JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Number 6, 2000
Mary Ann Liebert, Inc.
Pp. 819–837

Analysis of Variance for Gene Expression Microarray Data

M. KATHLEEN KERR,[1] MITCHELL MARTIN,[2] and GARY A. CHURCHILL[1]

ABSTRACT

Spotted cDNA microarrays are emerging as a powerful and cost-effective tool for large-scale analysis of gene expression. Microarrays can be used to measure the relative quantities of specific mRNAs in two or more tissue samples for thousands of genes simultaneously. While the power of this technology has been recognized, many open questions remain about appropriate analysis of microarray data. One question is how to make valid estimates of the relative expression for genes that are not biased by ancillary sources of variation. Recognizing that there is inherent "noise" in microarray data, how does one estimate the error variation associated with an estimated change in expression, i.e., how does one construct the error bars? We demonstrate that ANOVA methods can be used to normalize microarray data and provide estimates of changes in gene expression that are corrected for potential confounding effects. This approach establishes a framework for the general analysis and interpretation of microarray data.

Key words: Gene expression microarray, differential expression, analysis of variance, bootstrap.

INTRODUCTION

THE REGULATION OF GENE EXPRESSION in a cell begins at the level of transcription of DNA into mRNA. Although subsequent processes such as differential degradation of mRNA in the cytoplasm and differential translation also regulate the expression of genes, it is of great interest to estimate the relative quantities of mRNA species in populations of cells. The circumstances under which a particular gene is up- or down-regulated provide important clues about gene function. The simultaneous expression profiles of many genes can provide additional insights into physiological processes or disease etiology that is mediated by the coordinated action of sets of genes.

Spotted cDNA microarrays (Brown and Botstein, 1999) are emerging as a powerful and cost-effective tool for large scale analysis of gene expression. In the first step of the technique, DNA clones with known sequence content are spotted and immobilized onto a glass slide or other substrate, the microarray. Next, pools of mRNA from the cell populations under study are purified, reverse-transcribed into cDNA, and labeled with one of two fluorescent dyes, which we will refer to as "red" and "green." Two pools of differentially labeled cDNA are combined and applied to a microarray. Labeled cDNA in the pool hybridizes to

[1]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609.
[2]Genome and Information Sciences, Hoffmann-La Roche, Inc., Nutley, NJ.

complementary sequences on the array and any unhybridized cDNA is washed off. Hybridization efficiency may vary from clone to clone, confounding comparisons between genes. However, if we assume that the efficiency of an individual clone is not altered by the type of the dye label, then the relative abundance of a particular mRNA in the two samples can be measured.

Microarray technology has the potential to address many interesting questions in genetics by revealing patterns of expression for genes and classifying samples (such as tumor samples) based on such patterns. However, basic questions about microarray data persist without satisfactory answers. The simplest microarray experiment studies the variation in gene expression across the categories of a single factor, such as tissue types, strains of mice, or drug treatments. (We refer to the categories of the factors under study as *varieties*, as is common in the statistical design literature.) The purpose of such an experiment is to identify differences in gene expression among the varieties. Since there are other sources of variation in these experiments, such as the two dyes and the arrays themselves, how does one estimate the magnitude of differences for the spotted genes? Further, given that there is inherent "noise" in the data, how does one state one's confidence in the estimates? In particular, how does one determine what level of observed differential expression is statistically significant? Error estimates are necessary for making valid, rigorous inferences from the experiment (Fisher, 1935, p. 60). Looking ahead, we believe they will also be useful in assessing the quality of the results from higher-order analyses such as clustering (Eisen *et al*., 1999; Tamayo *et al*., 1999).

In this work, we perform analysis of variance on microarray data from two designed experiments that used independent arrays to study the same tissue samples. We employ a bootstrapping technique to construct confidence intervals for the estimates of interest. Comparing the results of the two separate analyses demonstrates the reproducibility of estimated changes in expression levels.

RESULTS

ANOVA models for microarray data

A microarray experiment may involve multiple arrays to compare multiple samples. Every measurement in a microarray experiment is associated with a particular combination of an array in the experiment, a dye (red or green), a variety, and a gene. Let y_{ijkg} denote the measurement from the i^{th} array, j^{th} dye, k^{th} variety, and g^{th} gene. To account for the multiple sources of variation in a microarray experiment, consider the model

$$\log(y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg}, \tag{1}$$

where μ is the overall average signal, A_i represents the effect of the i^{th} array, D_j represents the effect of the j^{th} dye, V_k represents the effect of the k^{th} variety, G_g represents the effect of the g^{th} gene, $(AG)_{ig}$ represents a combination of array i and gene g (i.e., a particular spot on a particular array), and $(VG)_{kg}$ represents the interaction between the k^{th} variety and the g^{th} gene. The error terms ϵ_{ijkg} are assumed to be independent and identically distributed with mean 0. The array effects A_i account for differences between arrays averaged over all genes, dyes, and varieties. These may arise, for example, because arrays are hybridized under slightly different conditions that result in a change in hybridization efficiency across an array. Similarly, the dye effects D_j account for differences between the average signal from each dye. One dye may be inherently "brighter" than the other, and this must be taken into account in the analysis. The terms V_k account for overall differences in the varieties. Such differences could arise if some varieties have more transcription activity in general, or simply because of differential concentration of mRNA in the labeled sample. The terms G_g account for average effects of individual genes spotted on the arrays in the experiment. The $(AG)_{ig}$ account for the average effect of the spot on array i for gene g. Essentially, these are "spot" effects and may arise because there is not complete control over the amount and concentration of cDNA immobilized from one array to the next. All of these effects are generally not of interest, but account for sources of variation in microarray data. It is also possible to include other effects, such as dye$\times$gene interactions. However, as we discuss below, including additional effects uses degrees of freedom that may need to be reserved to estimate the error variance in the experiment. The effects of interest in

Model (1) are the interactions between varieties and genes, $(VG)_{kg}$. These terms capture departures from the overall averages that are attributable to the specific combination of a variety k and a gene g. Nonzero differences in variety×gene interactions across varieties for a given gene indicate differential expression.

Our decision to analyze the data on the log scale was based on several considerations. The log transform is the natural method for analyzing data with an additive model where the effects in the data are believed to be multiplicative. The common use of ratios to analyze microarray data illustrates that this is a prevalent assumption, and in fact some tools for clustering genes based on microarray data advise using the log transform on ratios (Eisen, 1999). Further, exploration of untransformed data and the examination of other transformations (square-root, reciprocal, etc.) led us to conclude that the log transform is a good choice (Sapir and Churchill, 2000).

The terms A, D, and V in the ANOVA model are used to capture differences that occur between different arrays, dyes, and varieties. However, these terms also capture all of the higher-order interactions among these factors. This is a consequence of the constraints on the design of microarray experiments that are imposed by pairing samples on arrays. For example, if the array number and dye of an observation are given, one knows which variety is associated with that observation. In this situation, the array×dye interaction (AD) is said to be confounded with the variety main effect (V). Confounding is an advantage in this setting. If there is significant variation in the rate of dye incorporation from one labeling reaction to another, this will result in a large dye×variety interaction (DV) effect. In our first experiment, DV is confounded with array (A), and a large A effect is observed.

The A, D, and V terms effectively normalize the data without preliminary data manipulation. Thus we combine the normalization process with the data analysis. We believe this integrated approach has several advantages. First, the normalization is based on a clearly stated set of assumptions that can be evaluated using information in the data. Second, the ANOVA analysis systematically estimates the normalization parameters based on all of the relevant data, as opposed to a piecemeal approach. In so doing, it properly accounts for the degrees of freedom used to normalize. In the event that further study shows preprocessing is necessary, we believe that ANOVA methods will remain useful and valuable in some modified form.

Finally, the Model (1) is designed for experiments in which each gene is spotted only once on each array. Ideally, genes could be replicated on multiple spots on an array, providing a means to directly assess experimental error variance. Model (1) can be generalized to this situation by breaking down the "spot" effects AG to account for replication. As one would expect, replication would lead to more precise estimation. In addition, it would provide degrees of freedom that would allow one to assess the importance of additional effects in the model. Lack of replication limits our ability to assess some effects. We will return to this point in our data analysis examples.

The Latin square experiment

In the first experiment, we compared an mRNA sample obtained from human liver tissue to a second sample obtained from muscle tissue. The design used two arrays such that on array 1 the liver sample is assigned to the "red" dye and the muscle sample is assigned to the "green" dye. On array 2 the dye assignments were reversed (Table 1). We assigned the array index to be $i = 1, 2$; the dye index to be $j = 1, 2$ for red and green, respectively; and the tissue index to $k = 1, 2$ for liver and muscle, respectively. This design can be summarized by the index set $(i, j, k) \in \{(1, 1, 1), (1, 2, 2), (2, 1, 2), (2, 2, 1)\}$. Each clone index $g = 1, \ldots, N$ occurs once with each combination of (i, j, k). Notice that specifying any two of array, dye, and tissue automatically determines the third. With respect to the design factors, array and

TABLE 1. THE LATIN SQUARE DESIGN

	Array	
Dye	1	2
Red	Liver	Muscle
Green	Muscle	Liver

dye, the layout of the tissue varieties forms a 2 × 2 Latin square (Cochran and Cox, 1992). We therefore refer to this as the *Latin square* design (it is sometimes called a "dye-swap" experiment).

Given the factors in our model, there are sixteen possible effects when we consider interactions of all orders. It turns out that the latin square design has a particularly neat structure. Each of the sixteen effects is completely confounded with one other effect, meaning one effect is estimable only assuming the other is zero. Table 2 shows the pairs of confounded effects. Effects that are not completely confounded are *orthogonal* in the Latin square. Orthogonality arises when a factor is completely balanced with respect to another factor. For example, if every variety in a microarray experiment appears in the design labeled with the red and green dyes equally often, variety is orthogonal to dye. One consequence of orthogonality is that the estimates of the two factors are uncorrelated. A second consequence is that including or excluding one effect in the model does not alter the estimates obtained for the other effect. In general, effects that are neither confounded nor orthogonal are said to be *partially confounded*.

Examining Table 2, we see there is one pair of effects not represented in the model (1), $DG \sim AVG$. It is possible that DG effects could be present in a microarray experiment. However, leaving them out of the latin square analysis will not alter the estimates of other terms in the model. This is only true for designs in which the DG effect is orthogonal to the other effects. Omitting DG effects leaves degrees of freedom for estimating error. Assigning some effects to be "error" is essential when there is no replication of clones within the arrays. Otherwise, there is no basis for statistical inference.

We computed the least-squares fit of the Model (1) subject to the parameter constraints $\sum A_i = \sum D_j = \sum V_k = \sum G_g = \sum_g (AG)_{ig} = \sum_i (AG)_{ig} = \sum_g (VG)_{kg} = \sum_k (VG)_{kg} = 0$. Some details about estimating model parameters are provided in the appendix. Table 3 gives the analysis of variance. We present the sums of squares as a gauge of the relative contribution of each set of effects. For example, one sees from the sums of squares that there is a large difference between the two arrays, compared to a modest tissue effect and an even smaller dye effect. The large array effect may be due to variety×dye (i.e., labeling) variation—recall from Table 2 that A is completely confounded with DV.

Accounting for degrees of freedom, the smallest effects are the array×gene or "spot" effects. We did not want to rely on the F-distribution to test the significance of these effects as we have not established normally distributed error. Instead, we employed a nonparametric version of the F-test to determine the significance of these interactions, following an example in Manly (1997, p. 128) motivated by Still and White (1981). We first adjusted the data to remove the overall effects of the other factors. In other words, we created a dataset from the residuals from fitting the model $\log(y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (VG)_{kg}$. We then randomly assigned residuals to factor combinations by sampling with replacement, fit the full Model (1), and calculated the F-statistic testing for array×gene interactions. The F-statistics from 19,999 simulations

Table 2. Confounding Structure for the Latin Square Design[a]

mean	∼	ADV
A	∼	DV
D	∼	AV
V	∼	AD
G	∼	AD*VG*
VG	∼	*ADG*
AG	∼	D*VG*
DG	∼	*AVG*

[a]This design partitions the 16 experimental factor effects into eight pairs. The members of each pair are completely confounded; i.e., one member of a pair is estimable only by assuming the other is zero. The Latin square design results in uncorrelated estimates for all effects not in the same pair. The proposed Model (1) includes an effect from every pair except the last. Thus it accounts for all data effects except *DG* and *AVG*, which are assumed to be zero.

TABLE 3. ANALYSIS OF VARIANCE FOR THE LATIN SQUARE DESIGN[a]

Source	df	SS	MS
Array	1	92.34	92.34
Dye	1	0.74	0.74
Variety	1	2.97	2.97
Gene	1285	1885.89	1.47
Array×Gene	1285	160.01	0.12
Variety×Gene	1285	1357.28	1.06
Residual	1285	82.75	0.0644
Corrected Total	5143	3581.99	

[a]The correlation coefficient of the fitted model is $R^2 = 0.977$. Abbreviations: df—degrees of freedom; SS—sum of squares; MS—mean square.

ranged from 0.81 to 1.27, compared to 1.93 for the original data. We therefore concluded array×gene effects are statistically significant, although relatively small.

The estimated gene effects, G_g, and variety×gene interactions, $(VG)_{kg}$, are summarized as histograms in Figs. 1a and 1b, respectively. The estimated gene effects reflect the expression levels of individual genes averaged across varieties, dyes, and arrays. As noted in the introduction, these effects are confounded by variation in the hybridization properties of individual spotted clones. The value of such estimates is yet to be established and will depend on the magnitude of clone-to-clone variation. We simply present a summary of these estimates and note that they are skewed right, suggesting that the bulk of genes may be expressed at low levels with fewer genes being expressed at moderate and high levels in these samples. We note that unexpressed genes may have been eliminated when we prescreened the data for signal quality (see section on data preparation below). The variety×gene interactions are centered around zero with heavy tails in either direction, indicating differential expression of genes across the tissue samples.

For the Latin square design, the estimated differences in the variety×gene interaction terms for a given gene g_0 can be expressed as

$$(\widehat{VG})_{1g_0} - (\widehat{VG})_{2g_0} = \frac{1}{2}\log\left(\frac{y_{111g_0}y_{221g_0}}{y_{122g_0}y_{212g_0}}\right) - \frac{1}{2N}\log\left(\prod_g \frac{y_{111g}y_{221g}}{y_{122g}y_{212g}}\right). \quad (2)$$

We again note that this estimator does not change if we alter Model (1) by including *DG* effects or dropping *AG* effects or both. This is a consequence of the balanced (orthogonal) design.

Despite its rather intimidating appearance, the interpretation of (2) is straightforward. The second term is simply a centering constant that does not depend on the particular gene g_0. It corrects for the overall difference in treatments across genes. The first term is the log of the ratio of the geometric means of the observations for the gene g_0 in the two treatment groups. Thus, the exponentiated differences can be interpreted as estimates of "fold change." This interpretation is one motivation for working on the log scale. However, instead of relying on raw ratios as error-free measures of relative expression, we can further estimate the error-variation in the estimates (2) resulting from Model (1). We discuss this next.

We wish to determine which of the differences (2) are significantly different from zero. Least-squares estimates are averages, so under the assumption of independent, identically distributed error, the central limit theorem tells us that they are asymptotically normal. However, this justification is problematic for the variety×gene interactions because they are essentially averages of just two observations, too few to invoke large-sample arguments. Furthermore, the fitted residuals appear to be heavy-tailed, as illustrated in Fig. 2a. These observations suggest that the usual confidence intervals based on normal theory are not appropriate. Therefore, we employed a bootstrap analysis of the residuals (Efron and Tibshirani, 1986) to address this question.

Using the bootstrap, we produced a set of simulated datasets $\log(y_{ijkg})^*$, where

$$\log(y_{ijkg})^* = \hat{\mu} + \hat{A}_i + \hat{D}_j + \hat{V}_k + \hat{G}_g + (\widehat{AG})_{ig} + (\widehat{VG})_{kg} + \epsilon^*_{ijkg}.$$

The notation "^" indicates an estimated parameter value based on the original fit of the model. The ϵ^*_{ijkg} are drawn independently from $\sqrt{4N/(N-4)}\hat{F}$, where $\hat{F}$ is the empirical distribution of residuals from the original fit. Rescaling $\hat{F}$ produces an empirical distribution with the same variance as the true residuals (Wu, 1986). Thus, we are resampling, with replacement, from the rescaled fitted residuals to generate a new set of observations. We fit the Model (1) to each of 20,000 bootstrap data sets and recorded the parameter estimates. We then used the percentile method to obtain 99% confidence intervals for the differences $(VG)_{1g} - (VG)_{2g}$. The bootstrap confidence interval width was 1.61, which implies that an estimated fold change of $e^{1.61/2} = 2.24$ is significant at the 0.01 level. The normal confidence interval for these data has width 1.29. We note that multiple testing has not been taken into account, which may or may not be necessary depending on the intended purpose of the analysis.

The bootstrap procedure assumes that residuals are identically distributed. Fig. 2b shows a scatterplot of residuals against the fitted values $\widehat{\log(y_{ijkg})}$. There is no obvious trend in the residual plot to cast doubt on the assumption of constant error variance σ^2. To further examine the distribution of residuals, we plotted the absolute value of each residual against the fitted values $\widehat{\log(y_{ijkg})}$ and fit a local regression curve (Hastie and Tibshirani, 1990, p. 29). Fig. 3a shows there is no overwhelming trend in the absolute size of the residuals, with only a very slight trend towards larger residuals for the smallest and largest fitted values. The fact that the residual plot is unremarkable is also evidence that the log scale is the appropriate transform of the data.

The reference sample experiment

In the second experiment, we made an independent comparison of the same samples used in the first. Again we used two arrays, but in this case we used placenta as a "reference" sample. Each of the muscle and liver samples were directly compared to the placenta sample on one array such that the test samples (liver and muscle) were assigned to the green dye and the reference sample (placenta) was assigned to the red dye, as in Table 4. The array and dye indices are assigned as before. The tissue index is expanded to $k = 1, 2, 3$ for liver, muscle, and placenta respectively. This design can be summarized by the index set $(i, j, k) \in \{(1, 1, 3), (1, 2, 1), (2, 1, 3), (2, 2, 2)\}$. We refer to this design as the *reference sample* design.

One advantage of the reference sample design is that it is readily extendable. Additional varieties can be added to the experiment by adding another array on which a new variety is compared to the reference sample. A second advantage is that each sample needs only to be labeled with one dye. However, one can intuitively appreciate some of the drawbacks of this design. More data are collected on the reference variety than any other, although this variety will generally be of least interest. In this case, there is only one measurement per gene for liver and muscle tissues, as compared to two measurements in the Latin square design.

More problems with this design become apparent when one considers Model (1). First, varieties are completely confounded with dyes because each variety is labeled with only one dye. Thus, one cannot include both variety effects and dye effects in an ANOVA model. This in itself is not a great concern because these main effects are not of interest. The more substantial problem is the large cost in degrees of freedom that comes with the additional reference variety. With N genes on each array, there are $4N - 1$ degrees of freedom. Array main effects account for 1 degree of freedom, variety main effects account for 2, gene effects use $N - 1$, VG effects use $2(N - 1)$, and AG effects account for the remaining $N - 1$ degrees of freedom. At least one set of effects must be excluded to be able to estimate error and allow statistical inference. We note that if genes were spotted more than once on the arrays, it would be possible to include AG terms in the model and still have degrees of freedom to estimate error.

The confounding of effects in the reference design is more complex than for the latin square design. There is no counterpart to the simple confounding structure presented in Table 2. As mentioned, varieties are completely confounded with dyes. In addition, since the varieties are not balanced with respect to arrays, variety main effects and array main effects are partially confounded, as are variety×gene interactions (the effects of interest) and array×gene interactions. When effects are partially confounded instead

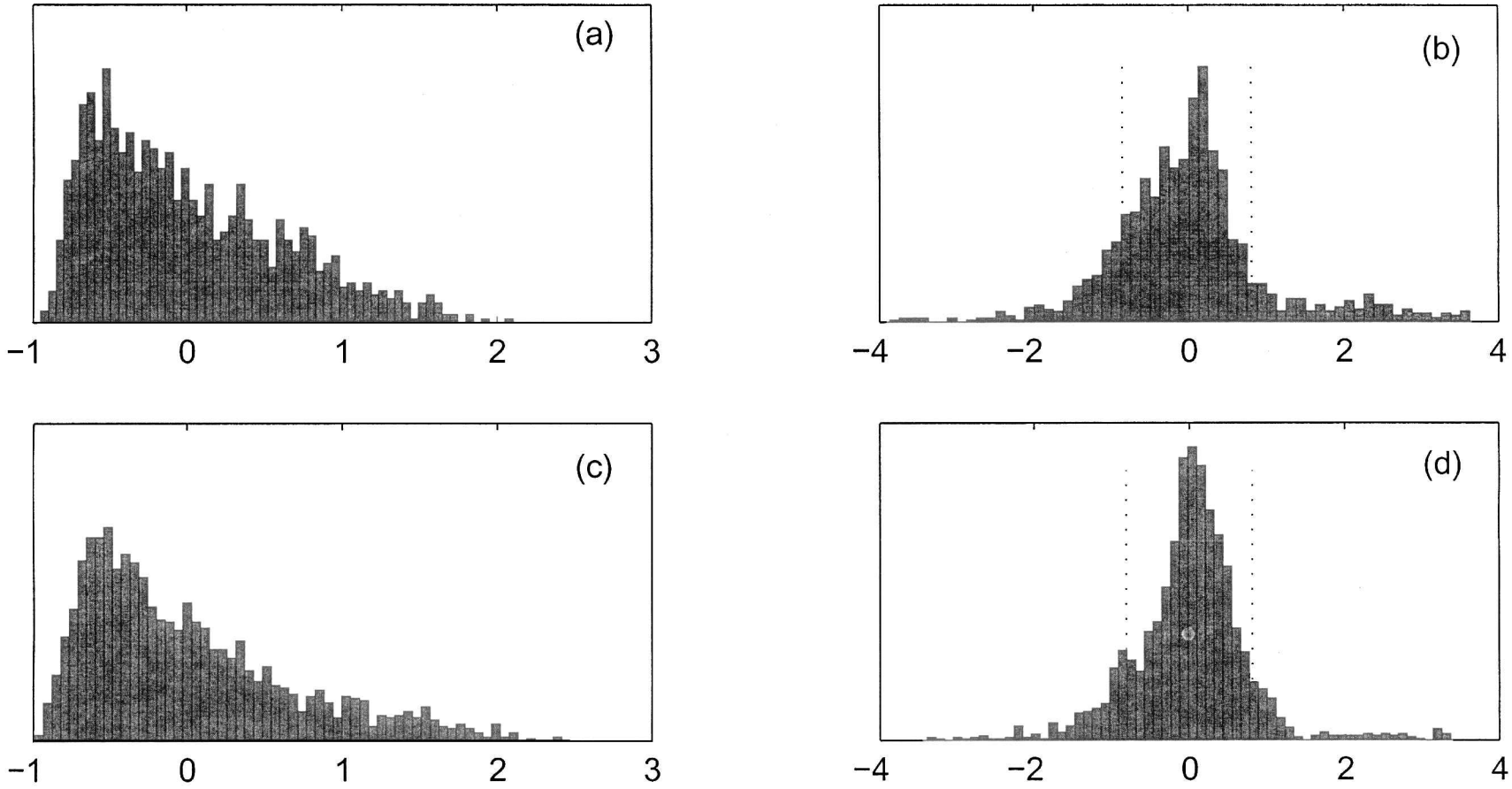

FIG. 1. Distribution of the estimated effects. Histograms of the estimated gene effects G_g are shown for (**a**) the Latin square and (**c**) the reference design. Histograms of the differences $(\widehat{VG})_{1g} - (\widehat{VG})_{2g}$ between variety×gene interaction effects for liver and muscle samples are shown for (**b**) the Latin square design and (**d**) the reference design. Dotted lines indicate the threshold for estimated difference that are significantly different from 0 according to the bootstrap 99% confidence interval.

of completely confounded, it is possible to obtain separate estimates for each effect, although they are correlated. Generally, the estimators have a more complicated functional form because the effects must be "disentangled." This usually means less precise estimation, i.e., larger error bars. Failure to account for potentially important effects, such as *DG* or *AG*, that are confounded or partially confounded with effects of interest can produce biased estimates.

We first consider the model

$$\log(y_{ijkg}) = \mu + A_i + V_k + G_g + (VG)_{kg} + \epsilon_{ijkg}, \tag{3}$$

where the V_k nominally represent tissue effects but are also measuring dye effects. The A_i, G_g, and $(VG)_{kg}$ terms are interpreted as in Model (1). Dye effects D_j cannot be explicitly included because they are completely confounded with variety effects V_k. It is possible to extend Model (3) to include *AG* effects, but this would leave no degrees of freedom to estimate error and thus it would not be possible to assess the significance of any effects or produce confidence intervals for estimated changes in expression. The limitations of the design force us to exclude at least one set of effects to be able to estimate error. The array×gene interactions were the smallest effects in our analysis of the Latin square experiment. However, because these are partially confounded with variety×gene effects, excluding them leads to potentially biased estimates of *VG* effects.

We fit Model (3) to the data by least squares, subject to the constraints $\sum A_i = \sum G_g = \sum_g (VG)_{kg} = V_1 + V_2 + 2V_3 = (VG)_{1g} + (VG)_{2g} + 2(VG)_{3g} = 0$. Because genes are balanced with respect to all other factors, the estimates of gene and variety×gene effects are uncorrelated with the other effect estimates in Model (3). We can partition the total sums of squares into four sources of variation, as in Table 5.

The estimated gene effects for the reference sample experiment are shown in Fig. 1c. The range and shape of the distribution are almost identical to the Latin square experiment. The distribution of estimated differences in the variety×gene effects is shown in Fig. 1d. The distribution is centered around zero with a mild left skew, but is somewhat tighter than the distribution obtained from the Latin square experiment. Long tails on this distribution indicate differentially expressed genes in the liver and muscle samples. Under Model (3), these estimates are given by

$$(\widehat{VG})_{1g_0} - (\widehat{VG})_{2g_0} = \log\left(\frac{y_{121g_0}}{y_{222g_0}}\right) - \frac{1}{N}\log\left(\prod_g \frac{y_{121g}}{y_{222g}}\right). \tag{4}$$

For a gene g, the estimates are based on a single pair of observations of the liver and muscle samples for the gene.

All of the measurements on liver and muscle tissue (half of the data) are fit exactly here, that is, with zero residual. This is because there is only one observation for these tissues for any given gene. A normal quantile plot of the nonzero residuals (from the reference sample) is shown in Fig. 2b. Again, we see that the residual distribution is heavy-tailed. To look for trends in the residuals contrary to our modeling assumptions, we plotted the absolute value of each residual against the corresponding fitted value and fit a local linear regression smooth (Fig. 3b). Other than the largest residuals that are more frequent for medium values, the smooth does not detect any remarkable nonuniformity. Overall, the residuals are smaller in this experiment. Even after adjusting for degrees of freedom, the estimate of error variance is half as large as for the latin square experiment. This is a property of these particular experiments, probably reflecting an overall difference in data quality. The smaller estimate of error variance is not a general property of the designs.

We again employed the bootstrap to obtain confidence intervals for the estimated differences in expression. The nonzero residuals, rescaled by $\sqrt{2N/(N-1)}$, were used in the bootstrap simulation. Fig. 4b shows the bootstrap 99% confidence intervals, based on 20,000 bootstrap simulations, for the liver–muscle differences. These intervals have width 1.62, as compared to 1.23 for normal intervals. Thus a 2.25-fold estimated difference is significant. Although the estimate of error variance is smaller for this experiment compared to the Latin square experiment, the confidence intervals for the comparisons of interest have about the same size because of the lesser efficiency of the reference design.

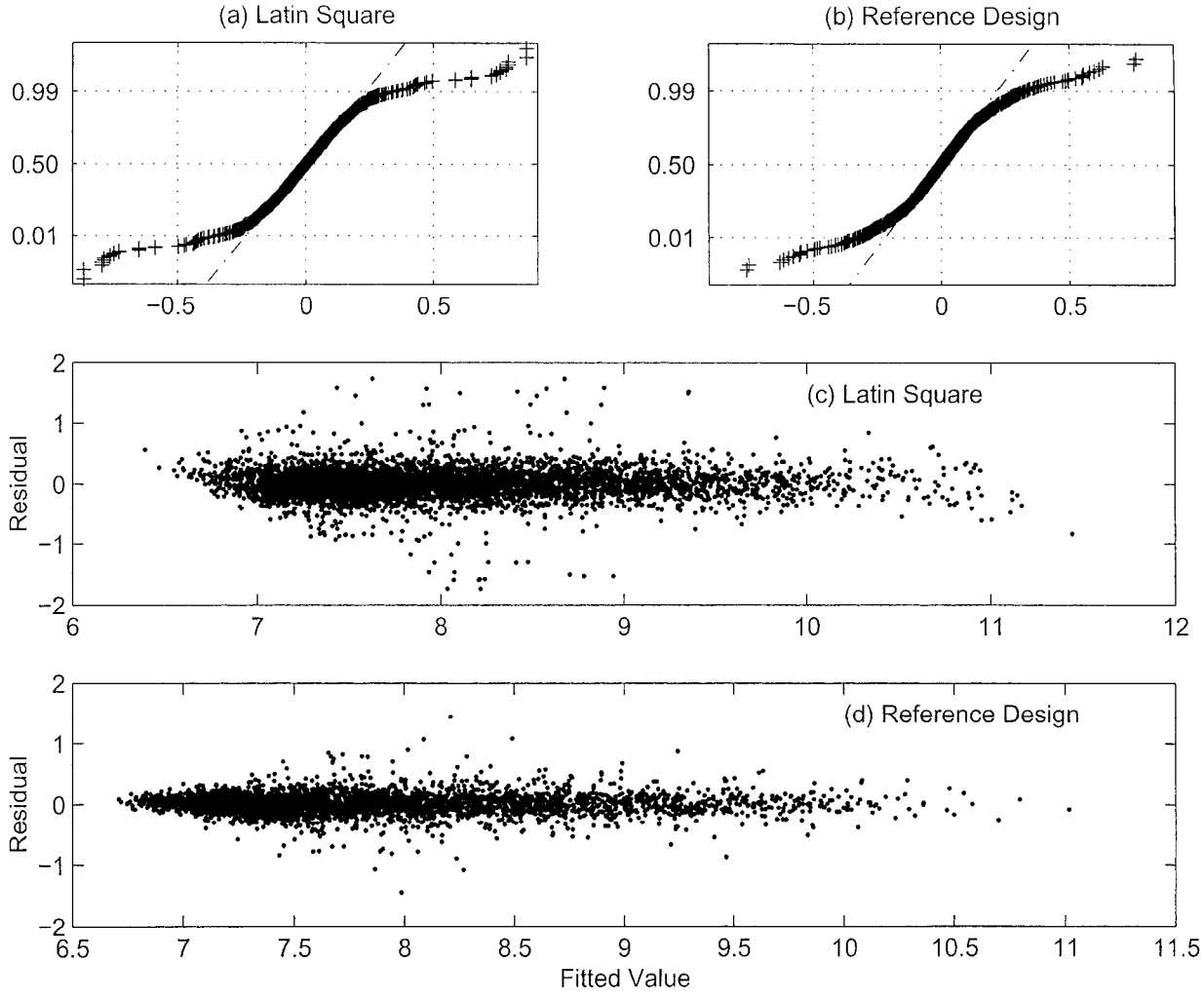

FIG. 2. Distribution of the fitted residuals. Normal quantile plots of fitted residuals are shown for the (**a**) Latin square and (**b**) reference sample experiments. The distribution of residuals is clearly heavier-tailed than normal. Scatterplots of the residuals by fitted values for the (**c**) Latin square and (**d**) reference design show no apparent trend. The residuals are re-scaled to adjust for the different degrees of freedom in the two analyses.

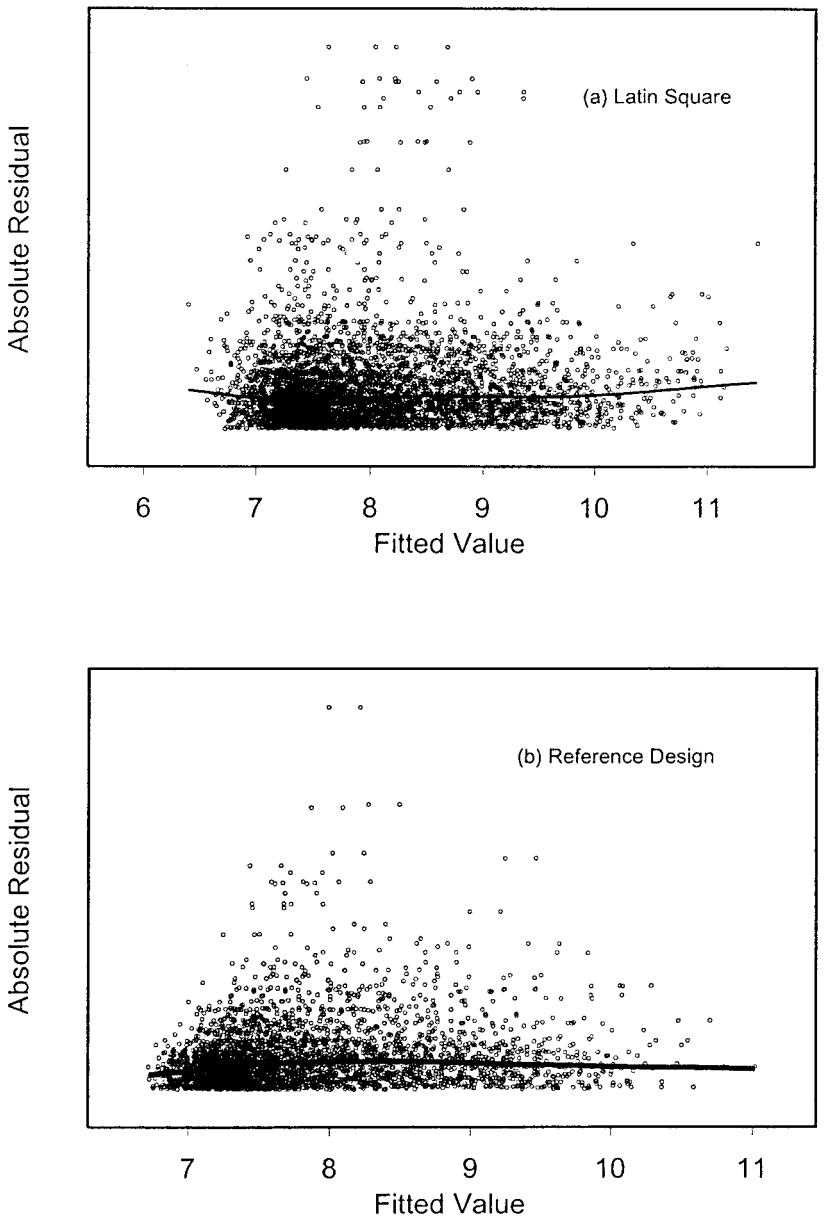

FIG. 3. Absolute value of residuals compared to fitted values. Plots (**a**) for the Latin square design and (**b**) for the reference design contain a loess smooth with span 0.35 (Hastie and Tibshirani, 1990, p. 29). In each case, the curve does not show any prominent departure from homoscedasticity.

TABLE 4. THE REFERENCE SAMPLE DESIGN

	Array	
Dye	1	2
Red	Placenta	Placenta
Green	Liver	Muscle

TABLE 5. ANALYSIS OF VARIANCE FOR THE REFERENCE DESIGN[a]

Source	df	SS	MS
Array, Variety	3	761.97	253.99
Gene	1904	3394.17	1.78
Gene × Variety	3808	1264.43	0.33
Residual	1904	55.21	0.0290
Corrected Total	7619	5475.78	

[a]The correlation coefficient of the fitted model is $R^2 = 0.990$. Abbreviations are as in Table 3.

We were concerned about possible bias in our estimates due to the omission of *AG* effects in Model (3). To examine the possible bias, we fit an extended version of Model (3), including array × gene effects.

$$\log(y_{ijkg}) = \mu + A_i + V_k + G_g + (VG)_{kg} + (AG)_{ig} + \epsilon_{ijkg}. \tag{5}$$

Although we could not evaluate the fit of this model because there are no residual degrees of freedom, we can compare the estimates ($\widehat{VG}$) from (5) with those from (3) to evaluate the extent to which the latter

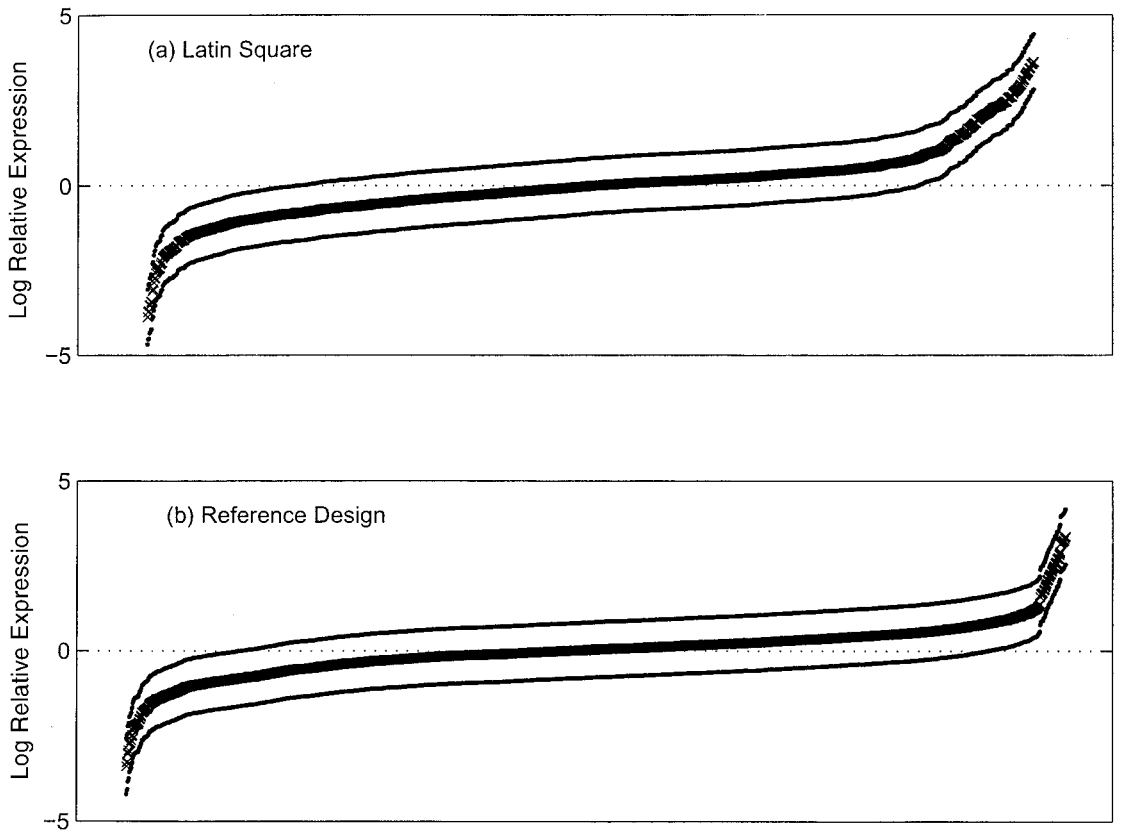

FIG. 4. Bootstrap confidence intervals. Estimated differences (liver − muscle) of the variety × gene interactions are shown for (**a**) the 1286 genes in the Latin square experiment fitting Model (1) and (**b**) the 1905 genes in reference sample experiment fitting Model (3). The estimates are plotted in increasing order along with their 99% bootstrap confidence limits. There is an optical illusion that the confidence intervals shrink at the ends because the lines are steeper, but the vertical distance between the upper and lower confidence bounds is constant in each plot.

might be biased. The estimates from Model (5) are different because they account for *AG* effects. The expression is more complicated because of the confounding structure in the design:

$$(\widehat{VG})_{1g_0} - (\widehat{VG})_{2g_0} = 2\left[\log\left(\frac{y_{121g_0}}{y_{222g_0}}\right) - \frac{1}{N}\log\left(\prod_g \frac{y_{121g}}{y_{222g}}\right)\right] - \left[\log\left(\frac{y_{113g_0}y_{121g_0}}{y_{213g_0}y_{222g_0}}\right) - \frac{1}{N}\log\left(\prod_g \frac{y_{113g}y_{121g}}{y_{213g}y_{222g}}\right)\right] \quad (6)$$

Notice that observations from variety 3 do not come into play at all in (4) because that estimator comes from a model that assumes no "spot" effects. Observations from variety 3 appear in (6) because this estimator corrects for spot-to-spot variation.

Figure 5 compares the differences of interest, $(\widehat{VG})_{1g} - (\widehat{VG})_{2g}$, for the Models (3) and (5), i.e., compares the estimators in Equations (4) and (6). There is a fairly substantial difference in the estimates for a handful of genes, suggesting there is some spot-to-spot variation biasing our estimates from (3). Fortunately, there appears to have been enough uniformity among spots from array to array so that most estimates of *VG* effects from fitting Model (3) are not too severely biased.

Comparison of the experiments

We used two approaches to compare the results from these two experiments. The first approach takes advantage of spots for which the same gene is represented by different clones on a single array. The second uses a subset of the clones that were common across all four arrays.

For those genes that are represented by two spots on an array, we were not able to determine whether the same clones were used and thus we treated these spots as distinct genes when fitting the ANOVA models. However, it seems desirable that a gene duplicated within an experiment should produce similar results. In particular, one would like the two confidence intervals for liver–muscle differences to either both contain 0 or both not contain 0. To study this question, we separately produced 1,000 bootstrap datasets for each experiment and recorded $(\widehat{VG})^*_{1g} - (\widehat{VG})^*_{2g}$ for the duplicated genes. Let g and g' be indices for spots that are the same gene. For each bootstrap dataset, we plotted the estimate of $(VG)_{1g} - (VG)_{2g}$ against the estimate of $(VG)_{1g'} - (VG)_{2g'}$. Fig. 6 presents the results. The first eleven plots (reading left to right, top to bottom) are for the genes that were duplicated on arrays in both experiments. Gene 931 was duplicated only in the Latin square experiment. The remaining eight genes were duplicated only in the reference design experiment. The clouds of symbols generally fall along the line of identity indicating that the two estimates from within an experiment are close to one another. However, comparison of the genes common to both experiments (the first 11 subplots) shows that different conclusions were obtained for three genes (116, 256, and 840) in the two experiments.

An alternative approach to assessing reproducibility that is not subject to doubts raised by nonidentity of clones is to compare 1,177 genes common to both experiments. For each experiment, we categorized genes into groups for which expression was higher in liver, not significantly different, or higher in muscle as determined by the bootstrap 99% confidence intervals. Table 6 shows the cross-tabulation for the two experiments. The two analyses agree on 88% of the 1,177 genes. The largest source of disagreement is genes for which the Latin square confidence interval contains 0 while the reference design interval does not. In general, when *AG* effects are accounted for, the confidence intervals for the reference design should be larger than those for the Latin square design by a factor of $\sqrt{3}$. However, we were not able to account for *AG* effects and estimate error in the same analysis of the reference design experiment. Without accounting for *AG* effects, confidence intervals for the reference design should still be larger by a factor of $\sqrt{2}$. However, in these particular experiments, the reference design yielded a smaller estimate of residual error, perhaps reflecting higher overall data quality, resulting in confidence intervals of about the same size for the two experiments.

Finally, we generated scatterplots of the estimated differences $(\widehat{VG})_{1g} - (\widehat{VG})_{2g}$ for the 1,177 common genes. On the whole, there is remarkable agreement between the two experiments. Fig. 7a shows the estimates for the reference design experiment using Model (3) and Fig. 7b shows estimates from Model (5). Each plot contains an orthogonal regression line (Casella and Berger, 1991, p. 581). In both plots,

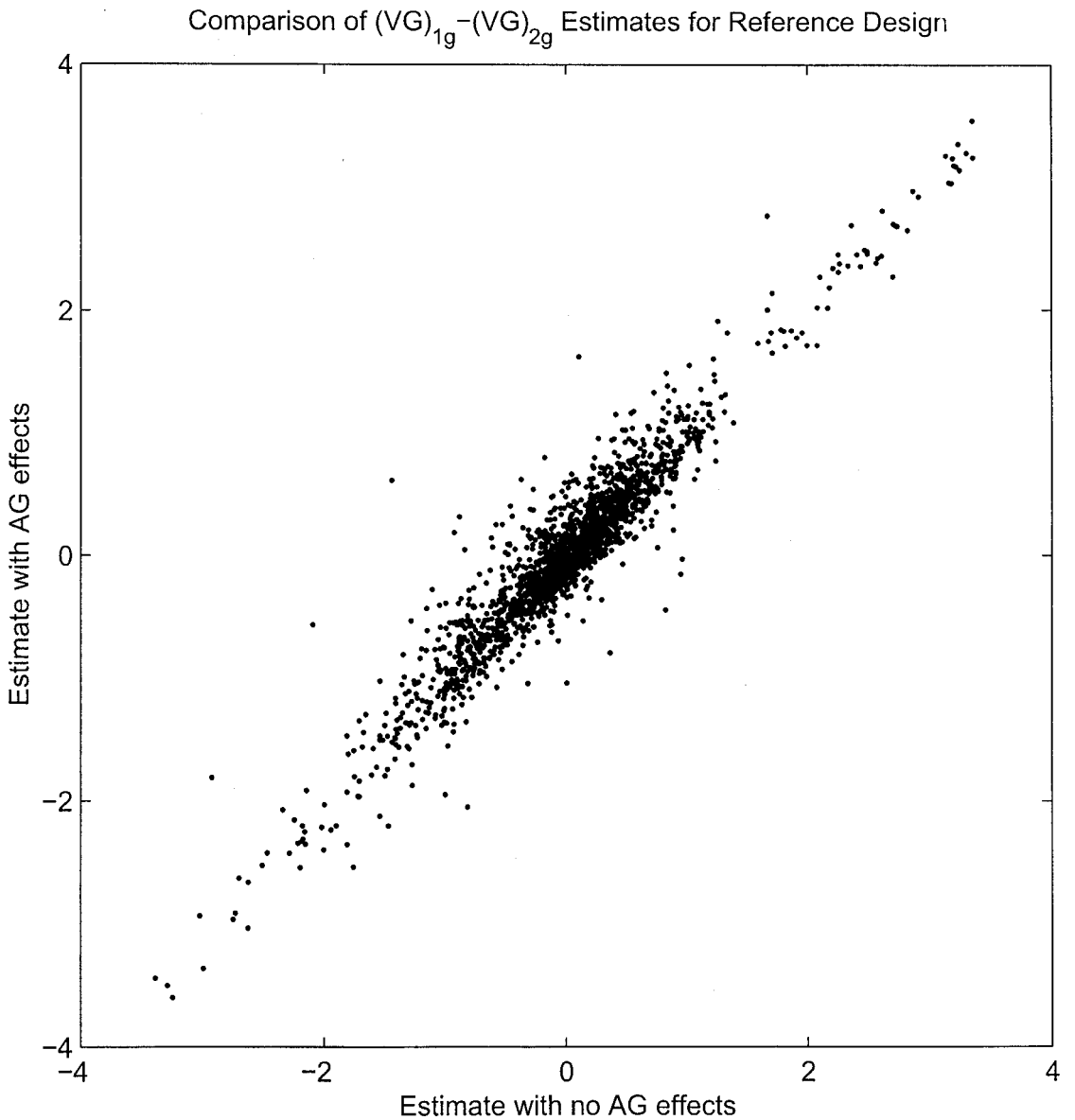

FIG. 5. Difference in estimated log fold change for the reference design when array×gene effects are taken into account. Comparison of estimated differences $(\widehat{VG})_{1g} - (\widehat{VG})_{2g}$ with estimator (4), which does not account for *AG* effects, and with estimator (6), which does account for these effects. The plot summarizes the magnitude of bias in estimates from (4) due to excluding *AG* effects. There is little change for most genes but notable change for a handful of genes.

the regression line has slope close to 1 and intercept close to 0. Agreement appears somewhat better for Model (5). In any case, the agreement of the independent estimates confirms our assertion that ANOVA analysis correctly normalizes microarray data and yields reproducible estimates.

DISCUSSION

A common practice with microarray data is to compute ratios of the raw signals as estimates of differential expression (Chen *et al.*, 1997). We find this approach to be inadequate for several reasons. It is natural and convenient to speak of fold change in expression, but it can also be misleading because

ratios expressing fold change in fluorescence do not necessarily correspond to fold changes in expression. Simple ratios do not necessarily account for differential behavior of dyes or variations between samples or arrays. These effects must be accounted for to obtain unbiased estimates of expression ratios. Indirect approaches to normalization require preprocessing steps, and ratios can be very sensitive to how these steps are carried out. ANOVA methods provide an automatic correction for the extraneous effects in a microarray experiment as an integral part of the data analysis.

Changes in gene expression across experimental samples are captured in the variety×gene interaction terms of the ANOVA model. In this study, we have simply looked at differences among these terms in order to test the hypothesis of differential expression for individual genes. The estimates $(\widehat{VG})_{kg}$ can be subject to alternative analyses, depending on the questions of interest. Hierarchical clustering (Eisen *et al*., 1999) and self-organizing maps (Tamayo *et al*., 1999) are two popular approaches to microarray data analysis that could be directly applied. The "normalized" estimates of differential expression obtained from ANOVA analysis should provide a more suitable and robust input for these analyses than raw ratios.

The properties of ANOVA estimates are tied to the experimental design. In practice, full factorial designs are impossible for microarrays because only one sample can correspond to each array–dye combination. However it is possible to derive efficient designs that satisfy the constraints imposed by this technology (John and Mitchell, 1977; Cheng, 1978). In general, for a given number of arrays, designs that are balanced across the samples of interest will provide the greatest efficiency. In our studies, using two arrays each, we prefer the Latin square design to the reference sample design. The Latin square design produces more data on the varieties of interest and allows more degrees of freedom for estimating error variance.

It is common practice in applied statistics to seek a transformation of the raw data to obtain normal residuals with constant variance (Draper and Smith, 1998). In this study we have applied a logarithm transformation to the fluorescent intensities. The residual distribution on the log scale is nonnormal, but we did not detect any dramatic evidence against our assumption of constant error variance. The ease of interpretation provided by the logarithmic transformation gives it a unique advantage over all other transformations. Biologists are quite accustomed to interpreting ratios and "fold change" for a good reason. Many natural phenomena occur on multiplicative scales; i.e., a system is more likely to "double" in response to a change of conditions than to shift by an additive constant amount. Nonnormality is a problem only in so far as it complicates the data analysis and results in inefficient estimators. In this study, we have used a bootstrap approach to obtain confidence intervals without relying on normality assumptions. Other approaches to obtain confidence intervals could be considered. The model fit and parameter estimates in our study were obtained by the method of least squares, which is most efficient for normal data. Alternative methods, such as minimum absolute deviation, can improve the efficiency of estimators for nonnormal data. Finally, we wish to note that, when large numbers of similar quantities are being estimated, the estimates of the highest and lowest effects will tend to be too extreme. This can be addressed by treating the gene and variety×gene terms as random effects in the ANOVA model (Robinson, 1991). This approach leads to "shrinkage" estimators for these terms (Newton *et al*., 2000). We view these problems as areas that are ripe for further investigation in the context of the analysis of well-designed microarray experiments.

METHODS

Tissue acquisition

Human liver and skeletal muscle samples from a 24-year-old male donor and placenta from a 26-year-old female donor were obtained from the BioChain Institute, Inc. (*www.biochain.com*). These tissues were collected expressly for mRNA isolation and were quick frozen within minutes of biopsy.

Probe preparation

Total RNA was isolated using a guanidine thiocyantate solution. For mRNA preparation, polyadenylated mRNA was isolated using oligo-dT cellulose. Fluorescently labeled cDNA was prepared from 3 ug mRNA by oligo dT-primed (21-mer) polymerization using SuperScriptII reverse transcriptase (LTI Inc.) and 0.5 mM dGTP, dATP, dTTP and 0.2 mM dCTP. Fluorescent nucleotides Cy3-dCTP ro Cy5-dCTP (Amersham) were present at 0.1 mM. Residual RNA was degraded by NaOH, neutralized and

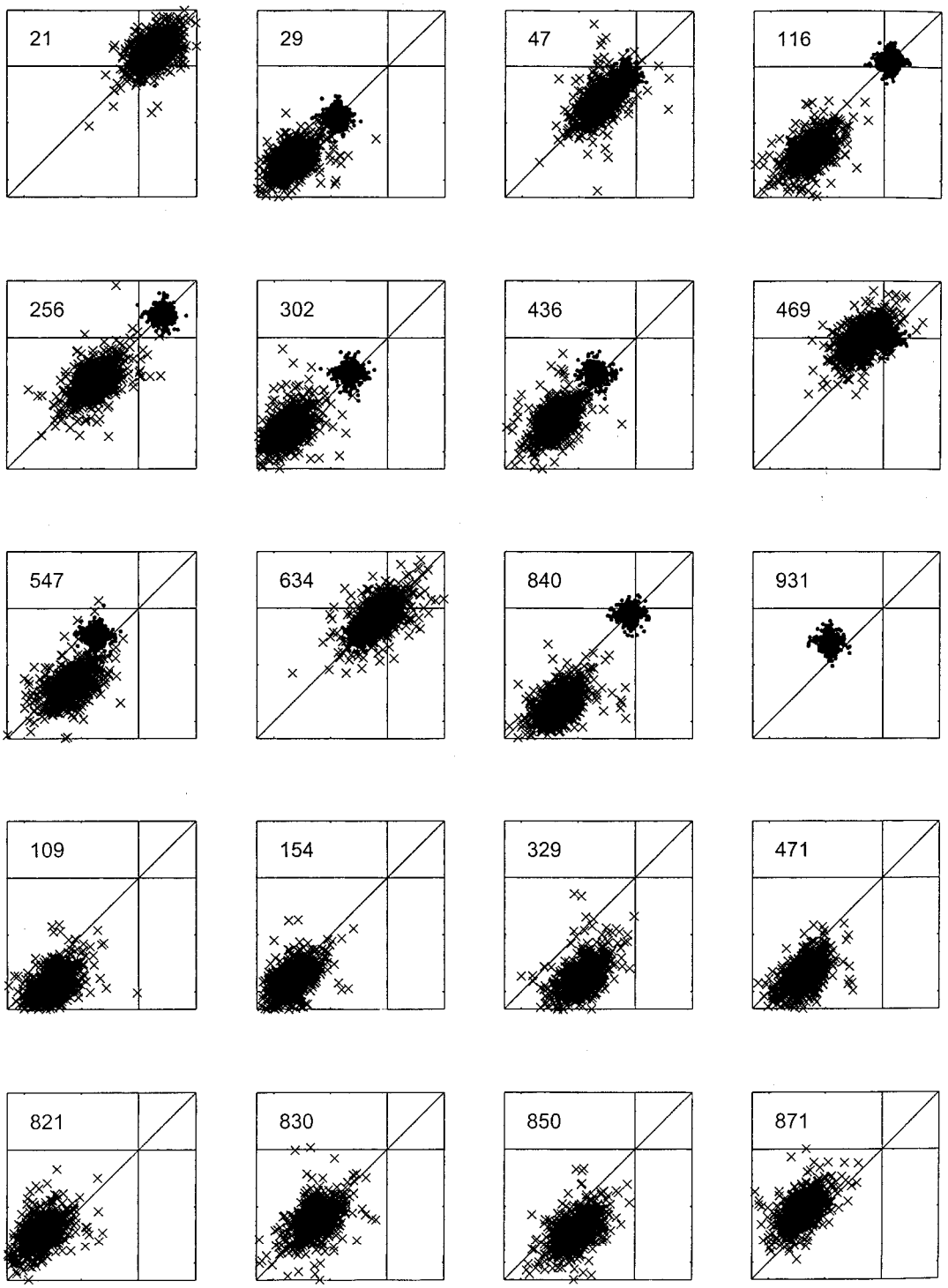

FIG. 6. Comparison of genes duplicated within an array. Bootstrap samples of estimated differences (liver − muscle) of the variety × gene interactions for the 20 genes that are duplicated in one or both experiments are shown. Each subplot corresponds to one of the duplicated genes, with the gene identifier shown in the upper left corner. Each point represents the estimated difference $(VG)_{1g} - (VG)_{2g}$ obtained for the two spots from one of 1000 bootstrap datasets. The Latin square estimates are indicated by dots and the reference sample estimates are indicated by crosses. The clouds of points generally fall along the line of identity, indicating that pairs of estimates from within an experiment are close to one another. There are eleven plots (containing both dots and crosses) for genes that were duplicated in both experiments. Disagreement between the experiments is noted for genes 116, 256, and 840.

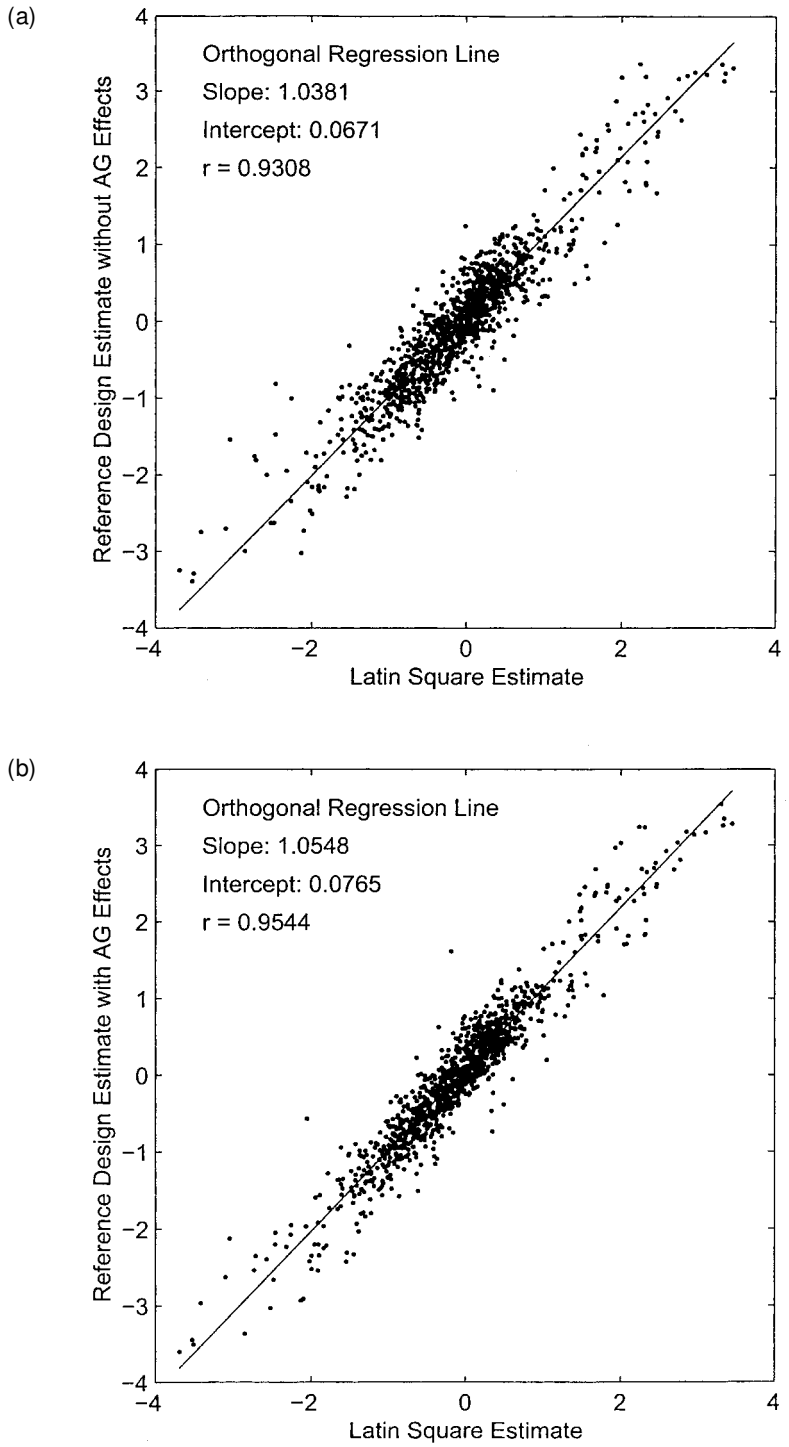

FIG. 7. Comparison of estimates for genes duplicated across experiments. A scatterplot of the Latin square and reference sample estimates of log fold-change for the 1177 genes common to the two experiments are shown. In (**a**) the estimates for the reference design come from fitting Model (3); in (**b**) estimates for the reference design come from Model (5), which includes *AG* effects. The orthogonal least squares regression line in both plots is essentially the line of identity. The high correlation confirms that the ANOVA results are reproducible and the near identity relationship demonstrates that the methodology properly normalizes the effect estimates.

TABLE 6. CONCORDANCE OF THE LIVER–MUSCLE DIFFERENCES, BY GENE, FOR THE 1177 GENES IN COMMON TO THE LATIN SQUARE AND REFERENCE DESIGN ANALYSES[a]

	Reference design			
Latin square	Liver<Muscle	Liver=Muscle	Liver>Muscle	
Liver<Muscle	164	37	0	17.1%
Liver=Muscle	49	780	43	74.1%
Liver>Muscle	0	17	87	8.8%
	18.1%	70.9%	11.1%	1177

[a]The genes are binned depending on whether the bootstrap 99% confidence intervals contain zero or do not. In the latter case, we conclude that there is significantly greater expression in either liver or muscle. The analyses agree that 780 of the genes do not have differential expression. There are no cases in which the experiments found differential expression in opposite directions. Overall, the experiments agree on 87.6% of the genes.

precipitated in ethanol. Washed pellets from 3 ug mRNA were suspended in 5 ml hybridization buffer (5X SSC, 0.2%SDS).

Hybridization and scanning

Labeled probe mixtures were aliquoted onto the cDNA microarray surface under a coverslip and incubated for 6–12 hours at 60 C in a hybridization chamber. Following washes, the arrays were scanned in 0.1X SSC using a fluorescence laser scanning device (D. Shalon, S.J. Smith, and P.O. Brown (1996) Genome Research 6, 639-645). A separate scan, at the appropriate excitation wavelength, was done for each fluorophore. Differential expression measurements were obtained by taking the average of the ratios of two independent hybridizations.

Data preparation

Data were prescreened for quality using Synteni "Gem Tools" software. We did not have access to raw images and thus excluded all data points marked by the software as unreliable. The data for the first experiment is comprised of red and green fluorescence readings for 1,556 spots on array 1 representing 1,540 different genes and 1,455 spots on array 2 representing 1,442 different genes. Spots that are indicated as representing the same gene may not contain the same clones. For each array, gene-identifiers were recorded to clone identifiers so that each dataset contained as many distinct clone identifiers as spots. This maintained a balanced design and also allowed an appraisal of the methodology. For analysis, a combined dataset was created containing readings for clone identifiers appearing for both array 1 and array 2. The final dataset had 1,286 clone identifiers representing 1,274 different genes.

The data for the second experiment is comprised of red and green fluorescence readings for 2,125 spots on array 1 representing 2,103 different genes and 2,098 spots on array 2 representing 2,078 different genes. As before, we assigned unique clone identifiers to different spots and created a combined dataset containing clone identifiers appearing on both arrays. The final data set had 1,905 clone identifiers representing 1,886 different genes.

Data analysis

All computations for the data analysis were carried using Matlab software (Mathworks Inc., Natick, MA). Data and routines are available at *www.jax.org/research/churchill/.*

APPENDIX: DERIVING LEAST-SQUARES ESTIMATORS

Generally, to fit a linear model it is not necessary to derive the functional form of least-squares parameter estimates because the estimates can be calculated by constructing the *design matrix X*, which depends on

the model and the experimental design (Draper and Smith, 1998). To fit the model, one inverts the $p \times p$ matrix $X^t X$, where p is the number of parameters in the model. In our case, p is very large because thousands of genes are spotted on microarrays and our models have G, VG, and AG effects for every gene. This makes inverting $X^t X$ computationally infeasible for general matrix inversion programs. To get around the hurdle, we derived the functional form of the parameter estimators.

The name "least-squares" comes from the fact that the estimates minimize the residual sum of squares RSS, the total squared difference between all data points and the estimated value under the fitted model. Let $t_{ijkg} = \log(y_{ijkg})$ be the log transformed data. For example, considering Model (1), $RSS = \sum_{ijkg} (t_{ijkg} - \mu - A_i - D_j - V_k - G_g - (AG)_{ig} - (VG)_{kg})^2$. The summation is over all combinations of indices i, j, k, and g that appear in the design. Estimators are derived by taking partial derivatives of RSS with respect to the parameters and setting them equal to zero. The result is a set of linear equations that can be solved for the least-squares estimates.

For example, taking partial derivatives with respect to the parameters of interest, VG, in Model (1) yields

$$\frac{\delta RSS}{\delta VG_{kg}} \propto \sum_{ij} (t_{ijkg} - \mu - A_i - D_j - V_k - G_g - (AG)_{ig} - (VG)_{kg}).$$

Note k and g are fixed, so the sum is over all pairs i, j such that i, j, k, g is a set of indices in the design. For the Latin square design and for any fixed k and g, i ranges over all arrays, so the constraints $\sum_i A_i = \sum_i (AG)_{ig} = 0$ cause the A and AG terms to drop out of this expression (similarly for the D_j terms). The resulting equation simplifies to $\sum_{ij} (t_{ijkg} - \mu - V_k - G_g - (VG)_{kg}) = 0$. Taking partial derivatives with respect to μ, V_k, and G_g yields similar equations whose simultaneous solution gives

$$\widehat{(VG)}_{kg} = t_{..kg} - t_{..k.} - t_{...g} + t_{....},$$

where a "$\cdot$" as an index means to average over that index. This expression for $\widehat{(VG)}$ does not depend on whether AG effects are included in the model because of the special orthogonality properties of the Latin square design.

In contrast, consider the reference design. For the smaller Model (3), the form of $\widehat{(VG)}$ is the same as above. Model (5) includes AG effects, which are partially confounded with VG effects in the reference design. The reference variety is balanced across arrays but variety 1 is only on array 1 and variety 2 is only on array 2. Calculating $\frac{\delta RSS}{\delta VG_{kg}}$ for $k = 1, 2$, none of the other effects in (5) drops out. In the end, one finds that

$$\widehat{(VG)}_{3g} = t_{.13g} - t_{.13.} - t_{...g} + t_{....},$$

while for $k = 1, 2$,

$$\widehat{(VG)}_{kg} = 2(t_{k2kg} - t_{k2k.} - t_{k..g} + t_{k...}) + t_{.13g} - t_{.13.} - t_{...g} + t_{....},$$

and the estimator (6) follows by taking differences.

REFERENCES

Brown, P.O., and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21(1 Suppl), 33–37.

Casella, G., and Berger, R.L. 1991. *Statistical Inference*, Duxbury Press, New York.

Chen, Y., Dougherty, E.R., and Bittner, M.L. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2, 364–374.

Cheng, C.S. 1978. Optimality of certain asymmetrical experimental designs. *Ann. Statist.* 6, 1239–1161.

Cochran, W.G., and Cox, G.M. 1992. *Experimental Designs*, Wiley, New York.

Draper, N.R., and Smith, H. 1998. *Applied Regression Analysis*, Wiley, New York.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nat. Genet.* 21(1 Suppl), 10–14.

Efron, B., and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1, 54–77.

Eisen, M. 1999. Cluster and Tree View Manual. (unpublished).

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA* 95, 14863–14868.

Fisher, R.A. 1951. *The Design of Experiments*, Sixth ed., Oliver and Boyd, London.

Hastie, T.J., and Tibshirani, R.J. 1990. *Generalized Additive Models*, Chapman and Hall, London.

John, J.A., and Mitchell T.J. 1977. Optimal incomplete block designs. *J. Royal Stat. Soc. Series B* 39, 39–43.

Manly, B.F.J. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, Chapman and Hall, London.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. 2000. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol*. (submitted).

Robinson, G.K. 1991. That BLUP is a good thing: The estimation of random effects. *Stat. Sci.* 6, 15–51.

Sapir, M., and Churchill, G.A. 2000. Estimating the Posterior Probability of Gene Expression from Microarray Data. (unpublished).

Still, A.W., and White, A.P. 1981. The approximate randomization test as an alternative to the F test in analysis of variance. *Brit. J. Math. Stat. Psychol.* 34, 243–252.

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci. USA* 96, 2907–2912.

Wu, C.F.J. 1986. Jackknife, Bootstrap, and other resampling methods in regression analysis. *Ann. Stat.* 14, 1261–1295.

Address correspondence to:
Gary A. Churchill
The Jackson Laboratory
600 Main St.
Bar Harbor, ME 04609

E-mail: garyc@jax.org