beta}$. Necessary and sufficient estimability conditions can be found in standard text books such as Searle [26].

It can be shown that if $\mathbf{c}'\boldsymbol{\beta}$ is estimable, $\mathbf{c}'\hat{\boldsymbol{\beta}} = \mathbf{t}'\mathbf{Z}$ is the best linear unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$ with variance $\text{Var}(\mathbf{t}'\mathbf{Z}) = \sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^- \mathbf{c}$, where $(\mathbf{X}'\mathbf{X})^-$ denotes some generalized inverse of $\mathbf{X}'\mathbf{X}$ [27]. For a given (unknown) technical error σ^2 and a contrast vector \mathbf{c} the variance $\text{Var}(\mathbf{t}'\mathbf{Z})$ depends only on the design matrix \mathbf{X} of the microarray experiment. Since it is desirable to obtain the most precise estimates for $\mathbf{c}'\boldsymbol{\beta}$, we select the design \mathbf{X}^* , which minimizes the variance $\text{Var}(\mathbf{t}'\mathbf{Z})$ for a given experimental question \mathbf{c} . The design \mathbf{X}^* is then called efficient design. It can be either determined from a pre-selected set of candidate designs (such as specific loop or swap designs) or from the entire space of possible designs with N arrays. Note that the ratio of variance factors for two designs \mathbf{X}_1 and \mathbf{X}_2 to be compared with each other, expresses the relative number of arrays required by the design \mathbf{X}_1 in order to achieve the same efficiency as in design \mathbf{X}_2 .

The ideas remain similar if more than one experimental question is of interest. In the example in Section 4, six conditions were included in the microarray experiment and the goal was to compare each condition with the average across all conditions. Thus, $p = 6$ experimental questions are of interest, which are summarized by the contrast matrix

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 5 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & -1 & 5 & -1 & -1 & -1 & -1 \\ 0 & 0 & -1 & -1 & 5 & -1 & -1 & -1 \\ 0 & 0 & -1 & -1 & -1 & 5 & -1 & -1 \\ 0 & 0 & -1 & -1 & -1 & -1 & 5 & -1 \\ 0 & 0 & -1 & -1 & -1 & -1 & -1 & 5 \end{pmatrix}.$$

Once such a contrast matrix \mathbf{C} has been established, the covariance matrix of the estimates is given by $\text{Cov}(\mathbf{T}'\mathbf{Z}) = \sigma^2 \mathbf{C}'(\mathbf{X}'\mathbf{X})^- \mathbf{C}$, where \mathbf{T} is a matrix contain-

ing the vectors \mathbf{t}_i associated with the i -th experimental question \mathbf{c}_i . Note that $\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}$ is now a matrix, in contrast to the scalar $\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$ obtained for a single contrast \mathbf{c} , and results from optimum design theory have to be used to solve the related minimization problem. We refer to Landgrebe et al. [11] and Pukelsheim [28] for further details related to this minimization problem.

In order to illustrate the previous methods, we briefly discuss their applications to 2×2 designs, which occur, for example, when comparing a mutant with the wildtype for two cell lines. Further discussions and other examples involving more complicated factorial treatment structures can be found in Landgrebe et al. [11].

Different designs can be compared with each other by calculating the variance factors for each single contrast and design; see Figure 2 for some sample designs and related abbreviations. In Table 1, variance factors of some common designs are shown for some relevant experimental questions (main A , main B and interaction effects). The table illustrates how the choice of an efficient design depends on the selected contrasts of interest. Main effects for one factor are estimated by averaging over the levels of the second factor. The CL design allows for the estimability of all contrasts. But the interaction effect $A \times B$ is the only comparison yielding a competitive variance factor. The other contrasts cannot be estimated efficiently in comparison with the competing designs. If only the main B effects and the interaction $A \times B$ are of interest, the BS design is clearly most efficient. However, if the experimenter changes his mind and becomes interested in the main A effect after completion of the experiment, this contrast cannot be estimated and the comparisons remain open. An appealing approach is to combine two or more basic designs from Figure 2. Consider, for example, the 2CL2XS design, where 2 CL designs are combined with 2 XS designs. This combined design is more efficient than the pure CL or XL designs when considering all three effects simultaneously, since it, on average, has lower variance factors. Other combinations of basic designs are possible and the selection of the final design for

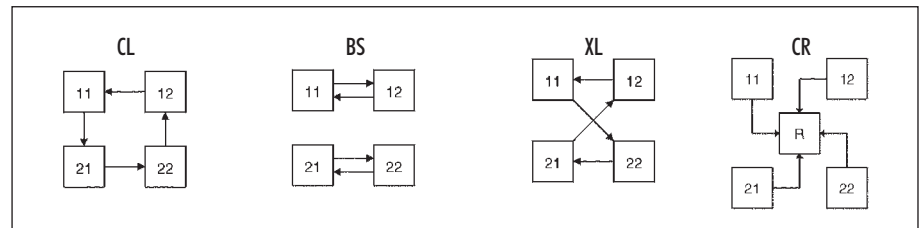


Fig. 2 Some examples of 2×2 designs. CL: circle loop; BS: swap over B; XL: cross loop; CR: common reference

the specific microarray experiment also depends on the relative importance of the experimental questions: if, for example, the comparison of a mutant with the wildtype is more important than the comparison of the two cell lines, a design emphasizing this comparison should be selected and vice versa. A general result from Table 1 is, however, that the CR design is substantially less efficient than the competitors.

So far we only considered the efficiency of experimental designs. A topic, which has not yet been covered in the literature, though being of equal importance, is the robustness of different microarray designs with respect to missing values. As long as the principles of dye balance and replication are applied, the common reference design seems to be robust to a loss of spots or arrays resulting from poor quality hybridization (or any other reason). Intuitively one could think that with more complex designs, the efficiency advantage of loop designs by creating multiple links among the samples is jeopardized by a potential non-robustness when one or more arrays are discarded. Since arbitrary pairs of samples can be contrasted only indirectly through chains of comparisons, already a loss of few arrays might cause a breakdown in the chain. This topic is currently under investigation by our group and preliminary results suggest that

the drop in efficiency can be substantial for some designs. However, when the experiment is properly designed, the resulting loops are efficient as well as robust and should be preferred to the common reference design.

Biological replicates can be incorporated in the setting of linear models as well. Initial results on experimental designs taking the biological variability into account are also available [13, 14, 18, 19]. Landgrebe et al. [13], for example, compared several two-sample designs with respect to the resultant degrees of freedom associated with the estimates of the biological and technical error terms. The authors give advice for experimental situations with differing group sizes and show the impact of different designs on the variance and degrees of freedom of the resultant test statistics. Similar results for one-factorial layouts comparing a treatments ($1 \times a$ designs) were obtained by [18, 19]. It should be noted that the final selection of the experimental design depends in particular on the magnitude of the biological variability. Based on our experience with hundreds of microarray experiments performed at our central eukaryote microarray facility, we noticed that biological variability strongly varies in a gene- and organism-dependent manner. Simple model organisms like yeast cells show little biological

Table 1
Variance factors for some examples of 2×2 designs

Effect	Design				
	4CR	4CL	4XL	2CL2XL	4BS
A	1	0.5	0.25	0.3333	n.e.
B	1	0.5	0.5	0.5	0.25
AB	1	0.25	0.5	0.3333	0.25

$n \times m Y = n$ times design X combined with m times design Y with X and Y taken from Figure 2; n.e. = not estimable

Table 2 Contrast vectors for the six-tissue experiment. *g*: labeling with cy3, *r*: labeling with cy5, *ta-ff* tissues 1 to 6

array	<i>g</i>	<i>r</i>	<i>ta</i>	<i>tb</i>	<i>tc</i>	<i>td</i>	<i>te</i>	<i>ff</i>
1	1	-1	1	-1	0	0	0	0
2	1	-1	1	0	-1	0	0	0
3	1	-1	0	1	-1	0	0	0
4	1	-1	0	1	0	-1	0	0
5	1	-1	0	0	1	-1	0	0
6	1	-1	0	0	1	0	-1	0
7	1	-1	0	0	0	1	-1	0
8	1	-1	0	0	0	1	0	-1
9	1	-1	0	0	0	0	1	-1
10	1	-1	-1	0	0	0	1	0
11	1	-1	-1	0	0	0	0	1
12	1	-1	0	-1	0	0	0	1

variability for the majority of genes, inbred mice show some more, whereas patient material shows high biological variability. For microarray experiments with patient material, the common reference design can be (nearly) as efficient as direct comparison designs as the advantage of the higher precision of the technical variance estimation diminishes.

4. Application to Experimental Data

In this section, we will exemplify the considerations from the two previous sections by looking at the design and analysis of a microarray experiment that was performed in our laboratory.

The goal of this experiment was to compare the expression profile of a defined set of about 150 genes in six different tissues from a rodent species. The aim was to obtain an “expression map” for the genes in the tissues under investigation. One could think of solving this problem with multiple pairwise tissue comparisons, but this approach would lead to a poorly interpretable output and aggravate the testing problems that are part of every microarray study [29, 30]. Instead, we

solved the problem of obtaining an expression map by comparing the expression of a certain gene in a given tissue type to the average expression of all remaining tissues leading to six one-against-all comparisons in total. The material constraint was to use a maximum of 12 microarrays. The design is a one way layout with six levels (1×6). As shown in Section 3, $1 \times a$ -layouts can be performed as loop-designs, which are more ef-

ficient than common reference designs. In this case, we used the nested loop design given in Table 2.

Using six arrays (the odd array numbers), each tissue was directly compared to the next tissue in a first loop, and in addition, there was a second six-array loop “skipping” every second tissue (even array numbers). Using these two nested loops, a design with improved robustness against missing observations and an efficient estimability of the technical error due to the inherent technical replication was obtained: in comparison to a common reference design with 12 arrays (two replicates per tissue), the efficiency is 2.31 times higher, i.e., to find the same number of differentially expressed genes the common reference design would require twice as many arrays [11].

The experiment was performed using cDNA-two-color microarrays manufactured in our laboratory. The arrays were spotted with 2000 presumably non-differentially expressed genes for the purpose of normalization and with the approx. 150 genes of interest. RNA was isolated from the six tissues and labeled and hybridized to 12 arrays according to the design shown above.

The data were log-transformed and normalized using the array/pinwise nonlinear regression normalization proposed by Yang et al. [12], the resulting normalized log-

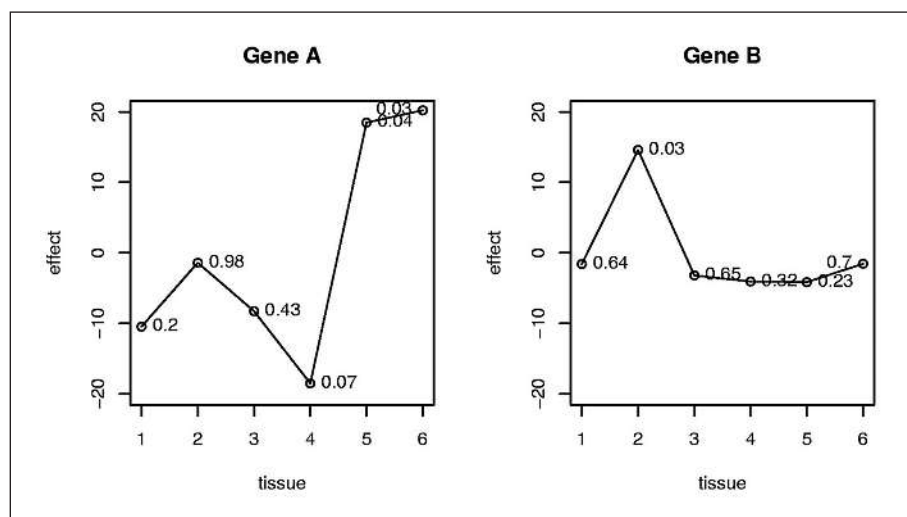


Fig. 3 Plot of the contrast effects for two genes (A and B). Abscissa: tissue, ordinate: effect tissue vs. the average of all other tissues (based on normalized gene expression log ratios). The numbers beside the plot dots indicate the adjusted p-values following Benjamini and Hochberg [31].

ratio residuals were modeled as described in (7).

The six T -statistics $T_i = \frac{\frac{5}{6}Z_i - \frac{1}{6}\sum_{j \neq i} Z_j}{MSE}$

(where MSE is the variance estimate) reflecting the experimental questions were expressed by the contrast matrix given in equation (9). These contrasts compare the expression of a gene in a given tissue against its average expression in the other tissues. As an example for the resulting analysis, Figure 3 shows plots of the six-contrast effects for two genes with the corresponding adjusted p -values according to Benjamini and Hochberg [31]. The plots show that gene A is strongly down-regulated in tissue 4 (compared to the remaining tissues) and upregulated in tissues 5 and 6. Gene B is strongly upregulated in tissue 2, while both genes do not show significant differential expression in the other tissues, respectively. This data presentation allows a straightforward biological interpretation and makes the planning of follow-up experiments easy. An analysis based on multiple pairwise tissue comparisons would have produced a hardly interpretable output instead of gene-wise tissue expression maps generated by the approach described here. Thus, this example demonstrates the value of using linear models for the design and analysis of a typical microarray experiment.

5. Conclusions

Linear models offer a flexible and powerful approach for the design and analysis of microarray experiments. Flexibility is provided at different levels for the statistical evaluation of microarray data: (i) potential systematic sources of variability are easily included or deleted according to the specific circumstances; (ii) linear models have been successfully applied at such different stages of a microarray experiment as experimental design, background correction, normalization and hypothesis testing; (iii) experimental questions of interest are easily translated to contrast statements and can therefore be answered in an efficient way.

A particular strength of linear models is the possibility to design a microarray experiment such that a minimum number of arrays are spent to obtain a maximum of information. Initial experimental design considerations prior to the actual data collection are crucial for a well-planned microarray experiment. Whether or not to apply a particular design thus also depends on the final objectives of the experiment. To our experience, experimental questions can often be incorporated into the linear model framework. As a particular result, it is seen that on many occasions the common reference design is substantially less efficient than alternative designs. Using variance factors as a measurement for efficiency shows that, for example, the common reference design requires three times more arrays in a 1×3 experiment than a loop design to achieve the same chance of finding differentially expressed genes if only technical replicates are used in the experiment and if inter-treatment comparisons are of main interest. Note, however, that for class discovery or class prediction objectives, common reference designs may be better suited than indirect comparison designs [18]. In any case, we pointed out in this article that besides the efficiency, further relevant issues such as robustness and the inclusion of biological replicates should always be taken into account when planning and performing a microarray experiment.

Acknowledgments

We thank Drs. L. Walter, R. Dressel and E. Günther (†) for providing the biological problem and material for the experiment described in Section 4. We thank Christiane Lach and Katharina Nau for their technical assistance.

References

- Speed T. *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton: CRC; 2003.
- Parmigiani G, Garrett ES, Irizarry RA, Zeger SL. *The Analysis of Gene Expression Data*. New York: Springer; 2003.
- Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and Analysis of DNA Microarray Investigations*. New York: Springer; 2003.
- Roche DM, Durbin B. A Model for Measurement Error for Gene Expression Arrays. *Journal of Computational Biology* 2001; 8 (6): 557-69.

- Kerr M, Martin M, Churchill G. Analysis of variance in microarray data. *Journal of Computational Biology* 2000; 7: 819-37.
- Kerr MK. Linear models for microarray data analysis: Hidden similarities and differences. *Journal of Computational Biology* 2003; 10 (6): 891-901.
- Lee MLT, Lu W, Whitmore GA, Beier D. Models for microarray gene expression data. *Journal of Biopharmaceutical Statistics* 2002; 12 (1): 1-19.
- Ittrich C. Normalization for Two-channel Microarray Data. *Methods Inf Med* 2005; 44: 418-22 (this issue).
- Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostatistics* 2001; 2: 183-201.
- Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 2001; 8 (6): 625-37.
- Landgrebe J, Bretz F, Brunner E. Efficient design and analysis of factorial two color microarray Experiments. *Computational and Statistical Data Analysis* 2004 (in press).
- Yang YH, Dudoit S, Luu P, Speed TP. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 2002; 30: e15.
- Landgrebe J, Bretz F, Brunner E. Efficient two-sample designs for microarray experiments with biological replications. *In Silico Biology* 2004; 4 (38). Available at <http://www.bioinfo.de/isb/2004/04/0038/>
- Kerr MK. Design considerations for efficient and effective microarray studies. *Biometrics* 2003b; 59: 822-8.
- Glonck G, Solomon P. Factorial and time course designs for cDNA microarray experiments. *Biostatistics* 2004; 5: 89-112.
- Vass K, Wit E. Background correction through ANOVA. Website <http://www.stats.gla.ac.uk/~microarray/research.html> [last access: August 21, 2004].
- Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* 2001; 29: 389-95.
- Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* 2002; 18: 1438-45.
- Dobbin K, Shih JH, Simon R. Statistical design of reverse dye microarrays. *Bioinformatics*. 2003; 19: 803-10.
- Churchill G. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* 2002; 32: 490-5.
- Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002; 18 (Suppl 1), S96-S104.

22. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 2002; 18 (Suppl 1), S105-S110.
23. Cui X, Kerr K, Churchill G. Data transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology* 2002; 2 (1): Article 4.
24. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* 2003; 19 (8): 966-72.
25. Ravishanker N, Dey DK. *A First Course in Linear Model Theory*. Chapman & Hall / CRC; 2002.
26. Searle S. *Linear Models*. New York: Wiley; 1971.
27. Rao CR, Mitra SK. *Generalized Inverse of Matrices and its Applications*. New York: Wiley; 1971.
28. Pukelsheim F. *Optimal Design of Experiments*. New York: Wiley; 1993.
29. Bretz F, Landgrebe J, Brunner E. Multiplicity issues in microarray experiments. *Methods Inf Med* 2005; 44: 431-7 (this issue).
30. Hommel G, Kropf S. Tests for Differentiation in Gene Expression Using a Data-Driven Order or Weights for Hypotheses. *Biometrical Journal* 2005 (to appear).
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B* 1995; 57: 289-300.

Correspondence to:

Edgar Brunner
Abteilung Medizinische Statistik, Universität Göttingen
Humboldtallee 32
37073 Göttingen
Germany
E-mail: brunner@ams.med.uni-goettingen.de