JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 6, 2001
Mary Ann Liebert, Inc.
Pp. 557–569

A Model for Measurement Error for Gene Expression Arrays

DAVID M. ROCKE[1] and BLYTHE DURBIN[2]

ABSTRACT

We introduce a model for measurement error in gene expression arrays as a function of the expression level. This model, together with analysis methods, data transformations, and weighting, allows much more precise comparisons of gene expression, and provides guidance for analysis of background, determination of confidence intervals, and preprocessing data for multivariate analysis.

Key words: cDNA array, microarray, oligonucleotide array, statistical analysis.

1. INTRODUCTION

JUST AS WITH ANY OTHER ANALYTICAL TECHNOLOGY, measurements of gene expression with cDNA or oligonucleotide arrays have measurement errors. It is commonly observed (e.g. Chen, Dougherty, and Bittner, 1997) that the standard deviation of measurements rises proportionately to the expression level. However, this proportionality cannot continue down to genes that are entirely unexpressed because that would imply zero measurement error, which is not observed. The model proposed in this paper was originally developed in the context of instrumental methods of analytical chemistry, but these methods also exhibit the same kind of behavior referenced above (Rocke and Lorenzato, 1995). This model resolves the difficulties by incorporating both types of error that are observed in practice into a single model. This model provides an obvious advantage over existing models by describing the precision of measurements across the entire usable range. We also discuss the application of the model to detection limits, categorization of genes as expressed or unexpressed, comparison of expression between conditions on the same gene, construction of confidence intervals, and transformation and weighting of expression data for use in comparisons and in multivariate applications such as classification or clustering.

In Section 2, we introduce a model for the statistical variation of measurements of gene expression. In Section 3, we give some methods of estimation of the parameters of this model from a given data set. Finally, in Section 4, we discuss the important issue of comparison between treatment and control of expression of a given gene on a slide or across slides.

[1]Department of Applied Science, University of California, Davis, 3011 Engineering Unit III, Davis, CA 95616.
[2]Department of Statistics, University of California, Davis, Davis, CA 95616.

2. THE MODEL

Most measurement technologies require a linear calibration curve to estimate the actual concentration of an analyte in a sample for a given response. We can incorporate into the linear calibration model the two types of errors that are observed in most analyses. The two-component model for analytical methods such as GC/MS is

$$y = \alpha + \beta\mu e^{\eta} + \epsilon \tag{2.1}$$

where y is the response (such as peak area) at concentration μ, $\eta \sim N(0, \sigma_\eta)$ and $\epsilon \sim N(0, \sigma_\epsilon)$. Here, η represents the proportional error that always exists, but is noticeable at concentrations significantly above zero, and ϵ represents the additive error that always exists but is noticeable mainly for near-zero concentrations. The normality of the error terms ϵ and η is assumed for convenience, but this is in practice often a reasonable assumption. This two-component model approximates a constant standard deviation for very low concentrations and approximates a constant relative standard deviation (RSD) for higher concentrations.[1] Note that y is the response of the measuring apparatus, for example peak area. This model was introduced in Rocke and Lorenzato (1995) and has been applied in analytical chemistry and environmental science (for example, Zorn, Gibbons, and Sonzogni, 1997, 1999).

For gene expression arrays, it is unusual to have calibration data (that is, samples of known expression levels); thus, we cannot actually discern the expression level in molecular units, but can only do so relatively. The model then looks like this:

$$y = \alpha + \mu e^{\eta} + \epsilon \tag{2.2}$$

where y is the intensity measurement, μ is the expression level in arbitrary units, and α is the mean background (mean intensity of unexpressed genes).[2] Our best estimate of μ is $y - \hat{\alpha}$, the background-corrected observed intensity. The first error term is $\epsilon \sim N(0, \sigma_\epsilon)$, which represents the standard deviation of the background (unexpressed genes), and the second error term is $\eta \sim N(0, \sigma_\eta)$, which represents the proportional error that always exists, but is noticeable mainly for highly expressed genes.

Under this model, the variance of the response y at concentration μ is given by

$$\mathrm{Var}(y) = \mu^2 e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) + \sigma_\epsilon^2 \tag{2.3}$$

(Rocke and Lorenzato, 1995). A derived quantity will be useful in interpretation of the results. $S_\eta = \sqrt{e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)}$ is the approximate relative standard deviation (RSD) of y for high levels.

Using this derived quantity, we can represent the variance of y as

$$\mathrm{Var}(y) = \mu^2 S_\eta^2 + \sigma_\epsilon^2. \tag{2.4}$$

We illustrate this with a data set from an experiment on the response of male Swiss Webster mice to a toxic substance (Bartosiewicz *et al.*, 2000). The treated animal received an intraperitoneal injection of 15mg/kg of β-naphthoflavone while the control mouse had an injection of the carrier (corn oil) of equal volume. The two-color spotted cDNA slides were constructed using Molecular Dynamics equipment, with data from a treated mouse and a control mouse on each slide. Data were replicated usually a total of eight times per slide (meaning eight spots from the same microplate well were placed on the slide). We use data from one slide. Figures 1–3 illustrate the phenomena. Figure 1 shows the close linear relationship between replicate mean and replicate standard deviation at high expression levels. The line shown is the predicted

[1] The RSD is also often called the coefficient of variation.

[2] Background is used here for the statistical distribution of overall intensity measurements for genes that are actually not expressed in the sample. We do not discuss here the image processing issues in which background may refer to the pixel distribution in areas of the slide in which there is no spot.

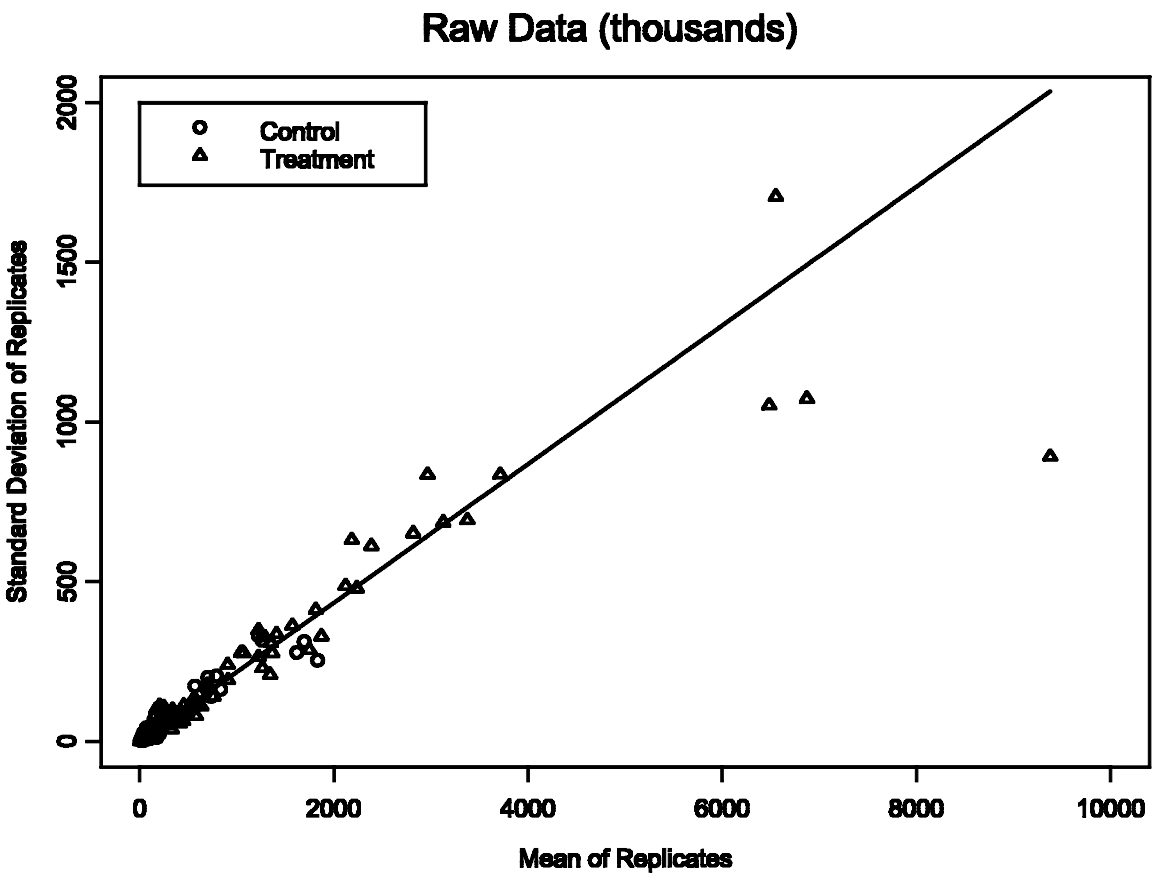

FIG. 1. Standard deviation vs. the mean of replicates for raw expression data. Each point represents the mean and standard deviation of eight replicates on a slide. The circles are the control measurements and the triangles are the treatment measurements. The solid line is the predicted standard deviation from the two-component model with parameters fit to the data as described in the text. This does not differ materially on this scale between control and treatment data.

standard deviation from the model; the slope of this line at high levels is the relative standard deviation. Figure 2 shows the approximately constant standard deviation of the natural logarithms of the data above a log intensity of about 13 or so. Note that when the RSD of the original set of replicates is not too large, the standard deviation of the natural logarithms is about the same as the RSD of the untransformed data. This figure also illustrates the vexing phenomenon that the logarithmic transformation, which nicely stabilizes the variance at high levels, produces highly variable results for low expression levels. The lines show the predicted standard deviation from the model for control data (solid line) and treatment data (dotted line) using (4.2). Figure 3 shows the approximately constant variance of the raw data below a measurement threshold as predicted by the two-component model.

It is important to realize that most of the variation observed on a cDNA or oligonucleotide array is caused by variations in μ, the actual expression. Variation within replicated spots at the same level μ of true expression is the measurement error that we model, and this is typically much smaller. For our example, the observed mean intensity varies across genes from about 6,000 units to over 9,000,000 units. Using the two-component model, the uncertainty in a mean of 9,000,000 units over eight replicates amounts to a confidence interval of about 7,650,000 to 10,590,000, far smaller than the variability between different genes. We also conjecture that variation in μ across individuals or experimental conditions for a fixed gene also has a two component structure in that the standard deviation at low expression levels is approximately constant while at high expression levels the RSD is approximately constant. This issue will be addressed in further work.

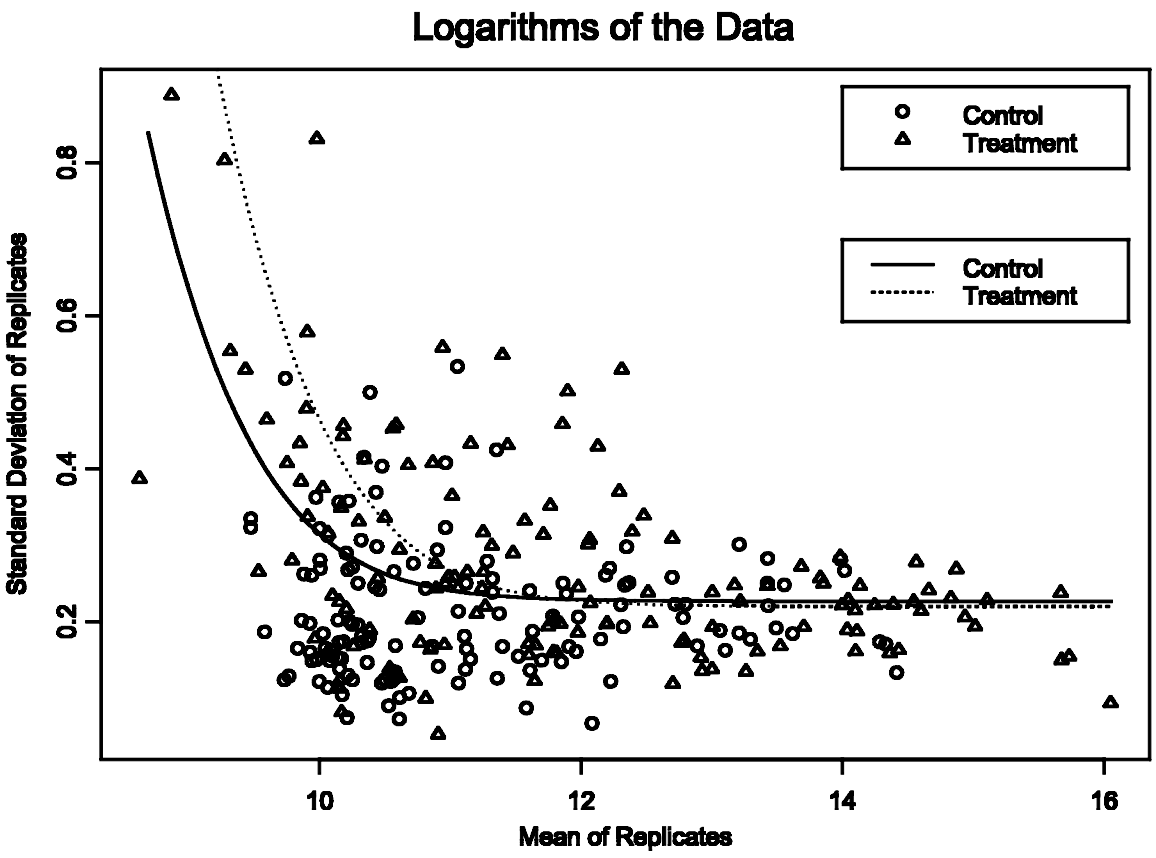

FIG. 2. Standard deviation vs. the mean of replicates for natural logarithms of expression data. Each point represents the mean and standard deviation of eight replicates on a slide. The circles are the control measurements and the triangles are the treatment measurements. The solid line (dotted line) is the predicted standard deviation from the two-component model for control (treatment) data with parameters fit to the data as described in the text.

3. ESTIMATION

A model such as (2.2) cannot be used in practice unless the parameters can be estimated. In this section, we discuss methods of estimation and necessary characteristics of the data for estimation to be possible.

3.1. Estimation of background using negative controls

The easiest way to estimate the mean α and standard deviation σ_ϵ of the low level measurements is from replicate blanks (negative controls). The standard deviation of the negative controls would be used as the estimate of σ_ϵ. The mean intensity of the negative controls is a suitable estimate of α, the mean background.

3.2. Estimation of background with replicate measurements

If we have replicated measurements, but no specific negative controls, we can still estimate the background mean and variance. According to (2.2), intensity measurements from unexpressed genes will be normally distributed with mean α and standard deviation σ_ϵ. This can be done separately for treatment and control in a two-color array.

1. Begin with a small subset of genes with low intensity, such as the 10% of genes with lowest intensity measurements. Compute the mean $\bar{x}_B$ of the genes including all replicates and the pooled standard deviation s_B of replicates of these genes that have been replicated. For each replicated gene in this

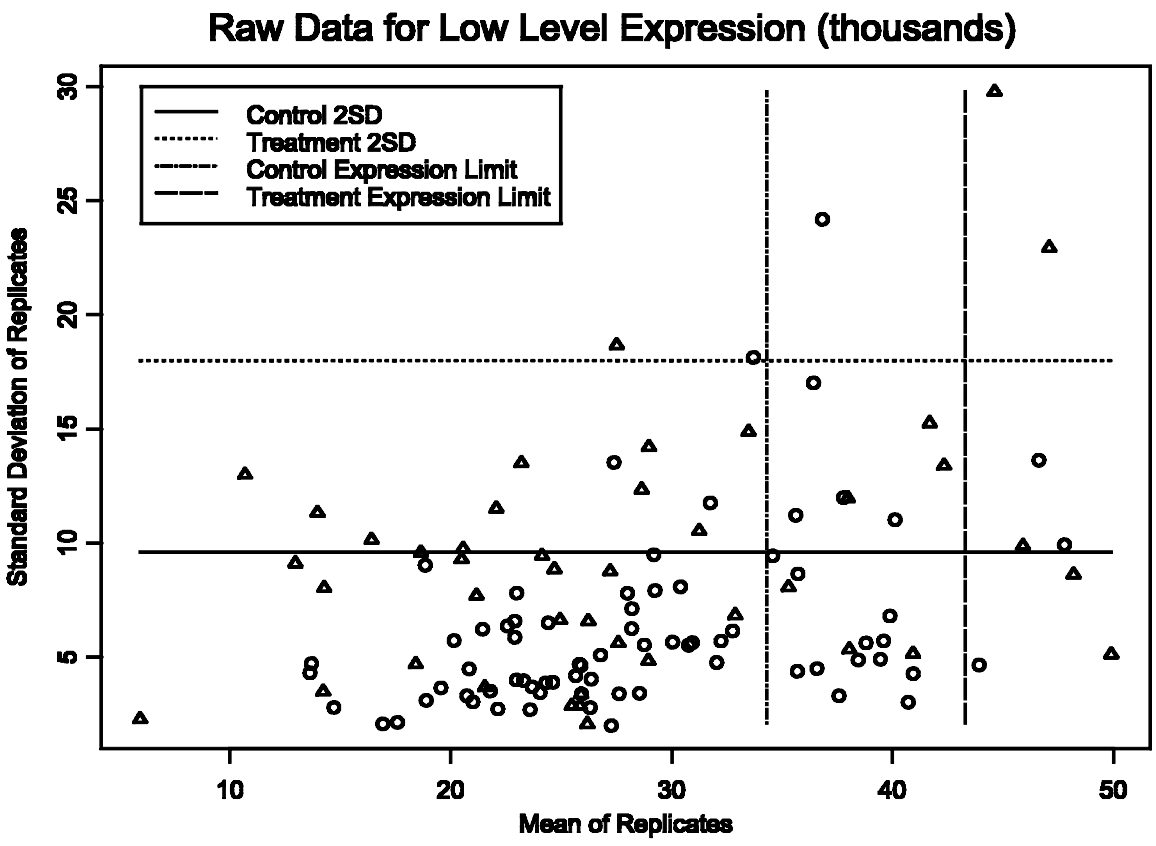

FIG. 3. Standard deviation vs. the mean of replicates for raw expression data for low-level expression. Each point represents the mean and standard deviation of eight replicates on a slide. The circles are the control measurements and the triangles are the treatment measurements. The solid line (dotted line) is at two estimated standard deviations for low-level data for control (treatment) data. The dash-dotted (dashed) line is the limit beyond which expression is statistically significant for the control (treatment) data.

group, compute the standard deviation s_i of the replicates. If there are m replicated genes, pool these estimates as follows:

$$s_B = \sqrt{(n-m)^{-1}\sum_{i=1}^{m} s_i^2(n_i - 1)}$$

where n_i is the number of replicates for gene i and $n = \sum_{i=1}^{m} n_i$. If there are a large number of such genes, or many replicates, it may be better to use only those replicated genes whose average expression is less than $\bar{x}_B$ to determine s_B.

2. Define a new subset consisting of genes whose intensity values are in the interval $[\bar{x}_B - 2s_B, \bar{x}_B + 2s_B]$. Recompute $\bar{x}_B$ and s_B.
3. Repeat the previous step until the set of genes does not change.

At the final step, the set of genes should include at least 95% of the unexpressed genes. If one includes the genes below $\bar{x}_B - 2s_B$, this would include about 97.5% of the unexpressed genes. Depending on the distribution of actual expression levels, this estimate could be biased up both in the mean and (slightly) in the standard deviation, because it is impossible in principle to distinguish an unexpressed gene from one with such a low expression level that it is below detection limits. Nonetheless, this estimate should be of considerable use in screening genes for expression.

The standard deviation of a replicated gene is at least σ_ϵ, and for genes that are unexpressed or expressed at very low levels, it will be essentially exactly this. This process cannot lead to the empty set, since at

least one replicated gene will be included. In practice, the process cannot converge to estimates of α or σ_ϵ that are systematically too small, and since the bias in the standard deviation at low levels is typically small, the estimates are rarely too large by much. Furthermore, the solution to which this process converges does not appear at all to depend on the details of the selection of the initial set, so long as it is fairly small. These observations have been confirmed by extensive simulations that are omitted here. In the next section, we present a method of estimating α and σ_ϵ even from unreplicated data.

3.3. Estimation of background without replication

In the absence of replicated measurements, it is still possible to estimate the mean and variance of unexpressed genes (background); the following procedure is recommended. This can be done separately for treatment and control in a two-color array.

1. Begin with a small subset of genes with low intensity, such as the 10% of genes with lowest intensity measurements. Compute the mean $\bar{x}_B$ and standard deviation s_B of these genes.
2. Define a new subset consisting of genes whose intensity values are in the interval $[\bar{x}_B - 2s_B, \bar{x}_B + 2s_B]$. Recompute $\bar{x}_B$ and s_B.
3. Repeat the previous step until the set of genes does not change.

At the final step, the set of genes should include at least 95% of the unexpressed genes. As before, this estimate may be biased upwards but nonetheless should be of considerable use in screening genes for expression. Note that this procedure is less reliable than the one to be used when there are an adequate number of replicates because the standard deviation is taken across genes. If some of these genes are actually expressed, the standard deviation is elevated by the variation in means, whereas this does not happen with replicated data.

A variant of this procedure may reduce the bias somewhat. In this variant, one uses the median of the expression levels of the subset of genes as the estimate of location and uses MAD/.6745 as the estimate of s_B, where the MAD is the median absolute deviation from the median. This is calculated by subtracting the median from each expression value in the subset, taking absolute values, and taking the median of the resultant set of absolute deviations. This is documented in detail in Nguyen and Rocke (2000).

3.4. Estimation of the high-level RSD

The parameter σ_η can be estimated from the standard deviation of the logarithms of high level replicated measurements in much the same way as the background standard deviation can be estimated from the low level data. For each replicated gene that is expressed at a high level, compute the standard deviation s_i of the logarithms of the replicate estimates $\hat{\mu} = y - \hat{\alpha}$ of μ. If there are m replicated genes, one then pools these estimates as follows:

$$s_H = \sqrt{(n-m)^{-1} \sum_{i=1}^{m} s_i^2 (n_i - 1)}$$

where n_i is the number of replicates for gene i and $n = \sum_{i=1}^{m} n_i$. This method works because for high expression levels, (2.2) is indistinguishable from

$$\hat{\mu} = \mu e^{\eta}$$

$$\ln(\hat{\mu}) = \ln(\mu) + \eta$$

which is a constant mean, constant variance model.

There is no method even in principle for estimating measurement error without at least some replication at high levels since it is impossible from an unreplicated sample to know if an intensity value is high because the expression is high or because of a positive measurement error. This fact of life should be an important determinant of experimental design in microarrays.

3.5. *What are "high" and "low" expression?*

Given the model, and preliminary estimates of the parameters, we can address the issue of high and low expression and the variability of genes in each group. The variance of y given by (2.4) can be compared with the variance of y at low levels. If the ratio is smaller than, say, 0.9, then most of the variance is due to the additive error component. Thus

$$\begin{aligned}\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2+\mu^2 S_\eta^2} &\geq 0.9\\ \sigma_\epsilon^2 &\geq 0.9\sigma_\epsilon^2+0.9\mu^2 S_\eta^2\\ \mu^2 &\leq \frac{0.1\sigma_\epsilon^2}{0.9 S_\eta^2}\\ \mu &\leq \sigma_\epsilon/3S_\eta. \end{aligned} \tag{3.1}$$

Thus, one can define "low level" data as ones where the observed expression is smaller than this threshold. Similarly,

$$\begin{aligned}\frac{\mu^2 S_\eta^2}{\sigma_\epsilon^2+\mu^2 S_\eta^2} &\geq 0.9\\ \mu^2 S_\eta^2 &\geq 0.9\sigma_\epsilon^2+0.9\mu^2 S_\eta^2\\ \mu^2 &\geq \frac{0.9\sigma_\epsilon^2}{0.1 S_\eta^2}\\ \mu &\geq 3\sigma_\epsilon/S_\eta \end{aligned} \tag{3.2}$$

gives a threshold above which the variance is mostly due to the multiplicative component.

An examination of the example data shows that the variance is approximately constant below about 25,000 and the variance of the logarithm is approximately constant above about 13. Use of the procedure given in Sections 3.2 and 3.4 yields parameter estimates as given in Table 1.

Using (3.1), for the control data, the logarithms have approximately constant variance when

$$\begin{aligned}\mu &\geq 3\sigma_\epsilon/S_\eta\\ \mu &\geq 61,000\end{aligned}$$

corresponding to a signal of 24,800 + 61,000 = 85,800 and a log signal of 11.4. For the treatment data, the equivalent values are $\mu \geq 118{,}400$, corresponding to a signal of 25,300 + 118,400 = 143,700 and a log signal of 11.9.

Table 1. Parameter Estimates for the Example Data[a]

Parameter	*Control*	*Treatment*
$\hat{\alpha}$	24,800	25,300
$\hat{\sigma}_\epsilon$	4,800	9,000
$\hat{\sigma}_\eta$	.227	.220
$\hat{\sigma}_\eta/\sqrt{r}$	.080	.078
S_η	.236	.228
Expression cutoff	34,300	43,300

[a]This shows the estimates for the treatment and control data separately; estimates for the combined model are given in Section 4.2. In this case, r, the number of replicates, is eight. The expression cutoff is the intensity above which the gene is expressed at a level that is statistically significantly above zero.

Using (3.2), the raw control data have approximately constant variance when

$$\mu \leq \sigma_\epsilon / 3S_\eta$$
$$\mu \leq 6,800$$

corresponding to a signal of 24,800 + 6,800 = 31,600. For the treatment data, the equivalent calculation give $\mu \leq 13{,}200$, corresponding to a signal of 38,500. In the range 31,600 to 85,800 for controls and 38,500 to 143,700 for treatment, both the variance and the coefficient of variation are changing substantially.

For data with calibration curves, the most effective estimation method is maximum likelihood, as described in Rocke and Lorenzato (1995). A method of applying maximum likelihood for replicated microarrays is in development, but the more heuristic methods given here may be satisfactory for many applications.

3.6. Uncertainty of a single measurement

The uncertainty of a single measurement is usually quantified using confidence intervals. There are two primary approaches to this problem, an exact solution and a normal or lognormal approximation. The exact solution requires numerical integration and will not be discussed here. Say we would like a 95% confidence interval for μ based on a single measurement, $\hat{\mu}$. The approximate method for low values of $\hat{\mu}$, using an estimated variance and a normal approximation is

$$\hat{\mu} \pm 1.96\sqrt{\mathrm{Var}(\hat{\mu})} \tag{3.3}$$

where $\mathrm{Var}(\hat{\mu})$ is estimated using (2.4). For high levels of $\hat{\mu}$, (those in which the second term in equation (2.4) dominates), $\ln \hat{\mu}$ is approximately normally distributed with variance σ_η^2. Hence, a 95% confidence interval for μ is

$$(\exp(\ln \hat{\mu} - 1.96\hat{\sigma_\eta}), \exp(\ln \hat{\mu} + 1.96\hat{\sigma_\eta})). \tag{3.4}$$

Note that this interval is symmetric on the log scale, but asymmetric on the original measurement scale.

We can also use this method to give confidence intervals for the average of a series of replicate measurements. For low levels, the average of r measurements will be approximately normally distributed with standard deviation $\sqrt{\mathrm{Var}(\hat{\mu})/r}$. For larger values of $\hat{\mu}$, the average of the natural log of the r measurements will have approximate standard deviation $\sigma_\eta/\sqrt{r}$. Confidence intervals can then be constructed as above, using the appropriate standard deviations.

4. COMPARING EXPRESSION LEVELS

The most common method of comparing gene expression is to examine the ratio of expression for treatment and control, or its logarithm. As we will see below, this procedure is effective when the gene is expressed at a moderate level or higher in both samples, and less effective when the gene is expressed only at a low level in one condition or the other. In the extreme case, a gene may not be expressed at all in the control sample (so that the true expression is zero). In this case, even a large change in expression is difficult to detect by ratios (since the standard deviation will be extremely large), and such a ratio is estimating a quantity that is not even well defined.

There is not a large previous literature on methods of comparing expression. The most developed previous treatment is that of Chen *et al.* (1997). Our approach expands on theirs in several ways—most importantly, addressing the genes with low expression where the assumption of constant relative standard deviation cannot hold. We do not repeat the details of their methods here, but instead refer the reader to their explication of the issues.

The specific methods we recommend for comparing measurements will depend on whether the measurements (e.g., control and treatment) are statistically independent (measured on different slides, for example) or share common errors. The second case is particularly represented by two-color arrays in which control and treatment are measured on the same spot. We first analyze the independent case, which is easier to explicate.

4.1. Comparing samples with independent errors

For a gene in which the two samples both have low level measurements, the significance of a difference can be assessed even when there is only one sample of each using the variance estimate consisting of the sum of the two separate background variances. Similarly, when the two samples both have high level measurements for that gene, then we can compare the log expression measurements using a variance estimate that is the sum of the squares of the high-level RSD's for the two cases. When there are several independent slides for each gene, a standard t-test on the raw data or the logarithms can be performed, depending on whether the expression is low or high.

If, however, the data are not obliging enough to conform to the stricture that both samples must have high expression or both must have low expression, then problems ensue. In general, the most interesting results will be those in which a gene is significantly expressed under at least one condition. In such a case, the use of the logarithmic transformation is commonly recommended. We must use a more complex analysis when the gene may not be expressed at a high level in both conditions. Using the delta method, it is straightforward to show that

$$\mathrm{Var}(\ln(\hat{\mu})) = \mathrm{Var}(\ln(y - \hat{\alpha})) \approx \sigma_\eta^2 + \sigma_\epsilon^2/\mu^2. \tag{4.1}$$

In spite of the many advantages of the logarithmic transformation, including stabilization of the variance for high level data, this transformation has several important problems. First, it is not defined for intensity measurements $y \leq \hat{\alpha}$. One would expect about half of the unexpressed genes to have intensity measurements below $\hat{\alpha}$, and these low level data will be missing in any analysis using the logarithms of the estimated expression. Second, the variance of data near but still above $\hat{\alpha}$ will be extremely high. This can make some log ratios look very large, when they are not even statistically significant. An alternative is to take logarithms of the intensity values y (or in general $\ln(y + c)$. In this case, the delta method yields the following:

$$\mathrm{Var}\{\ln(y)\} \approx \frac{\mu^2\sigma_\eta^2 + \sigma_\epsilon^2}{(\mu + \alpha)^2} \approx \frac{\mu^2\sigma_\eta^2 + \sigma_\epsilon^2}{y^2}. \tag{4.2}$$

This still looks like $\ln(\hat{\mu})$ at high levels. If one of the genes is expressed at a low level, this variance can also be quite high, making comparisons difficult. For example, using the parameter values of our example, and ignoring for the moment that the control and treatment have been measured on the same spot, if a gene is unexpressed in the treatment, and the actual measured value is 7,000, and if the control is expressed at a high level, the difference has a standard deviation of

$$(9000)^2 + (4800)^2 + (.236)^2(\mu_C)^2 \tag{4.3}$$

and the log ratio has a standard deviation of about

$$(9000)^2/(7000)^2 + [(4800)^2 + (.236)^2(\mu_C)^2]/y_C^2. \tag{4.4}$$

The difference is superior to the log ratio until y_C reaches about 750,000. The expected t-statistic for the log ratio does not reach 2 until y_C exceeds 145,000, whereas the expected t-statistic for the difference reaches 2 at about 28,000. In general, differences may be superior to log ratios for estimation and testing of changes when one of more of the measurements is quite low.

4.2. Two-color arrays

Perhaps a more important case is when the treatment and control measurements are correlated; we describe the situation when using a two-color array with treatment and control on the same spot. A version of the two-component model that accommodates this structure is

$$y_C = \alpha_C + \mu_C e^{\eta_S + \eta_C} + \epsilon_S + \epsilon_C \tag{4.5}$$

$$y_T = \alpha_T + \mu_T e^{\eta_S + \eta_T} + \epsilon_S + \epsilon_T \tag{4.6}$$

where the subscript S indicates a component of variance due to the spot that is shared by control and treatment, and C and T are the specific control and treatment components of variance, respectively.

At low levels of both treatment and control expression, this is approximately

$$y_C \approx \alpha_C + \epsilon_S + \epsilon_C \tag{4.7}$$

$$y_T \approx \alpha_T + \epsilon_S + \epsilon_T, \tag{4.8}$$

a standard-components-of-variance model. In this case,

$$\mathrm{Var}(y_C) \approx \sigma^2_{\epsilon_S} + \sigma^2_{\epsilon_C} \tag{4.9}$$

$$\mathrm{Var}(y_T) \approx \sigma^2_{\epsilon_S} + \sigma^2_{\epsilon_T} \tag{4.10}$$

$$\mathrm{Var}(y_C - y_T) \approx \sigma^2_{\epsilon_C} + \sigma^2_{\epsilon_T}. \tag{4.11}$$

Solution of these equations provides estimates of the variance components, although (as usual for method-of-moments estimation of variance components) the estimates may be negative, in which case the estimate is usually taken to be zero. Figure 4 shows the three estimated variance components for each of the 31 genes (out of 138 total) that had both treatment and control at a low level. From this one might conclude that the controls at low levels have errors basically due to spot variation and that the treatment has an additional component of variance. Rough estimates are that the spot standard deviation is about 5,000, the treatment-specific standard deviation is about 7,000, and the control-specific standard deviation is negligible (meaning that the variance of the control measurement is essentially the spot variance).

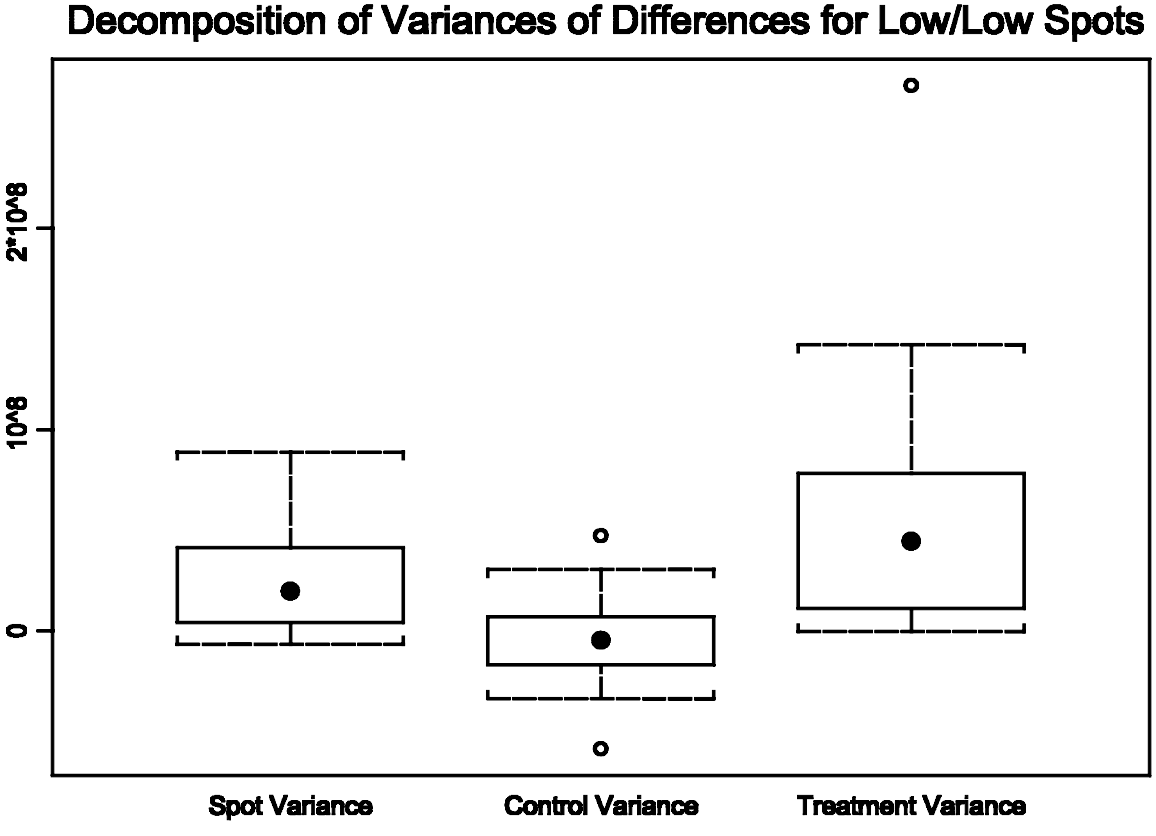

FIG. 4. Box plots of the components of variance of the differences of raw expression data when both treatment and control are expressed at a low level.

Similarly, if both treatment and control are expressed at high levels, we have approximately

$$\ln(y_C) \approx \ln(\mu_C + \alpha_C) + \eta_S + \eta_C \tag{4.12}$$

$$\ln(y_T) \approx \ln(\mu_T + \alpha_T) + \eta_S + \eta_T \tag{4.13}$$

and

$$\mathrm{Var}(\ln(y_C)) \approx \sigma^2_{\eta_S} + \sigma^2_{\eta_C} \tag{4.14}$$

$$\mathrm{Var}(\ln(y_T)) \approx \sigma^2_{\eta_S} + \sigma^2_{\eta_T} \tag{4.15}$$

$$\mathrm{Var}(\ln(y_C) - \ln(y_T)) \approx \sigma^2_{\eta_C} + \sigma^2_{\eta_T}. \tag{4.16}$$

Figure 5 shows the components of variance when both genes are expressed at a high level (16 out of 138), implying a large common spot effect. Roughly, the spot standard deviation is about .22, the treatment-specific standard deviation is about .03 and the control-specific standard deviation is about .06.

For the log ratios of the example data, we have an average standard deviation of .0674 when both genes are highly expressed, .3429 when neither gene is highly expressed, and .1780 in intermediate cases. The average standard deviation of replicates of the log ratios over all 138 genes is .2173. Since the variance of measurements depends on the levels of treatment and control separately, and not just on the value of the ratio, great care needs to be taken in analyses in which many genes are used, as is commonly the case in cluster analysis and discrimination. It is perhaps important to note that the log ratio is quite precise for highly expressed genes. A value of $2(.0674) = .1348$ would imply significance, and this corresponds to a ratio of only 1.14. If one used the overall standard deviation of the log ratios (.2173) to assess significance,

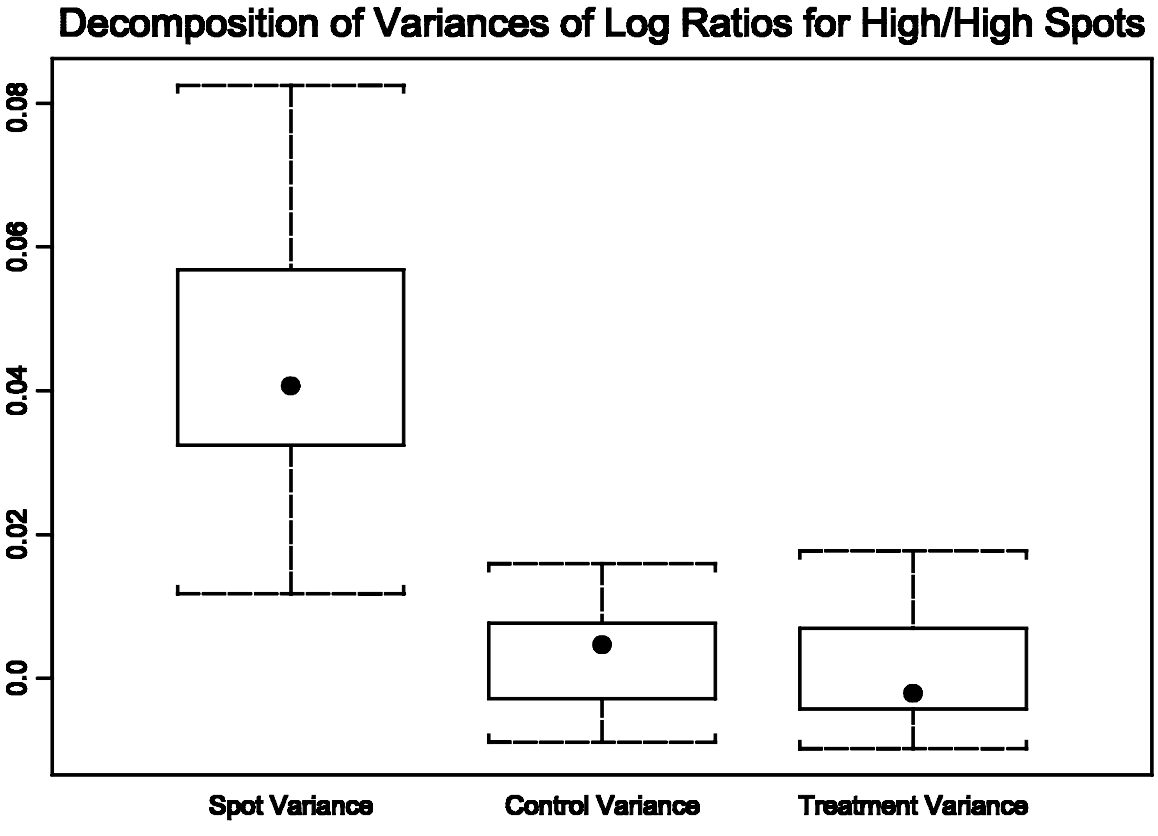

FIG. 5. Box plots of the components of variance of the logarithms of the ratios of expression data when both treatment and control are expressed at a high level.

the critical ratio is 1.54, which is much larger. The latter figure is actually an inappropriate mixture of high and low level behavior, which can be easily separated using the two-component model. This suggests that the repeatability of microarrays may be much better when used appropriately than is commonly thought.

Until better methods of coping with this problem are developed, one measure that should improve the performance of multivariate techniques is to use weights proportional to the reciprocals of estimated variances. If the data to be used are log ratios of same-spot data, then the approximate variance using the two-component model has a complex form. To determine this, we must provide an approximation that (unlike (4.14)–(4.16)) is valid even when one of the measurements is not at a high level. First, we linearize the two measurements in the form

$$\ln(y_C) \approx \ln(\mu_C + \alpha_C) + (\eta_S \mu_C + \eta_C \mu_C + \epsilon_S + \epsilon_C)/(\mu_C + \alpha_C) \tag{4.17}$$

$$\ln(y_T) \approx \ln(\mu_T + \alpha_T) + (\eta_S \mu_T + \eta_T \mu_T + \epsilon_S + \epsilon_T)/(\mu_T + \alpha_T) \tag{4.18}$$

so that approximately

$$\begin{aligned} \ln(y_T/y_C) \approx{}& \ln(\mu_C + \alpha_C) - \ln(\mu_T + \alpha_T) \\ &+ \eta_S[\mu_C/(\mu_C + \alpha_C) - \mu_T/(\mu_T + \alpha_T)] \\ &+ \eta_C \mu_C/(\mu_C + \alpha_C) - \eta_T \mu_T/(\mu_T + \alpha_T) \\ &+ \epsilon_S[1/(\mu_C + \alpha_C) - 1/(\mu_T + \alpha_T)] \\ &+ \epsilon_C/(\mu_C + \alpha_C) - \epsilon_T/(\mu_T + \alpha_T) \end{aligned} \tag{4.19}$$

and

$$\begin{aligned} \mathrm{Var}[\ln(y_T/y_C)] \approx{}& \sigma_{\eta_S}^2[\mu_C/(\mu_C + \alpha_C) - \mu_T/(\mu_T + \alpha_T)]^2 \\ &+ \sigma_{\eta_C}^2 \mu_C^2/(\mu_C + \alpha_C)^2 + \sigma_{\eta_T}^2 \mu_T^2/(\mu_T + \alpha_T)^2 \\ &+ \sigma_{\epsilon_S}^2[1/(\mu_C + \alpha_C) - 1/(\mu_T + \alpha_T)]^2 \\ &+ \sigma_{\epsilon_C}^2/(\mu_C + \alpha_C)^2 + \sigma_{\epsilon_T}^2/(\mu_T + \alpha_T)^2. \end{aligned} \tag{4.20}$$

Note that if μ_C and μ_T are both large, this reduces to $\mathrm{Var}[\ln(y_T/y_C)] \approx \sigma_{\eta_C}^2 + \sigma_{\eta_T}^2$, in agreement with (4.16).

All of the parameters in the expression above can be estimated from the data (as shown above), so that a derived weight can be given as the reciprocal of this estimated variance for each log ratio. These can then be used in a weighted analysis of many different types.

If we assume that the data have been adjusted so that $\alpha_C = \alpha_T$, and if we consider the case of testing the hypothesis of equality, so that we take $\mu_T = \mu_C$ for the purpose of estimating the variance, we obtain the simplified expression

$$\mathrm{Var}[\ln(y_T/y_C)] \approx [(\sigma_{\eta_C}^2 + \sigma_{\eta_T}^2)\mu^2 + \sigma_{\epsilon_C}^2 + \sigma_{\epsilon_T}^2]/(\mu + \alpha)^2. \tag{4.21}$$

This can be used to determine which log ratios are significantly different from zero. Note that the test using the log ratios for the example data will have a standard deviation of .0674 when μ is large and will rise to .287 when μ is 0. A standard deviation of .1 (corresponding to an approximate CV of 10%) occurs at $\mu = 56,000$. Note that this would imply that differences as small as 20% would be statistically significant at and above this expression level.

5. CONCLUSION

We have introduced a new model for measurement error in gene expression microarray data that greatly eases the interpretation and comparison of these data. This model provides a common framework for determination of background means and standard deviations, for estimation of the measurement error of single measurements and means of replicates, for comparison of expression data, and for pre-processing data for multivariate methods.

ACKNOWLEDGMENTS

Research reported in this paper was supported by grants from the United States Environmental Protection Agency (CR 825621-01-0) and the National Institute of Environmental Health Sciences (P42 ES 04699). We are grateful for data and collaboration from Matt Bartosiewicz and Alan Buckpitt. We also gratefully acknowledge the comments of a referee, which helped significantly improve the presentation of the paper.

REFERENCES

Bartosiewicz, M., Trounstine, M. Barker, D., Johnston, R., and Buckpitt, A. 2000. Development of a toxicological gene array and quantitative assessment of this technology. *Arch. Biochem. Biophys*., 376, 66–73.

Chen, Y., Dougherty, E.R., and Bittner, M.L. 1997. Ratio–based decisions and the quantitative analysis of cDNA microarray images. *J. Biomedical Optics* 2(4), 364–374.

Nguyen, D.V., and Rocke, D.M. 2000. Analysis of gene expression from microarrays using partial least squares, manuscript.

Rocke, D.M., and Lorenzato, S. 1995. A two-component model for measurement error in analytical chemistry. *Technometrics* 37(2), 176–184.

Zorn, M.E., Gibbons, R.D., and Sonzogni, W.C. 1997. Weighted least-squares approach to calculating limits of detection and quantification by modeling variability as a function of concentration. *Analytical Chem.* 69, 3069–3075.

Zorn, M.E., Gibbons, R.D., and Sonzogni, W.C. 1999. Evaluation of approximate methods for calculating the limit of detection and the limit of quantification. *Environmental Science and Technology* 33, 2291–2295.

Address correspondence to:
David M. Rocke
Department of Applied Science
University of California, Davis
3011 Engineering Unit III
Davis, CA 95616

E-mail: dmrocke@ucdavis.edu