Statistica Sinica **12**(2002), 203-217

STATISTICAL ANALYSIS OF A GENE EXPRESSION MICROARRAY EXPERIMENT WITH REPLICATION

M. Kathleen Kerr[1], Cynthia A. Afshari[2], Lee Bennett[2],
Pierre Bushel[2], Jeanelle Martinez[2], Nigel J. Walker[2] and Gary A. Churchill[3]

[1] *University of Washington,*
[2] *National Institute of Environmental Health Sciences and*
[3] *The Jackson Laboratory*

Abstract: Common ratio-based approaches for analyzing gene expression microarray data do not provide a framework for handling replication, although replication is clearly desirable for these noisy data. In contrast, replication fits naturally into analysis of variance (ANOVA) methods. We use ANOVA to analyze data from a microarray experiment to compare gene expression in drug-treated and control cells lines. We discuss issues that commonly arise in the analysis of microarray data, and present practical solutions to some common problems.

Key words and phrases: Analysis of variance, bootstrap, cDNA microarray, gene expression, orthogonal design.

1. Introduction

This paper describes an analysis of a cDNA microarray experiment to compare gene expression in treated and control human cell line samples. We present the analysis as a case-study in handling important issues that arise with microarray data (Kerr and Churchill (2001)). As noted by Lee, Kuo, Whitmore and Sklar (2000) and others, replication is crucial to microarray studies because there is inherent noise in the data, even after systematic sources of variation are removed. However, simple ratio-based approaches do not provide a framework for analyzing data with replicate measurements. Our primary tool for studying microarray data is the analysis of variance (ANOVA). The data analysis illustrates how ANOVA naturally handles experiments that incorporate replication. We will describe issues that arise in the analysis of microarray data concerning model selection, data scaling, and statistical inference. Our experience demonstrates that some relatively minor modifications to the modeling and bootstrapping used in Kerr, Martin and Churchill (2000) are necessary, useful, and valuable additions to the statistical analysis of microarrays.

Spotted cDNA microarrays are a tool for high-throughput analysis of gene expression (Brown and Botstein (1999)). In the first step of the technique, DNA is "spotted" and immobilized on glass slides or other substrate, the microarrays. Each spot on an array contains a particular sequence, although a sequence may be spotted multiple times per array. Next, mRNA from cell populations under study is reverse-transcribed into cDNA and one of two fluorescent dye labels, Cy3 and Cy5, is incorporated. Two pools of differently-labeled cDNA are mixed and washed over an array. Dye-labeled cDNA can hybridize with complementary sequences on the array, and unhybridized cDNA is washed off. The array is then scanned for Cy3 and Cy5 fluorescent intensities. The idea is that the mRNA sample that contained more transcript for a given gene should produce higher fluorescence in the corresponding label in the spot containing that gene. The experimental data consist of Cy3 and Cy5 measurements for every spot on every array.

The data studied here are from an experiment to study 2,3,7,8-tetrachlordibenzo-*p*-dioxin (TCDD) (Martinez and Walker, unpublished data). This compound is known to induce a wide range of biological and biochemical responses, including gene induction. The experiment used the human hepatoma cell line HepG2 as an *in vitro* model to study TCDD. HepG2 is an established cell line for which metabolic enzymes are known to be inducible (Kikuchi, Hossain, Yoshida and Kobayashi (1998); Li, Harper, Tang and Okey (1998)). Thus it can be considered a prototype of the TCDD response.

The experimental design included replication to control the noise that is associated with microarray data. Although each gene was spotted only once per array, replication was achieved by using six arrays to study the two samples instead of just one or two. A separate labeling reaction was performed for each hybridization. Each array was spotted with the same set of 1920 genes. Table 1 summarizes the "triple dye-swap" experimental design. As in Kerr and Churchill (2000), we refer to the TCDD-treated and control cell lines as "varieties." Control cells are variety 1 and treated cells are variety 2. We refer to the fluor Cy3 as dye 1 and Cy5 as dye 2. Ninety-eight entries in the datafile, corresponding to 13 genes, were exactly 1 and appeared to be artificial "floor" values. These genes were removed from the dataset for analysis. There was no other data pre-processing. Thus the cleaned dataset has complete data for 1907 genes.

Table 1. Experimental design: variety assignments to arrays.

	Dye 1	Dye 2
Array 1	Variety 2	Variety 1
Array 2	Variety 2	Variety 1
Array 3	Variety 1	Variety 2
Array 4	Variety 2	Variety 1
Array 5	Variety 1	Variety 2
Array 6	Variety 1	Variety 2

2. ANOVA Modeling

We analyzed the data on the log scale using ANOVA models (Kerr, Martin and Churchill (2000)). Every data value is identified by four factors, array (A), dye (D), variety (V), and gene (G). This experimental design has three- and four-way interactions of these factors confounded with main effects and two-way interactions. Therefore, considering only lower-order effects indirectly accounts for higher-order effects. Let y_{ijkg} be the log intensity reading from array $i = 1, \ldots, 6$, dye $j = 1, 2$, variety $k = 1, 2$, and gene $g = 1, \ldots, 1907$. We consider the ANOVA model

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

where A_i, D_j and V_k are "global" effects to account for overall differences in arrays, dyes, and varieties. The gene effects G_g account for the expression level of genes averaged over the other factors. The $(AG)_{ig}$ terms are the "spot" effects. The $(DG)_{jg}$ terms are gene-specific dye effects, which occur when subsets of genes exhibit higher fluorescent signal when labeled with one dye or the other, regardless of the variety. We have seen such effects repeatedly in microarray data. Finally, the variety×gene interactions $(VG)_{kg}$ capture the expression of gene g specifically attributable to variety k. The VG effects are the effects of interest for studying the relative gene expression in the two samples. The terms ϵ_{ijkg} represents independent random error with mean 0.

The global effects A, D, V do not saturate the entire "design space" spanned by arrays, dyes and varieties. We also consider an alternate model that replaces the 1 degree of freedom variety effect V_k with the 5 degree of freedom array×dye interaction $(AD)_{ij}$,

$$y_{ijkg} = \mu + A_i + D_j + (AD)_{ij} + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \epsilon_{ijkg}. \qquad (2)$$

This model saturates the design space so that every combination of arrays, dyes and varieties is directly or indirectly accounted for. In particular, variety effects are indirectly accounted for by the AD effects. Table 2 shows the analysis of variance for Models (1) and (2). The residual mean square for Model (2) is less than half that for Model (1), so Model (2) is preferred. Wolfinger et al. (2000) also use AD effects to model microarray data.

Proceeding with Model (2), we examine the fitted residuals to check our modeling assumptions. Figure 1(a) plots residuals vs. fitted values and shows some modest heteroscedasticity but no other clear pattern. This is misleading, however. Examining residuals separately for each array shows systematic patterns. Figure 1(b) shows the most dramatic example of this, which occurs for array 3. There are two discernable intersecting arcs of points in these plots, which correspond to the two dyes.

Table 2. Analysis of variance for log data.

Model (1)				Model (2)			
Source	SS	df	MS	Source	SS	df	MS
Array	328.28	5	65.66	Array	328.28	5	65.66
Dye	119.10	1	119.10	Dye	119.10	1	119.10
Variety	40.66	1	40.66	Array∗Dye	128.42	5	25.68
Gene	35285.23	1906	18.52	Gene	35285.23	1906	18.52
Spot	1671.35	9530	0.18	Spot	1671.35	9530	0.18
Variety∗Gene	230.60	1906	0.12	Variety∗Gene	230.60	1906	0.12
Dye∗Gene	316.84	1906	0.17	Dye∗Gene	316.84	1906	0.17
Residual	144.67	7628	0.0190	Residual	56.86	7624	0.0075
Adj Total	38136.69	22883		Adj Total	38136.69	22883	

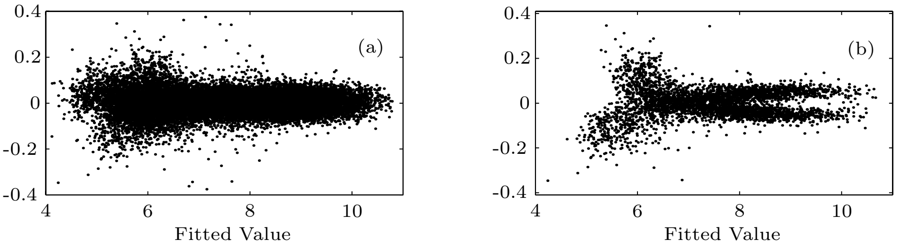

Figure 1. (a) Residual plot for the data analyzed on the log scale using Model (2). (b) The residuals for array 3 show systematic trends that are not visible in the plot of all residuals. The two discernable bands of points correspond to the two dyes.

Our first attempt to fix the analysis was to adopt the loess adjustment of Yang, Dudoit, Luu and Speed (2000). The approach is an array-by-array “normalization” method to account for non-linear effects of the dyes. For each array, one takes the two readings for each spot on the log scale and plots the differences in the two values, the log ratio R, versus the mean of the two values, the average log intensity I. In other words, $I = (y_{i1kg} + y_{i2kg})/2$ and $R = y_{i1kg} - y_{i2kg}$. Under the assumption that most genes are not differentially expressed, one expects this plot to be a horizontal band centered around 0. Instead, we see curvature (Figure 2(a)). The Yang et al. technique is to fit a loess curve through these plots and then re-define this curve to be 0. This forces the plots to straighten and center around 0 (Figure 2(b)). For any given gene, the loess adjustment does not change the intensity value I but produces an adjusted ratio value R'.

Yang et al. (2000) use the loess technique to adjust the log-ratios R. We adapt the technique to adjust the log signal y_{ijkg}. After obtaining the R' value for each gene, we use the mean log intensity value I and solve the equations

$(y'_{i1kg} + y'_{i2kg})/2 = I$, $y'_{i1kg} - y'_{i2kg} = R'$ to get adjusted log signal intensities y'_{i1kg} and y'_{i2kg}. We perform the loess in S-Plus using the default settings for the loess span parameters.

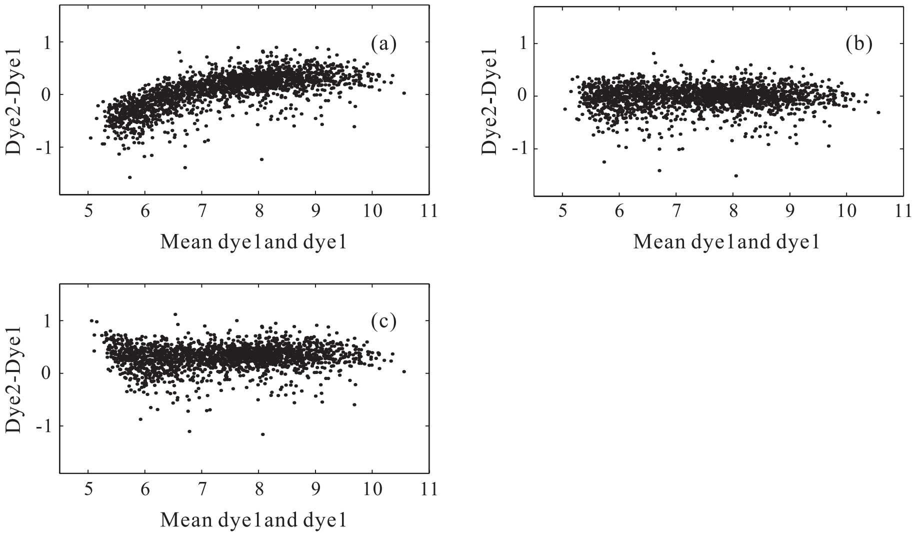

Figure 2. R vs. I plots for (a) the unadjusted log data, (b) the loess-transformed data, and (c) the shift-log data, all representing array 1.

Table 3. Analysis of variance for transformed data.

	(a) Loess			(b) Shift-Log		
Source	SS	df	MS	SS	df	MS
Array	328.28	5	65.66	337.07	5	67.41
Dye	0.02	1	0.02	4.47	1	4.47
Array∗Dye	2.34	5	0.47	245.62	5	49.12
Gene	35285.23	1906	18.52	34712.49	1906	18.22
Spot	1671.35	9530	0.18	1657.78	9530	0.17
Variety*Gene	221.28	1906	0.12	222.62	1906	0.12
Dye*Gene	61.52	1906	0.03	64.12	1906	0.03
Residual	35.96	7624	0.0047	39.62	7624	0.0052
Adj Total	37605.98	22883		37283.78	22883	

Table 3(a) gives the analysis of variance for the loess-transformed data. We see that the array×dye sum of squares is much smaller here than with the unadjusted log data. This is to be expected because of the way the loess adjustment centers the data. Figure 3 shows that the systematic trends in the residual plots have been removed. There seems to be a small degree of heteroscedasticity, especially on array 3. One possibility we had hoped for was that the loess adjustment would remove the need for DG effects in the ANOVA model. However, we see evidence for persistent dye×gene effects — the loess adjustment does not account for all the aberrant behavior of the dyes. Figure 4 shows that many genes at low levels of intensity and some genes at higher levels of intensity exhibit the largest dye interactions in the loess-adjusted data.

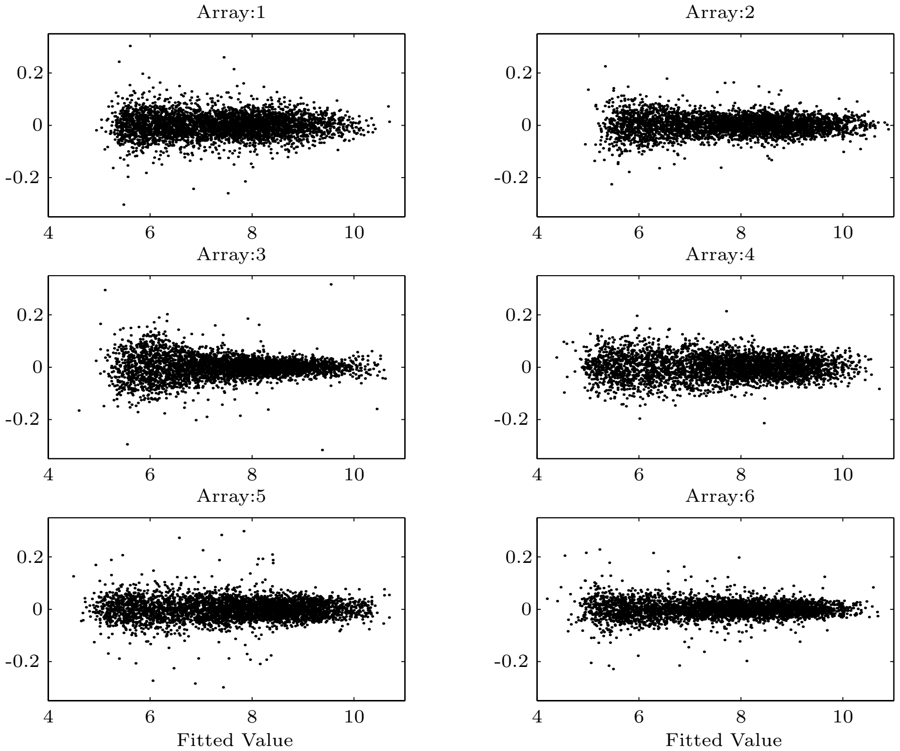

Figure 3. Residual plots for the loess-adjusted data analyzed with Model (2) show the loess adjustments remove systematic trends (compare to Figure 1(b)).

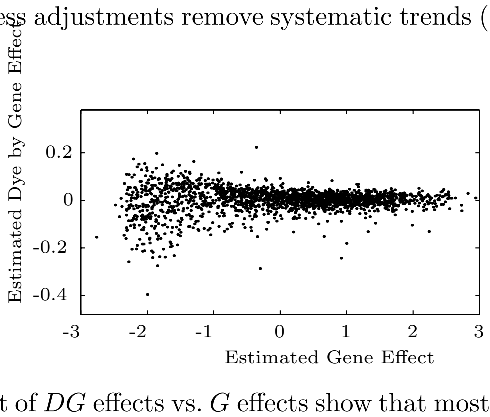

Figure 4. A plot of DG effects vs. G effects show that most, but not all, genes exhibiting the largest dye interactions have lower overall levels of expression.

Some transformation of the data, other than the simple log transform, is clearly needed in order for the ANOVA effects to be considered additive. The loess adjustment is not completely satisfactory: it is unclear how to choose the loess span parameter, a typical concern of statisticians using this kind of smoothing technique. Thus if the span is too small, there will be overfitting. In this case, differentially expressed genes will have high leverage for the local fit and their signal may be muted. On the other hand, if the span is too large the fit will be crude and the adjustment will not have the desired effect. A general concern is the large numbers of parameter used in estimating the loess curve. We decided to look more closely at other options for data transformations.

For all six arrays, the data associated with dye 2 have smaller values. Because the derivative of the log function is higher at the low end, the log transform magnifies this difference in the data for the two channels. One way to see this effect is by looking at standard deviations. Table 4 shows that the log data for dye 2 are more spread out than for dye 1. In contrast, the loess-transformed data have roughly equal standard deviations for each dye. We hypothesized that the curvature in the R vs. I plots could arise simply because of a single additive difference between the two channels on the raw (pre-log) scale. Of course, this hypothesis is not the only explanation for the observed patterns in the data. However, it suggests alternative data transformations that might replace the loess procedure.

Table 4. Data standard deviations.

	Log Data		Loess Data		Shift-Log Data	
Array	Dye 1	Dye 2	Dye 1	Dye 2	Dye 1	Dye 2
1	1.05	1.28	1.17	1.16	1.16	1.17
2	1.16	1.42	1.29	1.29	1.27	1.28
3	1.07	1.34	1.21	1.20	1.18	1.20
4	1.23	1.42	1.33	1.33	1.31	1.32
5	1.29	1.40	1.34	1.34	1.33	1.33
6	1.26	1.39	1.33	1.32	1.31	1.32

Assuming there is a single additive difference in the dyes on the raw data scale, the next question is how to estimate and correct for the difference. We do not want to simply align the data to have the same mean or median on the pre-log scale because then medium and large data values would have high influence on the transformation, whereas the effect of an additive shift is most pronounced for low data values. Therefore, we adopt the following procedure.

The goal is to estimate a shift for each array to be applied to the raw data so that, after taking logs, the R vs. I plots are as close as possible to a line with 0 slope. Taking logs reduces the influence of large data values. Let s_i be the shift for array i. For each i, estimate s_i to minimize the sum of the absolute deviations from the median of $\log(x_{i1k_1g} - s_i) - \log(x_{i2k_2g} + s_i)$, where $x_{ijkg} = \exp(y_{ijkg})$ is the data on the pre-log scale and $g = 1, \ldots, 1907$. We choose the absolute deviation criterion instead of a least-squares criterion to make the procedure robust against the influence of differentially expressed genes. The fitted values of the s_i are 101.0, 133.6, 120.5, 66.7, 38.6 and 39.5 for arrays $i = 1, \ldots, 6$. Although the shifts are rather small compared to the range of the data, applying the shifts before taking logs straightens the resulting R vs. I plots (Figure 2(c)). Table 4(c) shows that the standard deviations of the "shift-log" data are very close to the loess-transformed data. Table 3(b) gives the analysis of variance for the shift-log data. This ANOVA is remarkably similar to the ANOVA for the loess-transformed data in Table 3(a). The exceptions are the dye and array×dye effects, which the loess transform necessarily reduces to near zero.

As with the loess-transformed data, the residual plots for the ANOVA modeling of the shift-log data exhibit modest heteroscedasticity. The next question is how to proceed from modeling to inference. That is, given the variability in the data, how do we infer differential expression?

3. Bootstrap Confidence Intervals for Relative Expression

For the remainder of the paper, y_{ijkg} refers to the shift-log data (i.e., the ANOVA in Table 3(b)). The next step is to estimate confidence intervals for the difference in expression $VG_{2g} - VG_{1g}$. The bootstrapping technique (Efron and Tibshirani (1986)) used in Kerr, Martin and Churchill (2000) involves creating B simulated datasets $y^*_{ijkg} = \hat{\mu} + \hat{A}_i + \hat{D}_j + (\widehat{AD})_{ij} + \hat{G}_g + (\widehat{VG})_{kg} + (\widehat{AG})_{ig} + (\widehat{DG})_{jg} + \epsilon^*_{ijkg}$. In each simulated dataset, $\hat{\mu}, \hat{A}_i, \hat{D}_j, (\widehat{AD})_{ij}, \hat{G}_g, (\widehat{VG})_{kg}, (\widehat{AG})_{ig}$, and $(\widehat{DG})_{jg}$ are the parameter estimates from the original model estimation and ϵ^*_{ijkg} are drawn randomly and with replacement from the set of studentized residuals. We use studentized residuals so that the empirical distribution has the same variance as the corresponding theoretical distribution. Model (2) is fit to each bootstrap dataset, producing bootstrap distributions for the quantities of interest. Confidence intervals are obtained via the percentile method. This procedure yields estimated 99.9% confidence intervals $(\widehat{VG})_{2g} - (\widehat{VG})_{1g} +/- 0.116$ based on B=10,000 bootstrap datasets. This is about the same as intervals based on normal theory $(+/- 0.137)$. Since $e^{0.116} = 1.123$ this means an overall fold-change of approximately 12% appears significant. However this bootstrapping method is based on the assumption of homoscedasticity, which is contradicted by the heteroscedasticity in the residual plot, and these confidence intervals are unsatisfactory.

One alternative is to assume very generally that each gene has its own associated error distribution, that is $\epsilon_{ijkg} \sim F_g$ with mean 0. To incorporate this assumption into the bootstrapping procedure, we create a different set of simulated datasets y^*_{ijkg}, but this time ϵ^*_{ijkg} is drawn with replacement from the residuals corresponding to the observations for gene g. Obviously, this procedure produces different size confidence intervals for every gene. Although there is a fair amount of replication in the experiment, there are still only 12 residuals associated with each gene. The results are extremely narrow confidence intervals for some genes. Figure 5 shows that some very small estimated differences in expression appear significant. We found such strong confidence unconvincing. With 1907 genes and only 12 residuals per gene, we expect some genes will have small residuals by chance. This will lead to deceivingly high estimated precision for the estimates.

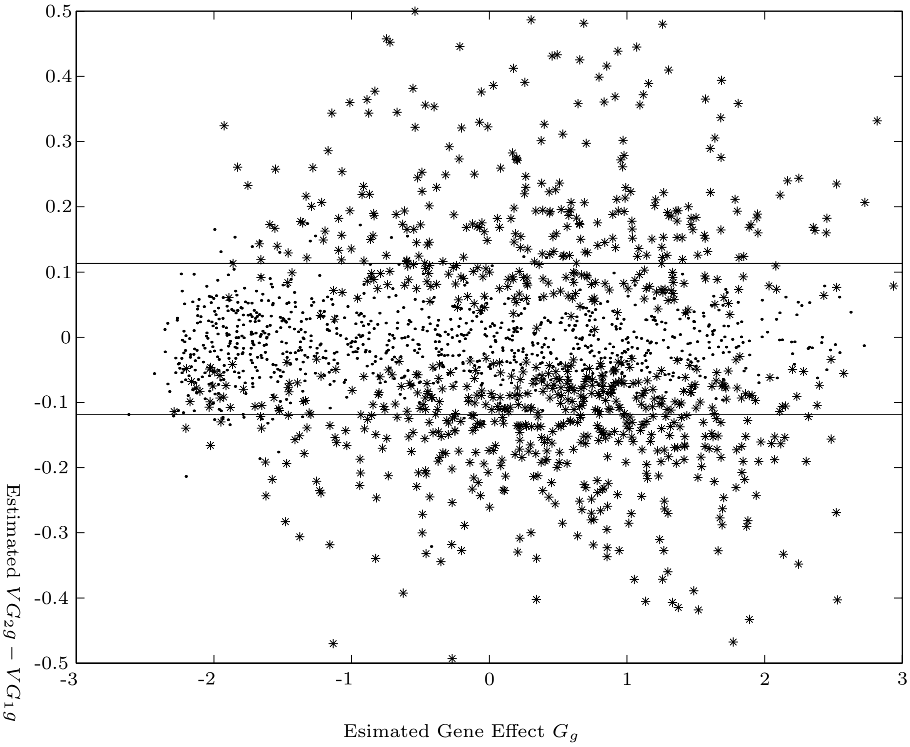

Figure 5. Differentially expressed genes according to 99.9% confidence intervals computed by assuming gene-specific error distributions. Genes whose 99.9% bootstrap confidence interval does not contain 0 are plotted with an asterisk. Horizontal lines show the cut-off obtained with the homoscedastic bootstrap. Some genes with extremely small estimated relative expression are found differentially expressed with this bootstrap procedure.

The residual plot shows that larger error is associated with smaller intensities. Rather than making the extreme assumption that each gene has its own error distribution, it seemed reasonable to assume that the magnitude of the error is intensity-dependent. By pooling the information about genes with similar average levels of expression, we get around the problem of few residuals per gene.

The procedure we adopted was to plot the standard-deviation of the 12 residuals per gene against the estimated gene effect $\hat{G}_g$. We then fit a loess curve through the plot. For each gene g we estimated the standard deviation of the residual distribution associated with gene g, denoted $\widehat{SD}_g$, to be the value of the loess curve corresponding to $\hat{G}_g$. Next, we re-scaled the studentized residuals by dividing each residual by the estimated standard deviation of the associated gene. Let $\hat{e}$ denote the rescaled residuals, so that $\hat{e}_{ijkg} = \hat{\epsilon}_{ijkg}/\widehat{SD}_g$. Since the estimated ϵ_{ijkg} are divided by their estimated standard deviation, the $\hat{e}_{ijkg}$ should have approximately unit variance. Third, we created B bootstrap datasets $y^*_{ijkg} = \hat{\mu} + \hat{A}_i + \hat{D}_j + (\widehat{AD})_{ij} + \hat{G}_g + (\widehat{VG})_{kg} + (\widehat{AG})_{ig} + (\widehat{DG})_{jg} + \widehat{SD}_g * e^*_{ijkg}.$, where e^*_{ijkg} is drawn with replacement from the $\hat{e}_{ijkg}$. This procedure accounts for the intensity-dependent heteroscedasticity while using the full set of fitted residuals to capture the uncertainty in the data.

Figure 6 shows that this procedure requires genes with lower average expression to exhibit larger differences in expression in order to be found to have significant evidence of differential expression. Table 5 compares the numbers of significant genes according to the nominal 99.9% confidence intervals found by the three bootstrap variations. The intensity-dependent bootstrap and the gene-specific bootstrap agree on 83% of genes and the former is more conservative, finding fewer genes to be differentially expressed. This is consistent with our suspicion that the gene-specific bootstrap is prone to error that leads to false-positives in finding differential expression. On the other hand, the intensity-dependent bootstrap finds more genes to be differentially expressed than the homoscedastic bootstrap; these two methods agree on 91% of genes. The gene-specific bootstrap and the homoscedastic bootstrap are the most discordant, agreeing on only 78% of genes.

Table 5. Pairwise comparisons of the three bootstrapping method. Each 2×2 table is a cross-tabulation comparing the number of genes found to be differentially expressed by the three methods. A '+' means the estimate of differential expression was statistically significant; '0' means it was not.

		Gene-specific		Homoscedastic				Homoscedastic	
		0	+	0	+			0	+
Intensity-	0	897	277	1138	36	Gene-	0	902	35
dependent	+	40	693	153	580	specific	+	389	581

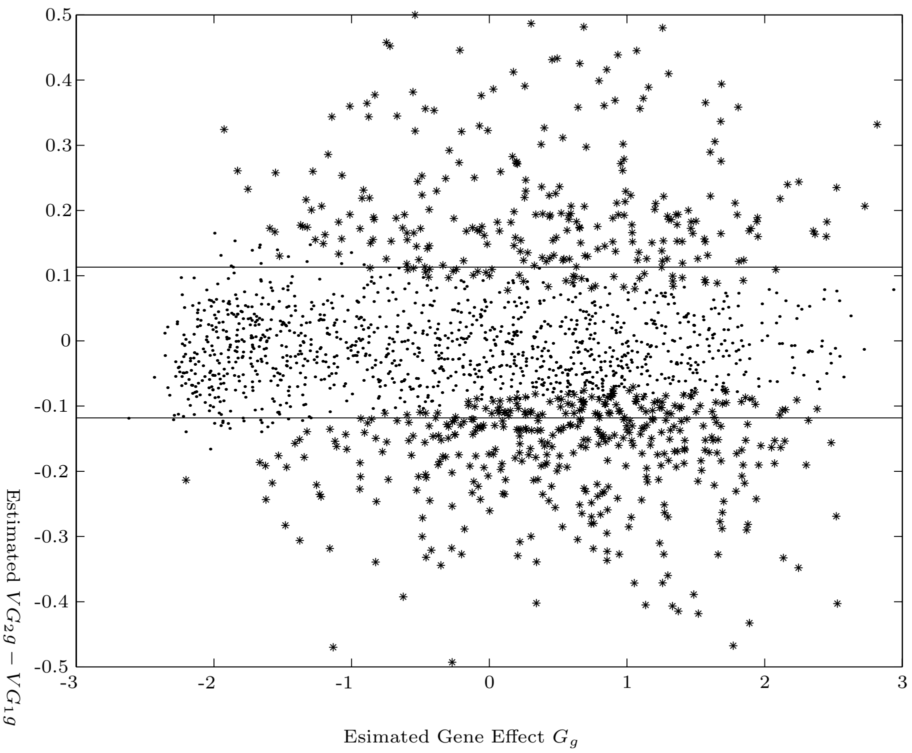

Figure 6. Differentially expressed genes according to 99.9% confidence intervals computed by assuming intensity-dependent error distributions. Genes whose 99.9% bootstrap confidence interval does not contain 0 are plotted with an asterisk. Horizontal lines show the cut-off obtained with the homoscedastic bootstrap. Genes with lower overall levels of expression are required to exhibit larger differences in expression in order for those differences to be found significant.

4. Discussion

This analysis illustrates some of the issues that arise with microarray data. As others have noted (Yang et al. (2000)), there are dye effects in microarray data that are not linear on the log scale. A more elaborate transform than the log transform is required to put the data on a scale on which the effects are additive. A simple additive shift in the data before taking logs removed the systematic dye trends in the data. We prefer this adjustment to the loess transform because it is a robust procedure with a simple interpretation — that there is an additive difference between the dyes on the raw data scale. However, it would be naive to expect such a simple adjustment to work for all datasets. We found it effective here and with other datasets, but we consider the question open as to why these patterns reoccur and how to find the proper scale for the analysis of microarray data. For both the loess and shift-log adjustments, there are issues about the consequences for the final analysis of estimating a transformation from the data rather than using a pre-determined one. In addition, both transforms eliminate straightforward interpretation of differences in variety×gene effects as log fold-changes.

After settling on a scale for the data and a model, we observed non-constant variance in the residuals. To incorporate this into our statistical inference, we used a modified version of the bootstrap. In light of the observed heteroscedasticity, a natural question is whether weighted least-squares would be more appropriate than simple least-squares for fitting the model in the first place. Since the observed heteroscedasticity was modest, incorporating weighted least-squares would complicate the model estimation without much effect on the results. Furthermore, there is error in estimating the weights for weighted-least-squares that could discount any gain in estimation efficiency.

To handle heteroscedasticity, we assumed that residual variance is a function of the intensity of the overall gene effect G_g. It may be the case that all genes of similar intensity are not alike and the general assumption of gene-specific error distributions F_g is closer to reality. In larger experiments involving more arrays and greater replication there may be sufficient data to investigate this question. In this experiment, with only twelve residuals per gene, we could not estimate a gene-dependent error distribution F_g with satisfactory precision. We found the intensity-dependent variance assumption to be a workable middle-ground.

It is instructive to consider what was achieved by the replication included in this experimental design. First, if the experiment had been done with only a single array, then the effects of interest, VG, would have been completely confounded with DG effects. The gene-specific dye effects would bias any estimates of relative expression based on a single slide. Suppose instead that a single dye-swap experiment had been done using two arrays. In this case VG and DG effects would be orthogonal so unbiased estimates of VG effects could be obtained. On the other hand, Model (2) would be completely saturated, leaving 0 degrees of freedom to estimate error and evaluate the results. In order to obtain non-zero residuals one would be forced to assume that some effects were error. To illustrate this, we randomly chose one of the nine dye-swap experiments contained within this experiment. The sub-experiment chosen by chance contains arrays 3 and 4. Assigning DG effects to be error, the residual mean square is 0.0145, 2.8 times larger than in the full triple dye-swap experiment. This translates into less power to detect genes with small differences in expression between the varieties. Because larger DG effects are associated with lower overall gene expression, the heteroscedasticity observed in this sub-experiment is more pronounced

(Figure 7). It is clearly desirable to employ an experimental plan with sufficient replication to be able to account for these systematic sources of variation and to reduce error variance.

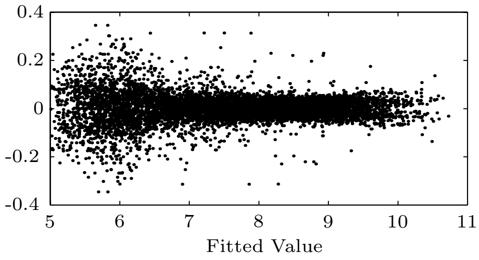

Figure 7. Residual plot for the sub-experiment containing only arrays 3 and 4. In order to have residual degrees of freedom, *DG* effects were assigned to error. Heteroscedasticity is somewhat more pronounced because the model does not account for *DG* effects, which are larger at low levels of expression.

We comment that the parameter estimates and residuals from this analysis have distributions that are notably heavier-tailed than normal (Figure 8(a)). A striking exception is the spot effects, whose distribution is extremely close to normal (Figure 8(b)). These effects are an obvious candidate to consider as random instead of fixed because they can been viewed as the result of the randomness associated with the robot that prints the spots on the arrays. Indeed, Wolfinger et al. (2000) use an ANOVA approach similar to ours but consider spot effects to be random with an underlying normal distribution. We agree that random effects are more appropriate, but are uncertain of proceeding with standard methods because of the non-normality in the residuals. This is an area of continuing research.

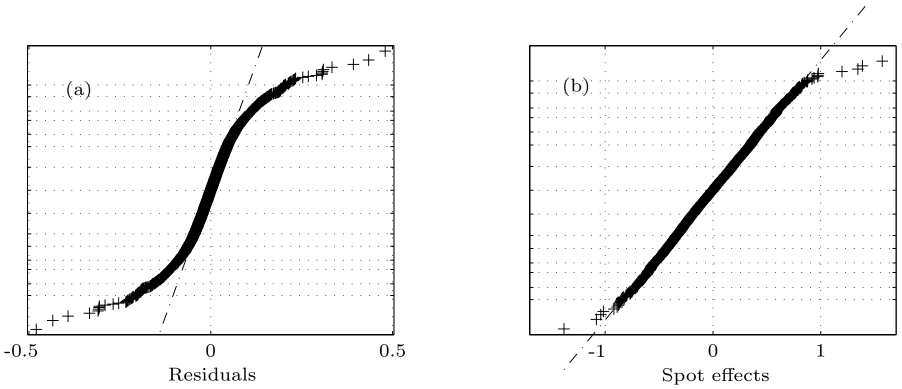

Figure 8. Normal probability plots for residuals and spot effects corresponding to Model (2) and the shift-log data. The residuals are much heavier-tailed than normal, while the spot effects are very close to normal.

Finally, this paper and others have stressed the importance of replication with microarray studies (Lee et al. (2000), Kerr and Churchill (2000), Kerr and Churchill (2001)). The "replication" in this study is multiple measurements of the same RNA samples. This kind of replication allows one to reduce the uncertainty introduced by experimental noise for more precise knowledge about the samples studied. In our study, inference extends to the two cell cultures that produced the RNA samples. This study does not include replication in the classical sense, which involves sampling multiple individuals from a population. Sampling multiple individuals allows one to make inferences about their populations. Callow, Dudoit, Gong, Speed and Rubin (2000) conducted a microarray experiment with this kind of replication. In that study, investigators used eight animals from each of three strains of mice to infer expression differences among the strains. A more accurate term for the replication we discussed in this paper might be "subsampling" (Snedecor and Cochran (1989, p.247)). Terminology is certainly not the primary concern, but it is unfortunate the language does not readily differentiate between these two types of studies. Careful consideration of the goals of a microarray experiment will help determine what level of "replication" is required to make desired inferences.

We hope the analysis described here will serve as a guide for analyzing microarray data and for computing unbiased estimates of relative gene expression with confidence bounds. We have presented practical solutions to some of the complications that regularly arise with this data. We do not claim to have found the best solutions to these problems. Rather, we have chosen methods that are theoretically reasonable but also realistic for routine implementation by conscientious data analysts.

The data discussed here and MATLAB scripts for ANOVA of microarray data are available at http://www.jax.org/research/churchill.

References

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**, 33-37.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research* **10**, 2022-2029.

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

Kerr, M. K. and Churchill, G. A. (2000). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183-201.

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

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

Kikuchi, H., Hossain, A., Yoshida, H. and Kobayashi, S. (1998). Induction of cytochrome P-450 1A1 by omeprazole in human HepG2 cells is protein tyrosine kinase-dependent and is not inhibited by alpha-naphthoflavone. *Arch. Biochemical Biophysics* **358**, 351-358.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Nat. Acad. Sci. U.S.A.* **97**, 9834-9839.

Li, W., Harper, P. A., Tang, B. K. and Okey A. B. (1998). Regulation of cytochrome P450 enzymes by aryl hydrocarbon receptor in human cells: CYP1A2 expression in the LS180 colon carcinoma cell line after treatment with 2,3,7,8-tetrachlorodibenzo-p-dioxin or 3-methylcholanthrene. *Biochemical Pharmacology* **56**, 599-612.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*, 8th edition. Iowa State University Press, Ames, Iowa.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. S. (2000). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biology* **8**, 625-637.

Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. P. (2000). Normalization for cDNA microarray data. Technical report 589, Department of Statistics, University of California, Berkeley. http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html

Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 98195, U.S.A.

E-mail: katiek@u.washington.edu

National Institute of Environmental Health Sciences Microarray Center, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.

E-mail: afshari@niehs.nih.gov

National Institute of Environmental Health Sciences Microarray Center/ITSS Contract, P.O. Box 12233, Research Triangle Park, NC 27709, U.S.A.

E-mail: bennet10@niehs.nih.gov

National Institute of Environmental Health Sciences Microarray Center, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.

E-mail: bushel@niehs.nih.gov

National Institute of Environmental Health Sciences Laboratory of Computational Biology and Risk Analysis, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.

E-mail: martinez@niehs.nih.gov

National Institute of Environmental Health Sciences Laboratory of Computational Biology and Risk Analysis, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.

E-mail: walker3@niehs.nih.gov

The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, U.S.A.

E-mail: garyc@jax.org

(Received February 2001; accepted October 2001)