BIOINFORMATICS

Vol. 20 no. 9 2004, pages 1436–1446
DOI: 10.1093/bioinformatics/bth118

Analysis of variance components in gene expression data

James J. Chen[1,*], Robert R. Delongchamp[1], Chen-An Tsai[1], Huey-miin Hsueh[3], Frank Sistare[4], Karol L. Thompson[4], Varsha G. Desai[2] and James C. Fuscoe[2]

[1]Division of Biometry and Risk Assessment, [2]Center for Functional Genomics, Division of Genetic and Reproductive Toxicology, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079, USA, [3]Department of Statistics, National Chengchi University, Taipei, Taiwan and [4]Division of Applied Pharmacology Research, Center for Drug Evaluation and Research, Food and Drug Administration, Laurel, MD 20708, USA

Received on April 29, 2003; revised on November 21, 2003; accepted on December 8, 2003
Advance Access publication February 12, 2004

ABSTRACT

Motivation: A microarray experiment is a multi-step process, and each step is a potential source of variation. There are two major sources of variation: biological variation and technical variation. This study presents a variance-components approach to investigating animal-to-animal, between-array, within-array and day-to-day variations for two data sets. The first data set involved estimation of technical variances for pooled control and pooled treated RNA samples. The variance components included between-array, and two nested within-array variances: between-section (the upper- and lower-sections of the array are replicates) and within-section (two adjacent spots of the same gene are printed within each section). The second experiment was conducted on four different weeks. Each week there were reference and test samples with a dye-flip replicate in two hybridization days. The variance components included week-to-week, animal-to-animal and between-array and within-array variances.

Results: We applied the linear mixed-effects model to quantify different sources of variation. In the first data set, we found that the between-array variance is greater than the between-section variance, which, in turn, is greater than the within-section variance. In the second data set, for the reference samples, the week-to-week variance is larger than the between-array variance, which, in turn, is slightly larger than the within-array variance. For the test samples, the week-to-week variance has the largest variation. The animal-to-animal variance is slightly larger than the between-array and within-array variances. However, in a gene-by-gene analysis, the animal-to-animal variance is smaller than the between-array variance in four out of five housekeeping genes. In summary, the largest variation observed is the week-to-week effect. Another important source of variability is the animal-to-animal variation. Finally, we describe the use of variance-component estimates to determine optimal numbers of animals, arrays per animal and sections per array in planning microarray experiments.

Contact: jchen@nctr.fda.gov

*To whom correspondence should be addressed.

1 INTRODUCTION

DNA microarray technology provides tools for studying expression levels for thousands of genes in a number of experimental samples (conditions) simultaneously. A microarray experiment is a multi-step process, and each step is a potential source of variation. Variability can be generally classified into three categories: biological variation, technical (process) variation, and residual variance (Novak *et al*., 2002; Churchill, 2002). Biological variation refers to the variation from different RNA sources. It reflects differences in host characteristics. Biological variation is due to inherent differences in gene expression, varying from subject to subject due to genetic or environmental factors. Technical variation refers to the variation arising from the use of the microarray system. Potential sources of technical variation include the sample preparation procedures such as RNA extraction and purification, cDNA synthesis, incorporation extent of dyes and the specific batch of dyes used; the microarray construction procedures such as the amount of probe applied to the slides, spot shape, pin geometry and fixation of the spotted DNA to the slides; the hybridization and washing procedures such as the amount of labeled cDNA applied to the slides and hybridization temperatures; the detection method such as scanner setting parameters; cross-hybridization within gene families; and outshining from neighboring spots. Laboratory environmental conditions, such as room temperature, are another source of variability during the lengthy, multi-step

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

Bioinformatics 20(9) © Oxford University Press 2004; all rights reserved.

process of performing a microarray experiment. Although the relative magnitude of effects of environmental conditions on the total variability of the experiment is generally not known, controlling these conditions to the greatest extent possible will obviously help reduce this source of variability. This variation is generally classified as a time or block effect. Other potential sources of variability such as concentrations of components in reaction and wash buffers can be controlled by using stock solutions and master mixes as much as possible.

A microarray experiment is generally a comparative experiment in which the experiment of interest is the comparison of the relative expression levels among samples rather than the determination of absolute intensity measures of each sample. But gene expression data produce a signal-to-noise ratio that must be assessed objectively. Replication allows for assessment of the variability of expression data such that formal statistical analysis methods can be applied. Without replication, one cannot distinguish between true differences in gene expression and random fluctuations. Replication can be incorporated at different levels of the experiment. For example, replications can be conducted for different tissues or different cell lines, each one can be hybridized to more than one arrays and each array can consist of replicated spots of the same gene. Yang and Speed (2002) described two types of replication: technical replicates and biological replicates. Technical replication refers to replication in which the mRNA is from the same pool (the same extraction). Biological replicates refer to hybridizations that involve mRNA from different extractions. If the purpose of the experiment is to determine the effects of a treatment on different biological populations, then statistical tests should be based on the biological replicate samples. If the purpose is to detect variations within the experimental groups, then tests can be based on technical replicate samples. In general, a researcher will want to use biological replicates to obtain averages of independent data and to validate a generalization of the conclusion and to use technical replicates to assist in reducing experimental variabilities.

Kerr *et al*. (2000) presented an analysis of variance (ANOVA) model for the analysis of microarray data. The ANOVA model is a popular statistical approach for modeling sources of variation. It considers all possible sources of variation in a microarray experiment and summarizes them in one equation. The ANOVA method provides an automatic correction for the nuisance effects in estimating the relative expression of genes across experimental samples. An advantage of the model-based ANOVA approach is that it estimates the magnitude of the sources of variation explicitly. A microarray experiment is conducted in steps; variation in each step potentially affects the measured gene expression intensity levels. Identifying and estimating different sources of variation are essential in designing an efficient experiment. Recently, Cui and Churchill (2003), Spruill *et al*. (2002) and Wernisch (2002) used the ANOVA model to evaluate the sources of variation in microarray data. In this paper, we present a variance-components approach to assessing sources of variation for two gene expression data sets.

2 MATERIALS AND METHODS

2.1 Sources of variation

The variation of the measured gene expression data can be categorized into three generic sources: biological, technical and residual variations. The biological variation in gene expression measured comes from different animals or different cell lines or tissues. It reflects the variability among the different target biological samples used in the experiment. The target biological samples are the experimental units. Different target samples are independent biological replicates to reflect the variability in the population of interest. Biological variation is estimable only when there are independent biological replicates. If all biological samples are pooled, the biological variation is minimized and inference can be made only as to the experimental conditions. The technical variation accounts for the variation associated with the use of microarray techniques unrelated to the biological samples (more details given below). The residual variation accounts for sampling or experimental variation or other unaccountable factors. The biological, technical and residual variations are mutually independent. The variation in a measured intensity is the sum of these three variations.

We distinguish two types of variations: within-array variation and between-array variation. Within-array variation refers to variation originating from array-specific spot effects. The within-array variation can be caused by scratches or dust on the surface of an array or by the printing, washing or image extraction processes. There are also systematic variations, such as differences in labeling efficiency, intensity or spatial dependency biases. For example, every grid in an array is printed using the same print-tip. Systematic differences may exist between the print-tips or printing order, such as slight differences in the length or deformation after many hours of printing, or the biological materials can decay with time. Within-array variation can be assessed by putting replicates at different locations of each array. Note that normalization or transformation (Yang *et al*., 2002; Durbin *et al*., 2002) is often performed to adjust for the systematic biases prior to statistical analysis. Since the target mRNA sample within an array is from the same extraction, within-array variation is attributed to the technical variation.

Between-array variation observed in an experiment can be due to biological factors or technical factors. They can be composed of array batch-to-batch variation (quality and homogeneity in manufacturing the array including gene sequence variance), array-to-array hybridization variance (such as sample preparation on different dates) and biological variation (i.e. target mRNA samples are from different biological samples hybridized to different arrays). In this paper, the between-array variation is restricted to the same target

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

(biological) sample. The target sample can be one individual or a mixture of individuals. Under this restricted definition, the between-array variation is attributable to technical variation. Thus, technical variation consists of two components: within-array and between-array variation. Technical variations can be assessed by technical replicates. Between-array variation can only be assessed by replication of the same RNA sample to more than one array. Depending on the microarray design, the within-array replication can consist of sub-components, illustrated in the first example data set below. The biological, between-array and within-array variations are nested in a hierarchical fashion (between-array variation nested in biological variation and within-array variation nested in between-array variation).

In addition to the biological and technical variations, another random variation often encountered is the time effect. Because of available resources, a large study (an experiment involves many arrays or many biological samples) is often divided into several small experiments. Each experiment is conducted at a different time (day or week). The data from the same experiment are more homogeneous because they are generated under similar experimental conditions. (Each experiment can be considered as a block.) In the hierarchical structure of the variance components, the time effect variance is generally at the highest level. The second example data set provides an illustration of the four hierarchical variance components: week-to-week, biological, between-array and within-array variations.

In the next section (Section 2.2), we give an overview of the basic experimental design for a two-color dye-flip microarray experiment and illustrate it with an ANOVA model to the partition of sources of variation. We use a simple variance-component model involving three layers of variations, animal-to-animal (biological), between-array (technical) and residual variations, to provide the background and methods for estimation of variance of components and calculation of degrees of freedom (Montgomery, 1991). The simple variance-component model will be generalized to more complex effects such as within-array variations in the example data sets (Sections 2.2.1 and 2.2.2).

2.2 Experimental design, ANOVA and variance components

Consider a dye-flip design. Assume that the treated sample is labeled with Cy5 (red) dye and the control sample with Cy3 (green) dye on the first array. The second array has the dye assignment reversed. Each observation (of a gene) is obtained from different combinations among these three factors: array, dye and treatment. The complete model for the intensity measure of a gene is given by the ANOVA model

$$Y = m + A + D + T + (AD) + (AT) + (DT) + (ADT) + \epsilon,$$

where m represent the overall effect (average), A represents the main effect for arrays, D represents the main effect for dyes (red or green), T represents the main effect for treatments, (AD), (AT) and (DT) represent the interactions between main effects and (ADT) are the interactions among the three main effects. To estimate all eight effects, it would need eight observations from the eight different combinations: (A1, Cy3, Control), (A1, Cy5, Treatment), (A2, Cy3, Treatment), (A2, Cy5, Control), (A1, Cy3, Treatment), (A1, Cy5, Control), (A2, Cy3, Control) and (A2, Cy5, Treatment). However, the dye-flip experiment, in which only the first four (or the last four) combinations are observed, is a one-half fractional design. The effects can still be estimated by linear combinations of the half observations. Then, every two effects would share the same linear combinations and are confounded. The indistinguishable effects are called aliases. The generator is the effect that confounds with the overall effect m. In this design, (ADT) is the generator, and the defining relation is then given by $m = ADT$. An efficient way of determining the alias structure is multiplying any effect on both sides of the defining relation. Any effect can be obtained by multiplying the overall effect. For example, the alias of A is obtained by multiplying A on both sides of the defining relation, $A = A \cdot m = A \cdot ADT = A^2DT = DT$, since the square of any two-level effect is the constant. Thus, the completed alias structure in this design is $m = ADT$, $A = DT$, $D = AT$, $T = AD$. When the interactions are negligible, the fractional design is applicable for investigation of the main effects of interest.

The ANOVA model was initially developed for the analysis of differences between means (e.g. Kerr *et al*., 2000). The ANOVA technique was later adapted to estimating variance components. In the analysis of variance components of a data set, we are interested in attributing variability of the data to various factors. There are different levels of an effect in a factor that impacts the measurements of interest. The factors include treatment, dye, animal, array, etc. There are two kinds of effects for a factor, fixed effects and random effects. Fixed effects refer to the effects attributable to a finite set of levels of a factor that occur in the data. Random effects refer to the effects attributable to a (usually) infinite set of levels of a factor. The effects for the factor dye (red and green) or treatment (exposed and unexposed samples) are fixed, and the effects for arrays or animals are random because they are considered randomly chosen from some infinite population of arrays or animals. Therefore, a typical microarray experiment consists of both fixed effects and random effect factors. Analysis of variance components involve estimation of the variance of random effects.

In order to separate the variance of random components from treatment effects, we consider a variance-component model of a repeated dye-flip experiment within each treatment group. Let n_r denote the number of animals within each treatment group, n_a denote the number of arrays per animal and n_g denote the number of genes on the array. The experimental design is shown in Figure 1 with $n_r = 2$ and $n_a = 2$.

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

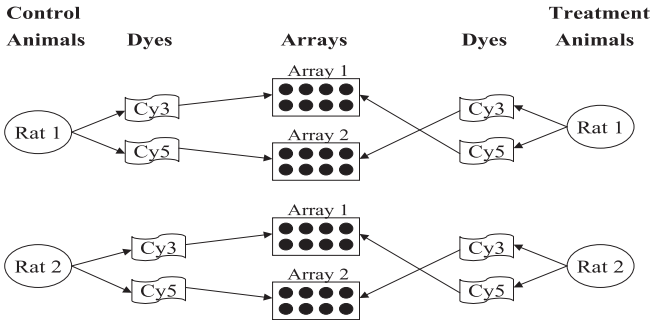

Fig. 1. Experimental design of a replicated dye-flip experiment.

A variance-component model that consists of array, dye, gene and animal (rat) effects is

$$Y = m + G + D + R + A(R) + \epsilon,$$

where G represents the effect of genes, R represents the effect of animals and $A(R)$ represents the between-array effect nested within animals, indicating the same RNA samples are hybridized with different arrays. Effects m, G and D are fixed, and effects R and $A(R)$ are random. Standard stochastic assumptions about the random effect components are that R, $A(R)$ and ϵ are independently normally distributed with mean 0 and with variances σ_r^2, $\sigma_{a(r)}^2$ and σ_ϵ^2, respectively. This model provides a simple structure of the three layers of variance components discussed before. The variance of a measured intensity is the sum of the three layers of the variance components, $\mathrm{Var}(Y) = \sigma_r^2 + \sigma_{a(r)}^2 + \sigma_\epsilon^2$. The variances σ_r^2, $\sigma_{a(r)}^2$ and σ_ϵ^2 can be estimated from the expectation of mean squares of the ANOVA table.

The expected mean squares of animal effect (MS_R), array effect ($\mathrm{MS}_{A(R)}$) and residual (MS_E) are $\mathrm{E}(\mathrm{MS}_R) = \sigma_\epsilon^2 + n_g n_a \sigma_{a(r)}^2 + n_g n_a n_r \sigma_r^2$, $\mathrm{E}(\mathrm{MS}_{A(R)}) = \sigma_\epsilon^2 + n_g n_a \sigma_{a(r)}^2$ and $\mathrm{E}(\mathrm{MS}_E) = \sigma_\epsilon^2$, respectively. Substituting the expectation with its ANOVA estimates, estimates of σ_ϵ^2, $\sigma_{a(r)}^2$ and σ_r^2 can be calculated by solving the above system of linear equations. This approach is known as the method of moments. In practice, an estimated variance can be negative. The restricted maximum likelihood (REML) method (Searle *et al*., 1992) is an alternative approach to estimating variance components. The REML estimates are maximum likelihood estimates with respect to the marginal likelihood excluding fixed effects by a vague prior. In balanced designs, REML and moment estimation give the same results under non-negative constrains.

A general rule for calculating the degrees of freedom is that for each effect, the degree of freedom is the product corresponding to the factors involved in the effect. Each term in the product is either the number of levels of the factor (for factors nested within parentheses) or the number of levels minus 1 (for factors not in parentheses). Thus, the degrees of freedom for G, D, R and $A(R)$ should be $(n_g - 1)$, $(2 - 1)$, $(n_r - 1)$ and $(n_a - 1)n_r$, respectively. However, due to the nature of the microarray design, each array consists of one dye (either red or green) for a given treatment. The effects of dye and array are, thus, confounded with each animal, i.e. the $A(R)$ and $D(R)$ are alias $A(R) = D(R) = D + DR$, where DR is the interaction between dye and animal. The $A(R)$ and D are then confounded. (The effect DR is not included in the above model.) Since the dye effect, D, is regarded as an important factor and is included in the model, this results in the loss of one degree of freedom for the $A(R)$ effect. The degree of freedom for $A(R)$ is $(n_a - 1)n_r - 1$.

2.2.1 Toxicogenomic data set The first data set is from a toxicogenomic study of gene expression changes of kidney samples from rats dosed with cisplatin, a known kidney toxin. Details of the study are given in Thompson *et al*. (2002). The array is a 700 gene cDNA rat chip from Phase-1 Molecular Toxicology (Santa Fe, NM). In each array there are four by four grids of 14×14 spots. The upper section (half of the array) consists of Grids 1–8, and the lower section is a replicate of the upper section, consisting of Grids 9–16. On each grid, genes were spotted in duplicate, and so each gene has four replicate values on each array. In addition to the 700 rat genes, sequences of four plant genes and one bacteria gene were also spotted on the array, each with four replications. A total of 705 genes were analyzed.

The control RNA was pooled from the kidneys of five rats. The treated RNA sample was pooled from the kidneys of five rats 7 days after being treated with 5 mg/kg of cisplatin. Control and treated samples were hybridized on six arrays (arrays A1–A6). Samples were all labeled and hybridized on one date. Each replicate sample was labeled independently (one treated and one control). There were 12 separate labeling reactions. On the arrays A1–A3, the control samples were assigned to the green dye and treated samples were assigned to the red dye. The dye assignments to the control and treated samples were reversed on the arrays A4–A6. Intensity was calculated by subtracting the local background intensities from each raw fluorescent value using the GenePix software package (Axon Instruments Inc., 1999). Since all biological samples were pooled, only technical and residual variances could be estimated. Control and treated samples were analyzed separately. Since the data set does not have biological replicates, only between-array and two levels of within-array variations are estimated.

Let y denote the background-subtracted base-2 logarithm of individual intensity measurements. Assume the three-factor interaction is negligible; a nested mixed-effects (ANOVA) model for the control samples can be expressed as

$$y_{aijws} = m + A_a + G_i + D_j + (GD)_{ij} + (GA)_{ia} + (DA)_{ja} + W(A)_{w(a)} + S(WA)_{s(wa)} + \epsilon_{aijws}, \quad (1)$$

where A represents the between-array effect ($a = 1, \ldots, 6$), G represents the gene effect ($i = 1, \ldots, 705$), D represents

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

the dye effect ($j = 1, 2$), $W(A)$ represents the between-section (within-array) effect attributed to upper and lower sections ($w = 1, 2$), $S(WA)$ represents the within-section effect (within-array) ($s = 1, 2$) and ϵ represents the random error in a given condition of a, i, j, w and s. As discussed, the gene and dye are assumed to be fixed effects, and the between-array, between-section (within-array) and within-section are assumed to be random effects. The within-array variations, $W(A)$, and the within-section (within-array) variation, $S(WA)$, are modeled as random effects. $S(WA)$ is the variation of intensities of two adjacent spots of the same gene within any section of an array. $W(A)$ is the variation of the intensities of the same genes between the upper section and the lower section. Both variations are likely due to the printing order. $S(WA)$ effect is interpreted as the printing order effect within any section within any array, which is independent of any particular gene or dye. $W(A)$ is the printing time effect within any array. The between-array effect, A, is likely due to the batch effect of preparing samples.

We restrict the model to the main effects and the interaction for two-factor fixed effects,

$$y_{aijws} = m + A_a + G_i + D_j + (GD)_{ij} + W(A)_{w(a)} + S(WA)_{s(wa)} + \epsilon_{aijws}. \quad (2)$$

The two random effect components, (GA) and (DA) [Equation (2)], are not included in this model; their variances $\sigma^2_{ga}, \sigma^2_{da}$ are negligible. Assume that the random effects A_a, $W(A)_{w(a)}$, $S(WA)_{s(wa)}$ and ϵ_{aijws} are independently normally distributed with mean 0 and with the variance components, σ^2_a, $\sigma^2_{w(a)}$, $\sigma^2_{s(wa)}$ and σ^2_ϵ, respectively. Under Model (2), the random error, ϵ, represents the overall variations of $\sigma^2_{ga}, \sigma^2_{da}$ and σ^2_ϵ. The variance of y_{aijws} is the sum of the variance components, $\sigma^2_a + \sigma^2_{w(a)} + \sigma^2_{s(wa)} + \sigma^2_\epsilon$. In order to investigate the distribution of variance components across genes, we consider the following mixed-effects model for gene-by-gene analysis,

$$y_{ajws} = m + A_a + D_j + W(A)_{w(a)} + \epsilon_{ajws}. \quad (3)$$

In the gene-specific model, the between-spot effect is not estimable, and it is confounded with the residual term, ϵ_{ajws}.

2.2.2 Circadian data set The second data set is a study of circadian changes in gene expression in liver samples from rats and was conducted at the National Center for Toxicological Research (NCTR), FDA. Details of the study are given in Desai *et al*. (2004). The rats were fed an *ad libitum* NIH-31 diet with a 12 h light/dark cycle (lights on at 0200 h and off at 1400 h). At 52 weeks of age, four rats were sacrificed at each of the following times: 0600, 1100, 1700 and 2300 h. Total RNA was extracted from the livers, generating 16 samples. A reference RNA was formed by mixing equal amounts of the 16 sample RNAs. In this study, the differences in gene expression among the four sacrifice times are the effect of interest. Microarrays were prepared using the rat 4K Ready-to-Print Long Oligos from Clontech (Palo Alto, CA; identities of the genes can be found at www.clontech.com). The RNA from each (test) sample was labeled with Cy3 and hybridized along with a Cy5-labeled reference sample to a glass slide containing duplicate sections of 3906 rat genes so that each gene had two measurements (upper and lower sections) on each array. In addition, dye-flip experiments were performed. After hybridization and washing, fluorescent signals were measured using an Axon 4000B scanner (Axon Instruments, Union City, CA). The Axon GenePix Pro software was used to quantify the signals from each gene spot for use in subsequent data analyses.

The 16 RNA samples were divided into four experimental blocks (four samples per block), each block consisting of a sample from each of the four sacrifice times. The samples within each block were labeled with one fluorophore and hybridized on a single day. The dye flip labeling and hybridization were conducted on the next day (one-half technical replicate). The four blocks were run in four different weeks, two consecutive days in each week, a total of eight arrays per week with two arrays per animal. Note that each day involved an animal from each of the four time points. In total, each gene had 128 intensity measurements (32 arrays × 2 dyes × 2 sections). For computational reasons, we only considered 955 genes where all 128 measurements were available.

For the reference samples, there is no biological effect (either animal-to-animal variation or sacrifice time) since the reference samples are all from the same pool. The mixed-effects model for the reference samples is

$$y_{abijwh} = m + B_b + H(B)_{h(b)} + A(HB)_{a(hb)} + G_i + D_j + (GD)_{ij} + W(AHB)_{w(ahb)} + \epsilon_{abijwh}, \quad (4)$$

where G_i represents the gene effect ($i = 1, \ldots, 955$), D_j represents dye effect ($j = 1, 2$), $(GD)_{ij}$ are the interaction between G_i and D_j, B_b represents the between-block effect ($b = 1, \ldots, 4$), $H(B)$ represents the between-day effect within block ($h = 1, 2$), $A(HB)$ represents the between-array effect within block and day ($a = 1, \ldots, 4$) and $W(AHB)$ represents the within-array (between-section) effect within block, day and array ($w = 1, 2$). The variance components of interest are the block variance, σ^2_b, between-day variance, $\sigma^2_{h(b)}$, between-array variance, $\sigma^2_{a(hb)}$, within-array variance, $\sigma^2_{w(ahb)}$ and residual variance, σ^2_ϵ. The gene-specific model is given by

$$y_{abjh} = m + B_b + H(B)_{h(b)} + A(HB)_{a(hb)} + D_j + \epsilon_{abjh}. \quad (5)$$

In this model, the within-array effect is not estimable, and it is confounded with the residual term ϵ_{ajbh}.

For the test samples, biological replicates enabled us to estimate biological variation (between-rat variation). Because there is only one biological sample for each sacrifice time in each week, the biological variability among

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

the four rats is confounded with biological differences at the sacrifice time. In order to estimate between-rat variance, we consider five housekeeping genes that are not expected to be changed with time among rats. The five genes were hypoxanthine-guanine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase, ornithine decarboxylase, polyubiquitin and tubulin alpha 1. Furthermore, the between-day effect (within an animal) and between-array effect are confounded in the 16 animals and four blocks. They cannot be estimated independently. We first consider the model without between-day effect. The nested mixed model for the five genes for the test samples is

$$y_{abrijw} = m + B_b + R(B)_{r(b)} + A(RB)_{a(rb)} + G_i + D_j + (GD)_{ij} + W(ARB)_{w(arb)} + \epsilon_{abrijw}, \quad (6)$$

where B represents the between-block effect, $R(B)$ represents the between-rat within block effect and $A(RB)$ represents the between-array effect within block and rat. The variance components of interest for the test samples are the block variance, σ_b^2, between-rat variance, $\sigma_{r(b)}^2$, between-array variance, $\sigma_{a(rb)}^2$, within-array variance, $\sigma_{w(arb)}^2$, and residual variance, σ_ϵ^2. Alternatively, we can model the between-day effect but without the between-rat effect,

$$y_{abhijw} = m + B_b + H(B)_{h(b)} + A(HB)_{a(hb)} + G_i + D_j + (GD)_{ij} + W(ARB)_{w(ahb)} + \epsilon_{abhijw}. \quad (7)$$

Finally, the variance-component ANOVA model for a gene-by-gene analysis with the rat effect and sacrifice time is

$$y_{tabrj} = m + T_t + B_b + R(B)_{r(b)} + A(RB)_{a(rb)} + D_j + \epsilon_{tabrj}, \quad (8)$$

where T represents the four sacrifice times effect ($t = 1, \ldots, 4$). This model is a generalization of the classical ANOVA model. For example, if there is no block or dye-effect, it becomes the split-plot ANOVA model (Montgomery, 1991, Chapter 14),

$$y_{tar} = m + T_t + R_r + A(R)_{a(r)} + \epsilon_{tar}.$$

In a typical split-plot model $R_r = R_r + (TR)_{tr}$ and $A(R)_{a(r)} = A_a + AR_{(ar)}$. The effects T and R are whole plot units and $A(R)$ is the sub-plot effect. When there is only one section per array (no within-array replicate), this model becomes the well-known one-way ANOVA model, $y_{tr} = m + T_t + \epsilon_{tr}$.

3 RESULTS

All analyses of the ANOVA model and variance component estimates were carried out using PROC VARCOMP of the SAS system for windows (SAS Institute, 1999). Although both PROC VARCOMP and PROC MIXED procedures can be used to analyze an ANOVA model with random effect components, PROC VARCOMP performs a type I analysis for a quantitative description of importance of each factor in the ANOVA model. The SAS output of the type I estimation, which uses the moment estimates, provides an analysis of variance table in addition to the variance component estimates. The analysis of variance table includes sum of square error, the degrees of freedom (df), mean square error and the expected mean square error (EMS) from each source of variation. PROC VARCOMP provides an alternative option of using restricted maximum likelihood (REML) estimation.

3.1 Analysis of toxicogenomic data set

Table 1 contains the ANOVA analysis and variance-component estimates from Equation (2). The ANOVA table (in upper panel) includes the degrees of freedom (df), mean square errors for the reference and treated samples and expected mean squares. The lower panel contains the variance-component estimates σ_a^2, $\sigma_{w(a)}^2$, $\sigma_{s(wa)}^2$ and σ_ϵ^2 with their respective asymptotic standard error estimate in brackets. Note that the number of degrees of freedom for the array effect is 4, instead of 5, due to the factional design for the dye-flip experiment. For instance, the control samples on array A1–A3 were labeled with green color and labeled with red color on array A4–A6. The dye effect and the array effect were confounded. The dye effect is the variation between the average of A1–A3 and the average of A4–A6. Estimating the dye effect results in the loss of one degree of freedom for estimating the array effect.

As discussed, the estimated residual variance, $\hat{\sigma}_\epsilon^2$, represents the pooled variance estimate for the experimental variation as well as higher ordered interaction effects such as Gene×Array, Dye×Array, etc. The estimated between-array variance, $\hat{\sigma}_a^2$, is much larger than both estimated within-array variances (between-section and within-section variances), $\hat{\sigma}_{w(a)}^2$ and $\hat{\sigma}_{s(wa)}^2$. The between-section variance itself is slightly larger than the within-section variance (variance between the adjacent spots, $\hat{\sigma}_{s(wa)}^2$). The smaller variance estimates appear to confirm that the within-array variance is due to printing order effect. It is worth mentioning that the dye effects and array effects are substantially different between treatment group and reference group (Table 1).

The distributions of the three variance components for treatment and reference samples are plotted in Figure 2. All variance components have positively skewed distributions. The shapes of the distribution of residual variance and within-array (within-section) variance are similar in both groups. The majority of the genes have small within-array as well as residual variance, the between-array effect being the major source of variation.

3.2 Analysis of circadian data set

Table 2 contains the ANOVA analysis and variance-component estimates for the reference sample, Equation (4).

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

Table 1. Analysis of variance and variance-component estimates [Equation (2)] for the toxicogenomic data set

Source	df	MS Control	MS Treated	EMS
Gene	704	56.8527	49.9613	$\sigma^2 + Q_1$
Dye	1	783.3850	74.9596	$\sigma^2 + 705\sigma^2_{s(wa)} + 1410\sigma^2_{w(a)} + 2820\sigma^2_a + Q_2$
Gene∗dye	704	0.7422	1.1134	$\sigma^2 + Q_3$
Array	4	360.4026	180.5637	$\sigma^2 + 705\sigma^2_{s(wa)} + 1410\sigma^2_{w(a)} + 2820\sigma^2_a$
Section(array)	6	15.9566	15.1519	$\sigma^2 + 705\sigma^2_{s(wa)} + 1410\sigma^2_{w(a)}$
Spot(section∗array)	12	3.8108	4.6545	$\sigma^2 + 705\sigma^2_{s(wa)}$
Residual	15488	0.1822	0.1838	
Variance-component estimates (standard error)				
Array: $\hat{\sigma}^2_a$		0.12214	0.05866	
		(0.09043)	(0.04538)	
Section(array) $\hat{\sigma}^2_{w(a)}$		0.00861	0.00745	
		(0.00663)	(0.00635)	
Spot(section∗array): $\hat{\sigma}^2_{s(wa)}$		0.00515	0.00634	
		(0.00221)	(0.00270)	
Residual: σ^2_ϵ		0.18221	0.18383	
		(0.00207)	(0.00209)	

Degrees of freedom (df), mean squares (MS) for the control and treated samples and the expected mean squares (EMS).

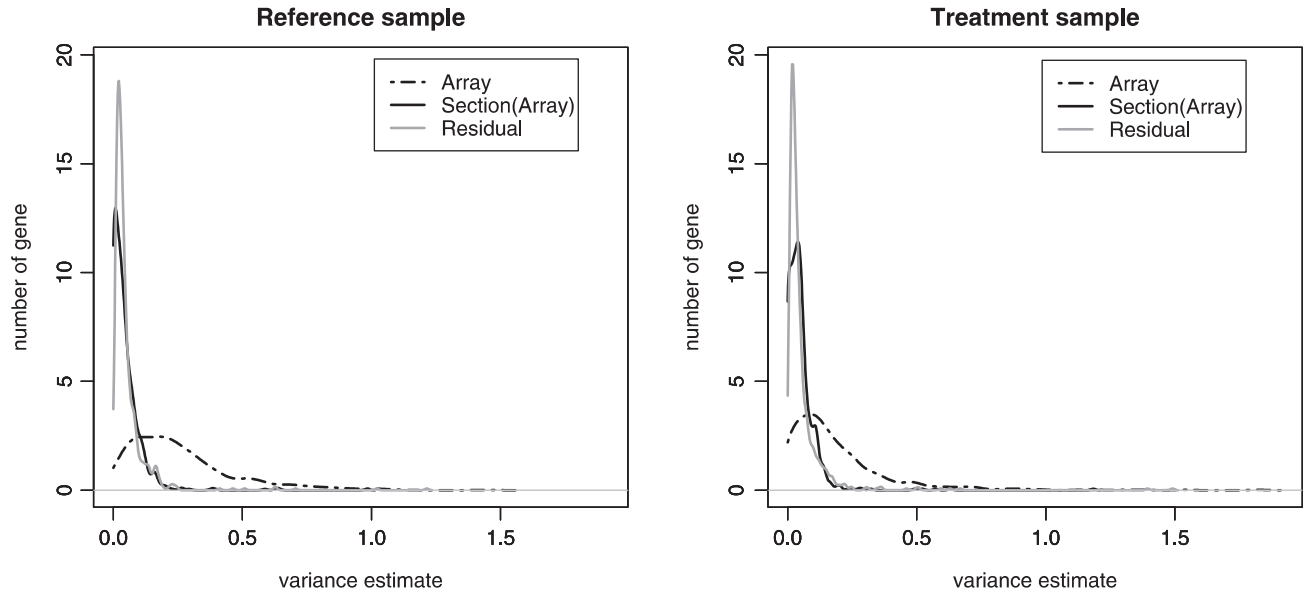

Fig. 2. Distributions of the estimated variance components for the toxicogenomic data set.

The variance-component estimates include block, day, between-array, within-array, and residual variances. Note that this design has 32 arrays with two dyes, four blocks and two days within a block. In this model, H and D are aliases within each block, that is, $H(B) = D(B) = D + DB$, where DB is the interaction between dye effect and the between-block effect. Thus, the $H(B)$ effect lost one degree of freedom for estimating the the dye effect.

The estimated variance due to day effects is small (0.0018). The day effect is the variation of intensities of the same gene hybridized in two different dates (with two arrays) within any week. The small variance suggests homogeneity of the hybridization procedure for the same biological batch. The three other variance component estimates, $\hat{\sigma}^2_b$, $\hat{\sigma}^2_{a(hb)}$ and $\hat{\sigma}^2_{w(ahb)}$, appear in a hierarchical pattern. The block effect has the largest variance and the between-array variance is

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

Table 2. Analysis of variance for the reference samples and variance component estimates, Equation (4), for circadian data

Source	df	MS	
Gene	954	40.3686	
Dye	1	665.5898	
Gene∗dye	954	2.0794	
Block	3	1599.8172	
Day(block)	3	201.7371	
Array(day∗block)	24	187.9890	
Section(array∗day∗block)	32	56.6009	
Residual	59148	0.1408	
Variance-component estimates (standard error)			
Block: $\hat{\sigma}_b^2$	0.09150	(0.08616)	
Day(block): $\hat{\sigma}_{h(b)}^2$	0.00180	(0.02270)	
Array(day∗block): $\hat{\sigma}_{a(hb)}^2$	0.06879	(0.02936)	
Section(array∗day∗block): $\hat{\sigma}_{w(ahb)}^2$	0.05912	(0.01482)	
Residual: $\hat{\sigma}_\epsilon^2$	0.14078	(0.00082)	

Degrees of freedom (df) and mean squares (MS) for the reference samples.

Table 3. Analysis of variance for five housekeeping genes in test samples and variance component estimates, Equation (6), for the circadian data

Source	df	MS	
Gene	4	12.7396	
Dye	1	10.0007	
Gene∗dye	4	1.9594	
Block	3	11.3173	
Rat(block)	12	1.7958	
Array(rat∗block)	15	0.6722	
Section(array∗rat∗block)	32	0.3235	
Residual	248	0.0809	
Variance-component estimates (standard error)			
Block: $\hat{\sigma}_b^2$	0.11902	(0.11589)	
Rat(block): $\hat{\sigma}_{r(b)}^2$	0.05618	(0.03866)	
Array(rat∗block): $\hat{\sigma}_{a(rb)}^2$	0.03487	(0.02584)	
Section(array∗rat∗block): $\hat{\sigma}_{w(arb)}^2$	0.04851	(0.01624)	
Residual: $\hat{\sigma}_\epsilon^2$	0.08094	(0.00727)	

Degrees of freedom (df) and mean squares (MS).

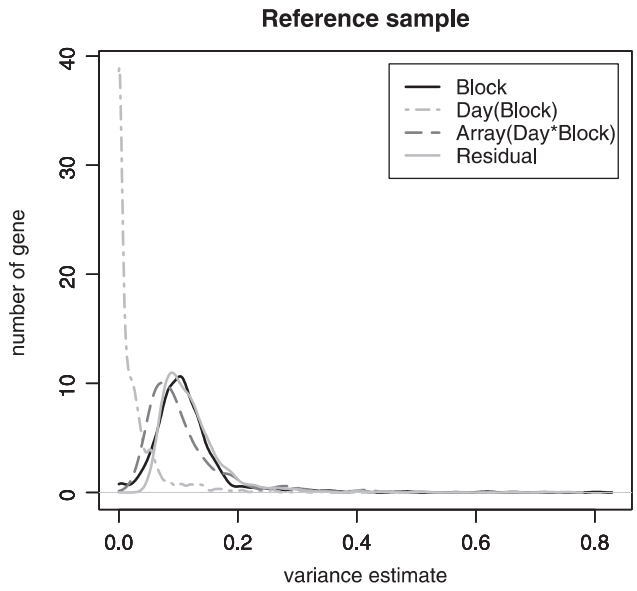

Fig. 3. Distributions of the estimated variance components for the circadian data set.

slightly larger than the within-array variance. The latter result is consistent with the result in Example 1. Figure 3 contains the plots of the distributions of the variance components from gene-by-gene analysis, Equation (5). Again, the day effect has small variances for most of the genes. The distributions for the remaining three variance components appear very similar.

Table 3 contains the ANOVA analysis and variance-component estimates for the test sample, Equation (6). The variance component estimates include block, between-rat, between-array, within-array and residual variances. Similar to the reference samples, the block effect has the largest variance. The between-rat (biological) variation is slightly larger than the between-array and within-array variation. Note that between-array variation consists of the two components: day effect and array effect. The technical variation consists of between-array (including day effect) and within-variation. These two component variances combined are larger than the biological variance.

For the alternative model, Equation (7), the variance-component estimates are

$$\hat{\sigma}_b^2 = 0.12682, \quad \hat{\sigma}_{h(b)}^2 = 0, \quad \hat{\sigma}_{a(hb)}^2 = 0.08481,$$
$$\hat{\sigma}_{w(ahb)}^2 = 0.04851, \quad \hat{\sigma}_\epsilon^2 = 0.08094.$$

The REML estimated variance for the day effect is 0 since the negative estimate is typically set to 0. Under this model, the between-array variance consists of the rat-variance and between-array variance. The estimated within-array and residual variance from (7) are identical to the estimates from (6). Since the estimated day-effect variance is 0, the estimate $\hat{\sigma}_{a(hb)}^2 = 0.08481$ probably reflects approximately the sum of between-rat variance ($\hat{\sigma}_{r(b)}^2 = 0.05618$) and between-array variance ($\hat{\sigma}_{a(rb)}^2 = 0.03487$) shown in Table 3 for the five housekeeping genes. In both reference and test samples, the between-block is the the major source of variation.

Table 4 gives the variance-component estimates (bottom panel) from the gene-by-gene analysis, Equation (8), for the five selected housekeeping genes. For Gene 1, Gene 2

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

Table 4. Analysis of variance for each of five housekeeping genes in test samples (degrees of freedom (df) and mean square) and variance component estimates, Equation (8), for the circadian data

Source	df	Mean square				
		Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Dye	1	0.5713	8.4281	3.3967	0.9039	4.5382
Time	3	0.4183	0.3225	0.3329	1.3160	0.7983
Block	3	2.7033	1.9111	2.4715	2.0921	2.4941
Rat (block)	9	0.2620	0.3077	0.3960	0.5873	0.5195
Array (rat∗block)	15	0.1180	0.1840	0.2295	0.3483	0.3208
Residual	32	0.0722	0.1137	0.1100	0.0937	0.0718
F-value (time)		1.60	1.05	0.84	2.24	1.54
P-value (time), from $F(3,9)$		0.2568	0.4168	0.5054	0.1529	0.2703
Variance-component estimates						
Block: $\hat{\sigma}_b^2$		0.15258	0.10021	0.12972	0.09405	0.12341
Rat (block): $\hat{\sigma}_{r(b)}^2$		0.03600	0.03093	0.04163	0.05975	0.04966
Array (rat∗block): $\hat{\sigma}_{a(rb)}^2$		0.02291	0.03518	0.05973	0.12730	0.12450
Residual: $\hat{\sigma}_\epsilon^2$		0.07219	0.11368	0.11004	0.09371	0.07184

Gene 1: hypoxanthine-guanine phosphoribosyltransferase; Gene 2: glyceraldehyde-3-phosphate dehydrogenase; Gene 3: ornithine decarboxylase; Gene 4: polyubiquitin; Gene 5: tubulin alpha 1.

and Gene 3, block effect dominates other sources of variation as seen in the previous results. Gene 4 has slightly larger biological variation. The technical variation in the level within-rat is mainly present in Gene 4 and Gene 5. This suggests that the majority of genes have moderate biological variance when compared with block variation and technical variation.

The ANOVA model, Equation (8), allows us to test the hypothesis of treatment effects. For purposes of illustration, assume that we are interested in knowing whether the expression profiles of, say, Gene 1 are different among the four sacrifice times. The significance of the hypothesis is revealed by the F-test (the last two rows of the upper panel in Table 4). Here the F-value is the ratio of the two mean squares of the Time effect and Rat(Block) effect, $MS_T/MS_{R(B)} = 0.4183/0.2620 = 1.60$. Note that the denominator of the F-statistic is not the conventional mean square of the residual because the biological samples are the basic experimental unit for inference from samples to population. The F-test measures the variation of the sacrifice times with the variation of rats. The mean square of residuals here represents the within-array variance. The p-values indicate no differences in the sacrifice times for all five genes. It should be noted that the above F-test presented for differences among sacrifice times was for illustrative purpose; only the test samples were used in the analysis. In practice, the data would include reference samples. Typically, the variable y in Equation (8) is the ratio of test to reference samples. Also, the permutation test or re-sampling method may be used instead of parametric t- or F-tests since the data often do not follow the normal distribution (Tsai *et al*., 2003).

4 DISCUSSION

Variability in microarray data is expected and unavoidable. The identification and estimation of different sources of variation are fundamental to the design of cost-efficient microarray experiments. The basic principles of experimental design are randomization, blocking and replication. The purpose of randomization is to reduce the likelihood of systematic biases due to selection or assignment. For example, the biological samples should be randomized to a treatment using a predetermined scheme so that the underlying characteristics of subjects are equally representative across treatments. Randomization can also be applied to the dye assignments in technical replicates. Replication, to minimize technical artifacts and to assess biological variability, is the key to the accuracy and reliability of the data. Replication enables us to understand and interpret the significance of observed changes for thousands of genes. Blocking is used to increase the precision of estimates; a block is a subset of experimental units that are more homogeneous than the entire experiment itself. Below we describe the use of variance-component estimates in conjunction with replication and blocking techniques in planning microarray experiments.

In the circadian data set, the basic measurement for a gene and sacrifice hour is the average of the measurements of a rat made on a particular week. These averages are then averaged across rats and week. Denote n_b (=4) as the number of weeks, n_r (=4) as the number of animals per week, n_a (=2) as the number of arrays per animal and n_l (=2) as the number of sections per array. Let $x_{rb\cdot\cdot} = \sum_{a,l} x_{rbal}/(n_a n_l)$ represent the basic experimental measurement for a given week, that

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

is, the average for the n_l sections for a gene and sacrifice hour for the r-th rat in the b-th week. The total number of measurements for the mean value of $x_{rb..}$ is $n_r \times n_b$. The variance of the mean of the $x_{rb..}$ is

$$\mathrm{Var}\left(\frac{\sum_{r,b} x_{rb..}}{n_r n_b}\right) = \frac{\sigma_b^2}{n_b} + \frac{\sigma_{r(b)}^2}{n_b n_r} + \frac{\sigma_{a(rb)}^2}{n_b n_r n_a} + \frac{\sigma_{w(arb)}^2}{n_b n_r n_a n_l} + \frac{\sigma_\epsilon^2}{n_b n_r n_a n_l}. \quad (9)$$

The variance estimates can be obtained from the variance-component estimates (Tables 2–4).

Equation (9) can be used to calculate the number of animals, of arrays per animal and of sections per array when the variance component estimates for σ_b^2, $\sigma_{r(b)}^2$, $\sigma_{a(rb)}^2$, $\sigma_{w(arb)}^2$ and σ_ϵ^2 are available. In general, if the information on the relative cost of measurements for sections within arrays and the cost of arrays (per rat) are available, then an optimum number of sections per array (possible only 1 or 2) and optimum number of arrays per rat can be calculated. In the special case, assume an equal cost of measurements for arrays within a rat and in different rats, for a fixed total number of rat, array and section combinations ($n_r \times n_a \times n_l$), the variance is minimized when $n_s = n_w = 1$. That is, the design with one gene per array per biological sample (rat) is the most efficient design.

In the context of a microarray experiment, replicated spots of the same gene are not only used for averaging and for reduction of variance [Equation (9)] but have also been used for evaluation of overall quality of the data. For example, the four plant genes and one bacteria gene in the toxicogenomic data set are regarded as replicates and are used to monitor non-specific background binding of labeled cDNA. Replicates of the same genes on an array are not only useful for evaluation of the quality of data but also offer protection against missing data (e.g. not-hybridized data points). However, replicated spots should be well spaced, not adjacent, so that the true variability within an array can be estimated.

Equation (9) provides a general formula to determine the number of replicates at different stages in the experiment. It is worth remembering that extrapolation of statistical inferences from sample to population is valid only through biological replication. Variability of the population of interest in a study can only be estimated by biological replication. Biological samples should reflect the variability of the population. In the experiment, biological samples can be pooled to reduce the biological component of variation and increase statistical power to detect a treatment difference (Kendziorski *et al*., 2002). Pooling, however, does not reduce the variation of the technical component. Technical replicates are needed to estimate or to reduce measurement variation. Furthermore, making a large pool from all biological samples for each treatment will minimize the biological variation. However, this design losses information on the variability of the population and statistical inference can be made only on the experimental samples.

In summary, we present an analysis of estimating sources of variation and their relative contributions to the overall variation in microarray studies. A large week-by-week variation is observed. That is, the performance of procedures from one week run (block) to the next run accounts for much of the overall variability. Reduction of this variability would increase the precision of the estimates of gene expression changes. The number of runs required to perform a study is determined by the size of the resources available on a daily basis to conduct the experiments. With limited resources, dividing the entire study into multiple experimental runs is unavoidable. Development of rigid standard operating procedures, as well as limiting the number of experimental runs, would be expected to have a big impact on reducing this run-to-run variability and therefore the overall variability. In addition, it is critical that each treatment is represented on each experimental block in order to avoid confounding and biases that would be introduced due to the different blocks. Furthermore, the technical variation, that is, the sum of the between-array and within-array variations, appears to be larger than the biological variation. The between-array variance is slightly larger than the within-array variance in both data sets. Technical replicates can be made on replicate arrays or replicate spots (of the same gene) on the array. Since the cost of replicate spots is less than the cost of replicate arrays, an array with well-spaced replicated spots is a cost-effective design to reduce technical variation. Another source of variability is animal to animal variation, although we found that the effect is less substantial in the second example. Equation (9) shows that replication of biological samples is generally the most effective way of reducing this variability with a fixed number of blocks and increasing the power for identifying differentially expressed genes. Thus, by identifying the sources and magnitude of variability in two microarray data sets, we show that reduction in overall variability may be obtained by modification of experimental protocols and inclusion of more biological replication.

REFERENCES

Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarray. *Nature Genetics Supplement*, **32**, 490–495.

Cui,X.Q. and Churchill,G.A. (2003) How many mice and how many arrays? Replication in mouse cDNA microarray experiments. CAMDA Competition available from http://www.jax.org.

Desai,V.G., Moland,C.L., Branham,W.S., Delongchamp,R.R., Fang,H., Duffy,P.H., Peterson,C.A., Beggs,M.L. and Fuscoe,J.C. (2004) Changes in expression level of genes as a function of time of day in the liver of rats. *Mut. Res.* (in press).

Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gen-expression microarray data. *Bioinformatics*, **18**, S105–S110.

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024

Kendziorski,C.M., Lan,H. and Attie,A.D. (2002) The efficiency of pooling mRNA in microarray experiments. *Technical Report #168*. Department of Biostatistics, University of Wisconsin, Madison, WI.

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

Montgomery,D. (1991) *Design and Analysis of Experiments*. Third Edition. John Wiley and Sons, Inc., New York.

Novak,J.P., Sladek,R. and Hudson,T.J. (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, **79**, 104–113.

Searle,S.R., Casella,G. and McCulloch,C.E. (1992) *Variance Components*. John Wiley and Sons, Inc., New York.

Spruill,S.E., Lu,J., Hardy,S. and Weir,B. (2002) Assessing sources of variability in microarray gene expression data. *Biotechniques*, **33**, 916–923.

Thompson,K.L., Mirsky,M.L., Kadyszewski,E. and Sistare,F.D. (2002) Concordance of degree of renal injury with gene expression in individual animals treated with the nephrotoxicant cisplatin. *The Toxicologist*, **66**, 297.

Tsai,C.A., Chen,Y.J. and Chen,J.J. (2003) Testing for Differentially Expressed Genes with Microarray Data. *Nucleic Acids Res.*, **31**, e52.

Wernisch,L. (2002) Can replication save noisy microarray data. *Comp. Funct. Genom.*, **3**, 372–374.

Yang,Y.H. and Speed,T.P. (2002) Design issues for cDNA microarray experiments. *Nature Rev. Genet.*, **3**, 579–583.

Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Downloaded from https://academic.oup.com/bioinformatics/article/20/9/1436/195662 by guest on 25 April 2024