BIOINFORMATICS

Vol. 20 no. 5 2004, pages 660–667
DOI: 10.1093/bioinformatics/btg464

Variance-stabilizing transformations for two-color microarrays

Blythe P. Durbin[1,*] and David M. Rocke[2]

[1]Center for Image Processing and Integrated Computing and [2]Department of Applied Science, University of California, Davis, CA 95616, USA

Received on August 20, 2003; accepted on October 2, 2003

ABSTRACT

Motivation: Authors of several recent papers have independently introduced a family of transformations (the generalized-log family), which stabilizes the variance of microarray data up to the first order. However, for data from two-color arrays, tests for differential expression may require that the variance of the difference of transformed observations be constant, rather than that of the transformed observations themselves.

Results: We introduce a transformation within the generalized-log family which stabilizes, to the first order, the variance of the difference of transformed observations. We also introduce transformations from the 'started-log' and log-linear-hybrid families which provide good approximate variance stabilization of differences. Examples using control–control data show that any of these transformations may provide sufficient variance stabilization for practical applications, and all perform well compared to log ratios.

Contact: bpdurbin@ucdavis.edu

1 INTRODUCTION

A number of recent papers have addressed the importance of constant variance in the analysis of gene-expression microarray data (Durbin *et al*., 2002; Huber *et al*., 2002; Munson, 2001; Hawkins, 2001; Rocke and Durbin, 2003). These authors have generally approached variance stabilization in the context of one-color arrays or a single channel from a two-color array. However, variance stabilization is also crucial for comparison of pairs of samples from two-color microarrays. Because two observations from the same spot on a two-color array will be correlated, techniques intended for statistically independent data (such as for comparison of one-color arrays) will not always apply in the two-color case.

A key purpose of a two-color microarray experiment is the comparison of two samples in order to determine which genes are differentially expressed. As with many statistical techniques, hypothesis tests for differential expression may be more effectively performed on data that have been transformed so that they have constant variance. Of course, one could attempt to circumvent the issue of non-constant variance by, e.g. testing for differential expression on a gene-by-gene basis via a two-sample t-test not assuming equal variances. However, one may soon find oneself left with few significant differences after compensating for hundreds or thousands of multiple tests. Should one wish to 'borrow' power for testing from other genes by using an ANOVA model, non-constancy of variance quickly becomes an obstacle.

A common approach for determining differential expression is to examine the ratio z_T/z_C, or its logarithm $\ln(z_T/z_C)$. However, Durbin *et al*. (2002) show that $\ln(z)$ has a greatly inflated variance for μ close to 0. Due to this non-constancy of variance, a log ratio that is statistically significant for one pair of true expression values (μ_T, μ_C) may not be significant for a different pair of values, even if the log ratio itself remains the same. Therefore, log ratios do not appear to provide an optimal means of determining differential expression. Extending previous work on one-color arrays (Rocke and Durbin, 2003), we present three different families of transformations as alternatives to log ratios.

2 THE TWO-COMPONENT ERROR MODEL FOR TWO-COLOR ARRAYS

Our choice of transformation in each family will be motivated by a model describing the variance–covariance structure of a pair of observations from the same spot on a two-color array. This error structure can be modeled by an extended version of the two-component error model of Rocke and Durbin (2001). Now, in a two-color microarray experiment, mRNA from two different biological samples is reverse-transcribed and labeled with two different fluorescent dyes, usually Cy3 and Cy5. The two samples are then hybridized to the same spotted cDNA array, resulting in two correlated measurements for each spot. This correlation requires the case of two-color arrays to be treated differently from, say, data from a pair of one-color arrays.

Rocke and Durbin (2001) model this pair of treatment and control observations for a single spot as

$$\begin{aligned} y_T &= \alpha_T + \mu_T e^{\eta_S+\eta_T} + \varepsilon_S + \varepsilon_T, \\ y_C &= \alpha_C + \mu_C e^{\eta_S+\eta_C} + \varepsilon_S + \varepsilon_C, \end{aligned} \tag{1}$$

*To whom correspondence should be addressed.

where y_T and y_C are the raw signal intensities for the control and treatment samples, respectively, μ_T and μ_C are the true expression levels of the gene in question, η_S and ε_S are spot-specific multiplicative and additive error terms shared by y_T and y_C, and η_T, η_C, ε_T and ε_C are multiplicative and additive error terms unique to control and treatment. Each error term is assumed to have mean 0 and to be stochastically independent from the others, with its own variance.

For the purposes of the following discussion, it will be more convenient to work with z_T and z_C rather than with y_T and y_C, where $z_T = y_T - \hat{\alpha}_T$. This presumes that sufficient background correction and normalization have already been applied to the data so that, for σ_{η_C}, σ_{η_T} and σ_{η_S} all small, $\mathrm{E}(z_T) \doteq \mu_T$ and $\mathrm{E}(z_C) \doteq \mu_C$. Preprocessing of the data prior to transformation should be limited to addition or subtraction of constants, as other changes have the potential to obscure the variance structure of the data.[1]

3 THE GENERALIZED-LOG TRANSFORMATION

Durbin *et al*. (2002), Huber *et al*. (2002) and Munson (2001) independently introduced the application to gene-expression microarray data of a transformation that stabilizes, to the first order, the variance of a random variable z satisfying

$$\mathrm{Var}(z) = a^2 + b^2\mu^2,$$

where $\mu = \mathrm{E}(z)$. (By 'to the first order' we mean that the first-order Taylor expansion of the function has constant variance not depending on μ.) This transformation may be written in several equivalent forms but we will use

$$h_{\lambda_0} = \ln\left(\frac{z + \sqrt{z^2 + \lambda_0}}{2}\right), \tag{2}$$

where $\lambda_0 = a^2/b^2$. This transformation converges to $\ln(z)$ for large z and is approximately linear at 0 (Durbin *et al*., 2002). The transformation and its inverse are monotonic functions with derivatives of all orders. Because its behavior for large values of μ is identical with the natural logarithm, and following Munson (2001), we will call this transformation a generalized logarithm.

Since, there exist transformations of the family $h_\lambda(z) = \ln[(z + \sqrt{z^2 + \lambda})/2]$ that stabilize the variances of z_T and z_C individually, it seems reasonable to search within this family for a transformation $h_\lambda(\cdot)$ such that

$$\Delta h_\lambda(z_T, z_C) = h_\lambda(z_T) - h_\lambda(z_C),$$

has constant variance. For the purpose of testing for differential expression, we need to know the variance of a test statistic in the null case, i.e. no differential expression. Therefore, for each of these families of transformations we will focus on the behavior of $\mathrm{Var}[\Delta h(z_T, z_C)]$ when $\mu_T = \mu_C = \mu$.

The approximate variance of $\Delta h_\lambda(z_T, z_C)$ for an unspecified parameter λ may be determined using the multivariate delta method. We may approximate the variance of $\Delta h_\lambda(z_T, z_C)$ by taking its first-Taylor expansion and evaluating the variance of the expansion. However, since $\Delta h_\lambda(z_T, z_C)$ is a function of the six independent random variables η_S, η_T, η_C, ε_S, ε_T and ε_C, we use an expansion in six variables rather than one, as would be the case with the univariate delta method. [The interested reader is referred to Chapter 7 of Ferguson (1996) for details.]

Once we have calculated the delta-method variance, we may solve for lambda such that $\mathrm{AV}_{\mu_C=\mu_T=\mu}[\Delta h_\lambda(z_T, z_C)]$ does not vary with μ, adopting the notation $\mathrm{AV}(X)$ to denote the delta-method approximated variance of a random variable X.

Using this technique we find that

$$\mathrm{AV}[\Delta h_\lambda(z_T, z_C)] = \frac{\mu^2(\sigma^2_{\eta_T} + \sigma^2_{\eta_C}) + \sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\mu^2 + \lambda}. \tag{3}$$

At $\mu = 0$ this becomes $(\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C})/\lambda$, and as $\mu \to \infty$,

$$\mathrm{AV}[\Delta h_\lambda(z_T, z_C)] \to \sigma^2_{\eta_T} + \sigma^2_{\eta_C}.$$

If the variance is to be constant, at the very least it should be equal at $\mu = 0$ and as $\mu \to \infty$. Setting

$$\frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\lambda} = \sigma^2_{\eta_T} + \sigma^2_{\eta_C},$$

and solving for λ yields the candidate value

$$\lambda^\star = \frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\sigma^2_{\eta_T} + \sigma^2_{\eta_C}}. \tag{4}$$

Inserting this value into (3) we find that

$$\mathrm{AV}[\Delta h_{\lambda^\star}(z_T, z_C)] = \sigma^2_{\eta_T} + \sigma^2_{\eta_C}, \tag{5}$$

which does not depend on μ. This member of the family of transformations

$$h_\lambda(z) = \ln\left(\frac{z + \sqrt{z^2 + \lambda}}{2}\right),$$

exactly stabilizes the delta-method variance of $h_\lambda(z_T) - h_\lambda(z_C)$, allowing meaningful hypothesis tests to be performed on the differences. One may compare (4) with the expression for one-color arrays of the optimal transformation parameter $\lambda = \sigma^2_\varepsilon/\sigma^2_\eta$ (Durbin *et al*., 2002).

4 THE STARTED-LOG TRANSFORMATION

While the generalized-log transformation of Section 3 is the exact delta-method variance-stabilizing transformation for

[1]Of course, further normalization may be performed on the data following transformation. For example, the loess normalization procedure of Yang *et al*. (2002) could be performed on data that have been transformed via a generalized-log transformation (2) rather than a log transformation.

data with a quadratic variance structure, transformations that only approximately stabilize the delta-method variance may occasionally be more convenient to use. In particular, log ratios are occasionally touted as providing better interpretability than alternatives, despite their inherent problems with inflation of the variance of low-level observations. However, one problem with log ratios that is more difficult to ignore is that of negative observations. When μ_T or μ_T is near 0, z_T or z_C will often be negative, in which case the log ratio is not defined. An *ad hoc* solution is simply to discard data for which z_T or z_C is less than zero; however, this approach can result in the loss of valuable biological information.

Should one insist on using log ratios to determine differential expression, a modified version of the logarithm, called the 'started logarithm' by Tukey (1964, 1977), can mitigate some of the problems with negative observations. This transformation takes the form

$$h_c(z) = \ln(z + c), \tag{6}$$

where $c > 0$. The delta-method variance of

$$\begin{aligned}\Delta h_c(z_T, z_C) &= h_c(z_T) - h_c(z_C),\\ &= \ln\left(\frac{z_T + c}{z_C + c}\right),\end{aligned}$$

under the null hypothesis $\mu_C + \mu_T = \mu$ is

$$\mathrm{AV}_{\mu_T=\mu_C=\mu}[\Delta h_c(z_T, z_C)] = \frac{\mu^2(\sigma^2_{\eta_C} + \sigma^2_{\eta_T}) + \sigma^2_{\varepsilon_C} + \sigma^2_{\varepsilon_T}}{(\mu + c)^2}, \tag{7}$$

$$= \frac{q^2 + \mu^2 r^2}{(\mu + c)^2}, \tag{8}$$

where

$$q = \sqrt{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}},$$

and

$$r = \sqrt{\sigma^2_{\eta_T} + \sigma^2_{\eta_C}}.$$

While no member of this family will exactly stabilize the delta-method approximated variance, we may ask for the choice of c that minimizes the maximum deviation of the variance from constancy. As $\mu \to \infty$,

$$\mathrm{AV}[\Delta h_c(z_T, z_C)] \to r^2,$$

which does not depend on c, so we will focus on the deviation of the variance from this limiting value. Following a lengthy derivation [which is exactly as in Rocke and Durbin (2003), and thus is not reproduced here], we find that the value of the shift constant minimizing the maximum deviation from constancy is

$$c = \frac{q}{2^{1/4} r}.$$

The minimized maximum deviation of the variance from constancy is

$$\frac{q^2}{c^2} - r^2 = r^2\sqrt{2} - r^2,$$

and the ratio of the SD at 0, which is $2^{1/4}r$, to the limiting SD r is about 1.2. For one-color arrays, the optimal shift constant is

$$c = \frac{\sigma^2_\varepsilon}{2^{1/4}\sigma^2_\eta},$$

which has the same structure as the optimal constant for differences but with q replaced by σ_ε and r replaced by σ_η (again, see derivation in Rocke and Durbin, 2003).

5 THE LOG-LINEAR-HYBRID TRANSFORMATION

A third class of transformations that may prove useful in the analysis of microarray data is the log-linear hybrid (Holder *et al.*, 2001). As described in Rocke and Durbin (2001), for μ close to 0, the untransformed data have approximately constant variance, and for μ large, $\ln(z)$ has approximately constant variance. This suggests that we might use a linear transformation for small z and a log transformation for large z.

Let

$$h_k(z) = \begin{cases} c + dz, & z \le k \\ \ln(z), & z > k. \end{cases} \tag{9}$$

If we choose c and d so that $h_k(z)$ is continuous with continuous derivative at k, we get $c = 1/k$ and $d = \ln(k) - 1$, yielding

$$h_k(z) = \begin{cases} z/k + \ln(k) - 1, & z \le k \\ \ln(z), & z > k. \end{cases} \tag{10}$$

It remains to choose k to minimize the maximum deviation of the variance of

$$\Delta h_k(z_T, z_C) = h_k(z_T) - h_k(z_C), \tag{11}$$

from constancy. The delta method variance of (11) takes four different forms, depending on the values of z_T and z_C relative to the splice point k. Therefore, under the null

hypothesis $\mu_T = \mu_C = \mu$,

$$\mathrm{AV}[\Delta h_k(z_T, z_C)]$$
$$= \begin{cases} \dfrac{\mu^2(\sigma^2_{\eta_T} + \sigma^2_{\eta_C}) + \sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{k^2}, & z_T, z_C \le k \\ \sigma^2_{\eta_T} + \sigma^2_{\eta_C} + \dfrac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\mu^2}, & z_T, z_C > k \\ \left(1 - \dfrac{\mu}{k}\right)^2 \sigma^2_{\eta_S} + \sigma^2_{\eta_T} + \dfrac{\mu^2}{k^2}\sigma^2_{\eta_C} & \\ \quad + \left(\dfrac{1}{\mu} - \dfrac{1}{k}\right)^2 \sigma^2_{\varepsilon_S} + \dfrac{\sigma^2_{\varepsilon_T}}{\mu^2} + \dfrac{\sigma^2_{\varepsilon_C}}{k^2}, & z_T > k,\ z_C \le k \\ \left(\dfrac{\mu}{k} - 1\right)^2 \sigma^2_{\eta_S} + \dfrac{\mu^2}{k^2}\sigma^2_{\eta_T} + \sigma^2_{\eta_C} & \\ \quad + \left(\dfrac{1}{k} - \dfrac{1}{\mu}\right)^2 \sigma^2_{\varepsilon_S} + \dfrac{\sigma^2_{\varepsilon_T}}{k^2} + \dfrac{\sigma^2_{\varepsilon_C}}{\mu^2}, & z_T \le k,\ z_C > k. \end{cases} \tag{12}$$

When $\mu = 0$,

$$\mathrm{AV}[\Delta h_k(z_T, z_C)] = \frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{k^2},$$
$$= \frac{q^2}{k^2},$$

where $q = \sqrt{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}$, as in Section 4. As $\mu \to \infty$,

$$\mathrm{AV}[\Delta h_k(z_T, z_C)] \to \sigma^2_{\eta_T} + \sigma^2_{\eta_C}$$
$$= r^2,$$

where $r = \sqrt{\sigma^2_{\eta_T} + \sigma^2_{\eta_C}}$, also as in Section 4.

Notice that when $\mu = k$, all four expressions become

$$\sigma^2_{\eta_T} + \sigma^2_{\eta_C} + \frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{k^2} = r^2 + \frac{q^2}{k^2}.$$

It can be seen that the value of k that minimizes the maximum deviation of the variance from constancy will be the one for which the variance at 0 is as much below the limiting value r^2 as the variance at the splice point is above r^2. Setting

$$r^2 - \frac{q^2}{k^2} = r^2 + \frac{q^2}{k^2} - r^2,$$

yields

$$k = \frac{q\sqrt{2}}{r}.$$

With this value of k, the maximum deviation of the variance from constancy is $r^2/2$, and the ratio of the SD of the difference at 0 to the limiting value r is about 0.7. For one-color data, the optimal transformation parameter is

$$k = \frac{\sigma_\varepsilon\sqrt{2}}{\sigma_\eta},$$

which has the same structure as the optimal splice parameter for one-color data, with q replaced by σ_ε and r replaced by σ_η (Rocke and Durbin, 2003).

6 COMPARISON OF ONE- AND TWO-COLOR CASES

As we pointed out above, the optimal transformation in each family has the same structure as that of the optimal transformation for one-color data. We will show that this occurs because, under the null hypothesis $\mu_C = \mu_T = \mu$, the variance of the transformed observations exhibits a similar structure. However, this similarity between variances in the one- and two-color cases does not hold in the case where $\mu_C \neq \mu_T$.

Now, if we assume (as in Rocke and Durbin, 2001), that a single untransformed microarray observation has variance

$$\mathrm{Var}(z) = \mu^2\sigma^2_\eta + \sigma^2_\varepsilon,$$

that observation transformed using an arbitrary function $h(\cdot)$ will have delta-method approximated variance

$$\mathrm{Var}[h(z)] = \dot{h}^2(\mu)\mu^2\sigma^2_\eta + \dot{h}^2(\mu)\sigma^2_\varepsilon. \tag{13}$$

Compare this with the variance of the difference of two transformed observations, not assuming $\mu_T = \mu_C$:

$$\begin{aligned} \mathrm{Var}[\Delta h(z_T, z_C)] = {} & \dot{h}^2(\mu_C)\mu_C^2\sigma^2_{\eta_C} + \dot{h}^2(\mu_T)\mu_T^2\sigma^2_{\eta_T} \\ & + \left[\dot{h}(\mu_C) - \dot{h}(\mu_T)\right]^2 \sigma^2_{\eta_S} \\ & + \dot{h}^2(\mu_C)\sigma^2_{\varepsilon_C} + \dot{h}^2\sigma^2_{\varepsilon_T} \\ & + \left[\dot{h}(\mu_C) - \dot{h}(\mu_T)\right]^2 \sigma^2_{\varepsilon_S}. \end{aligned} \tag{14}$$

When $\mu_T = \mu_C$, the second and fourth terms become 0 and the first and third combine to yield

$$\begin{aligned} & \mathrm{Var}_{\mu_C=\mu_T=\mu}[\Delta h(z_T, z_C)] \\ & \quad = \dot{h}^2(\mu)\mu^2(\sigma^2_{\eta_C} + \sigma^2_{\eta_T}) + \dot{h}^2(\mu)(\sigma^2_{\varepsilon_C} + \sigma^2_{\varepsilon_T}) \\ & \quad = \dot{h}^2(\mu)\mu^2 q^2 + \dot{h}^2(\mu)r^2, \end{aligned}$$

which has the same structure as (13). However, when $\mu_T \neq \mu_C$ (a setting that would be of interest for power calculations) the spot-specific variances $\sigma^2_{\eta_S}$ and $\sigma^2_{\varepsilon_S}$ fail to drop out and we are left with the more complicated variance structure of (14).

7 EXAMPLES

We illustrate the performance of these transformations with additional data from Bartosiewicz *et al*. (2000). We will use a small subset of the data presented in that paper, featuring control versus control experiments, in order to determine the behavior of the transformed data when there is no differential expression. For these data, two groups of three mice were each treated with 0.10 mg/kg of corn oil. mRNA from the livers of the mice was extracted, pooled and reverse-transcribed into fluor-labeled cDNA, with one group labeled with Cy5 and one group labeled with Cy3. Notice that this is not true self–self data, since three different mice were used for each group.

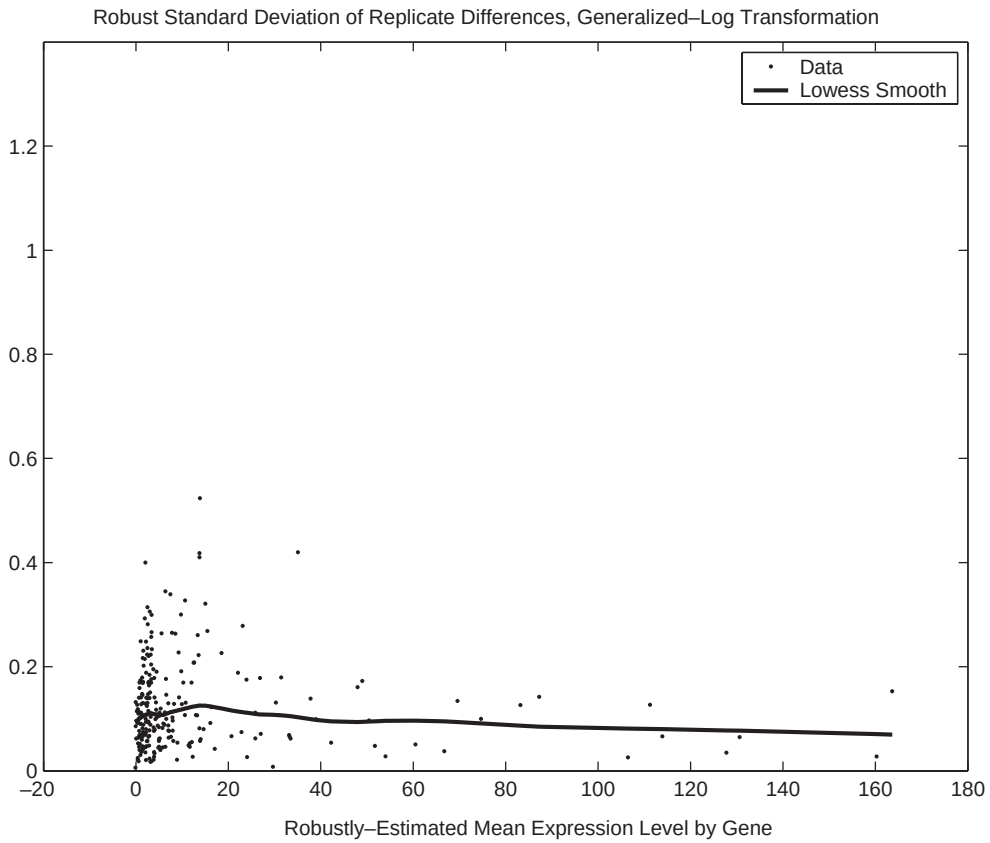

Fig. 1. Robustly estimated SD of differences of transformed observations versus robustly estimated mean expression, generalized-log transformation. The solid line on the plot is a lowess smooth of the data.

The cDNA was then hybridized to a spotted array in which each gene was replicated between 6 and 14 times. (We will use the term 'replicate' to refer to replicated spots of the same cDNA clone on the same array).

Parameters for the two-component model were estimated as described in Rocke and Durbin (2001). In this procedure, a set of observations close to background for both samples and a set of genes with replicated observations expressed well above background in both samples is identified via an iterative procedure. Now, according to (1), the variance of a control observation close to the expression background will be approximately

$$\mathrm{Var}(z_C) = \sigma^2_{\varepsilon_S} + \sigma^2_{\varepsilon_C},$$

the variance of a treatment observation close to background will be

$$\mathrm{Var}(z_T) = \sigma^2_{\varepsilon_S} + \sigma^2_{\varepsilon_T},$$

and the variance of the difference of paired low-level observations will be approximately

$$\mathrm{Var}(z_T - z_C) = \sigma^2_{\varepsilon_C} + \sigma^2_{\varepsilon_T}.$$

We may calculate the pooled sample variance for each color for each of the genes in the near-background group and the pooled sample variances of their differences in order to calculate the variances above. The three equations may then be solved for $\sigma^2_{\varepsilon_C}$, $\sigma^2_{\varepsilon_T}$ and $\sigma^2_{\varepsilon_S}$. The same procedure is repeated on logarithms of high-level data in order to calculate $\sigma^2_{\eta_C}$, $\sigma^2_{\eta_T}$ and $\sigma^2_{\eta_S}$. This method of estimating variance components may yield a negative variance estimate when the true value is small. By convention, negative estimates are set to 0 (Searle *et al.*, 1992).

This procedure yielded $\hat{\sigma}_{\varepsilon_C} = 0.335$, $\hat{\sigma}_{\varepsilon_T} = 0.0585$, $\hat{\sigma}_{\varepsilon_S} = 0.0747$, $\hat{\sigma}_{\eta_C} = 0$, $\hat{\sigma}_{\eta_T} = 0.135$ and $\hat{\sigma}_{\eta_S} = 0.143$. These model parameters yield the transformation parameters $\hat{\lambda} = 6.33$ for the generalized-log transformation, $\hat{c} = 2.12$ for the started-log transformation, and $\hat{k} = 3.56$ for the log-linear-hybrid transformation.

Figures 1–3 show the robustly estimated replicate SD of differences of transformed observations against the robustly estimated mean expression for the generalized-log, started-log and log-linear-hybrid transformations. The robust mean was estimated for each gene from pooled raw treatment and control observation using the S-Plus function location.m,

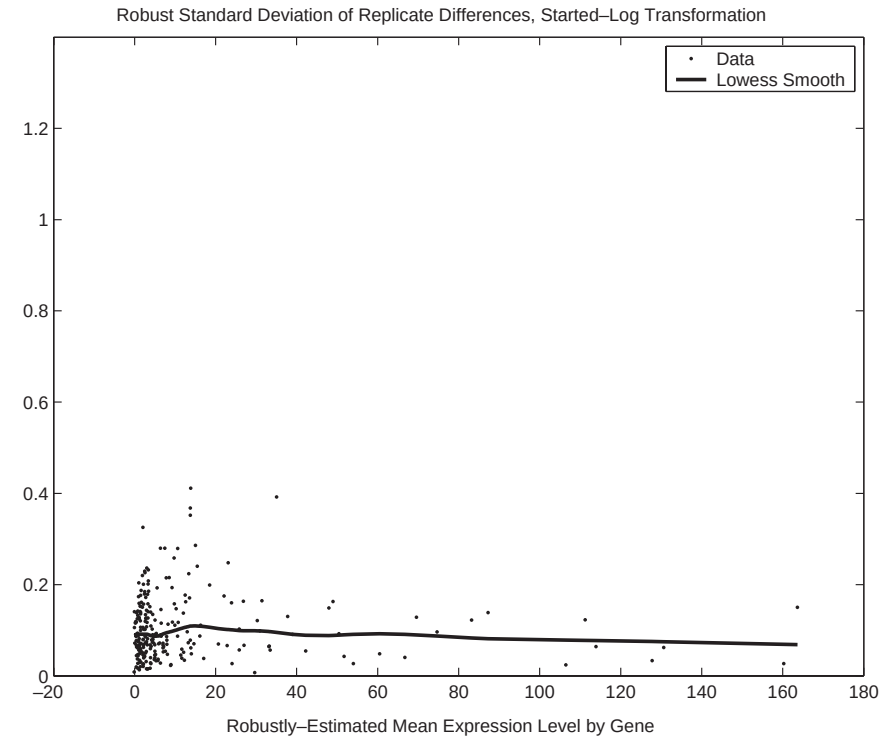

Fig. 2. Robustly estimated SD of differences of transformed observations versus robustly estimated mean expression, started-log transformation. The solid line on the plot is a lowess smooth of the data.

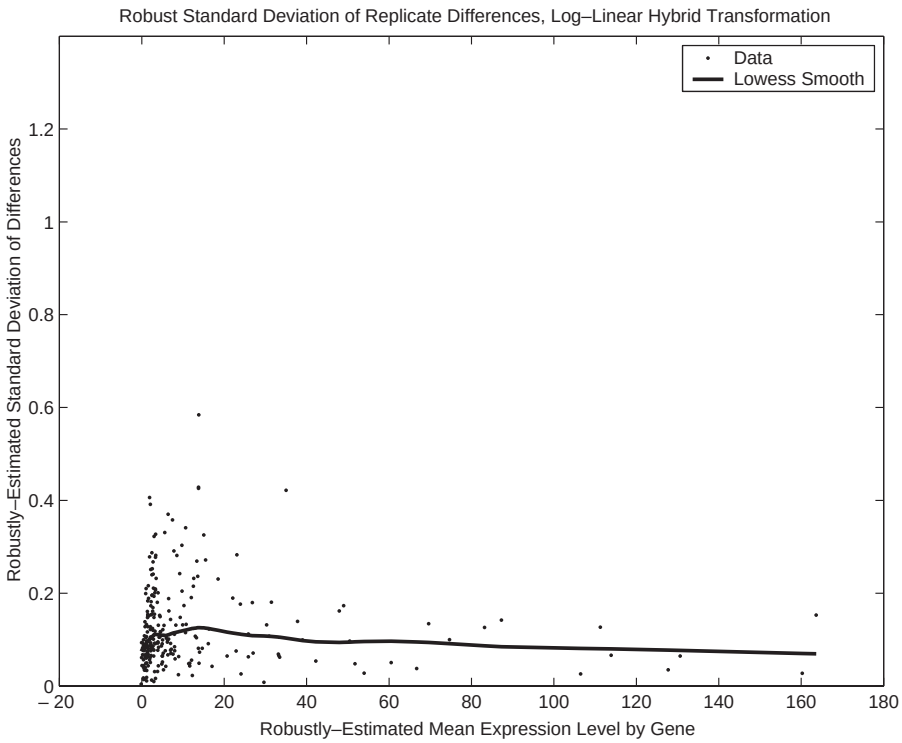

Fig. 3. Robustly estimated SD of differences of transformed observations versus robustly estimated mean expression, log-linear hybrid transformation. The solid line on the plot is a lowess smooth of the data.

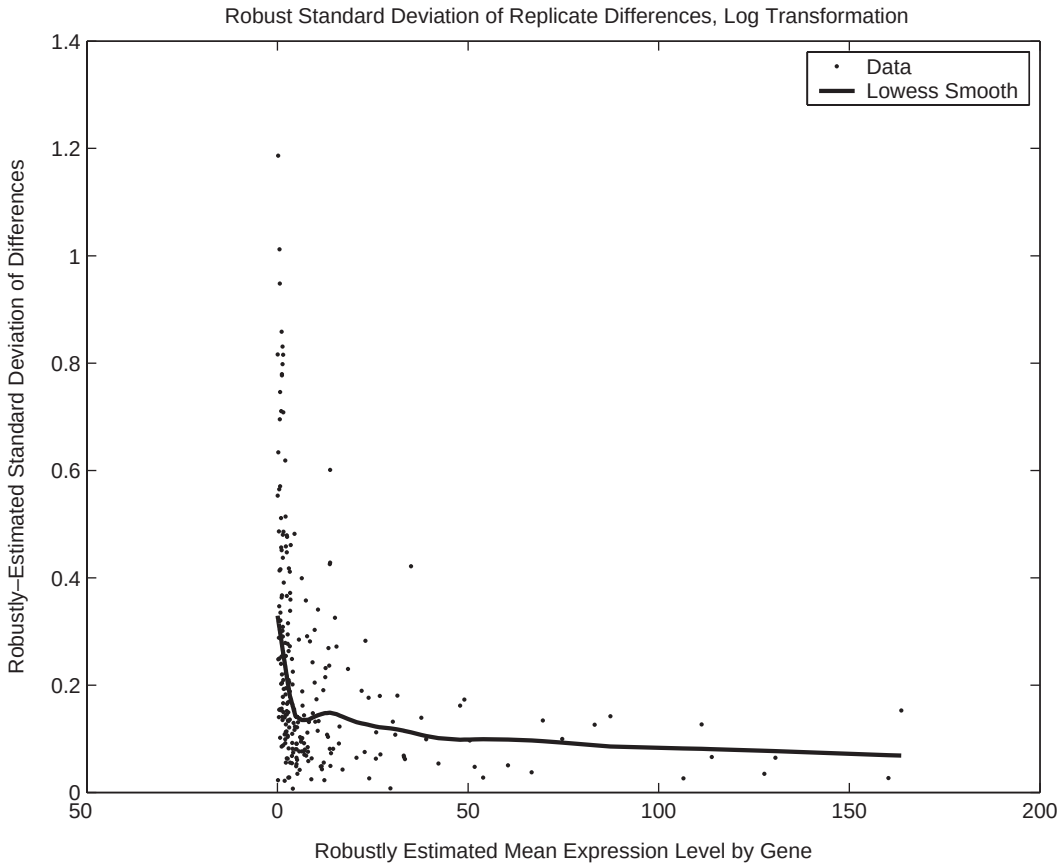

Fig. 4. Robustly estimated SD of differences of transformed observations versus robustly estimated mean expression, log transformation. The solid line on the plot is a lowess smooth of the data.

and the robust SD of differences of transformed observations was estimated using the S-Plus function scale.a. The solid line on each plot shows a lowess smooth that was fit to the robust means and SD.

In each case, the SD appears relatively constant when compared with the mean expression. Furthermore, the three plots look quite similar, indicating that each of these transformations does an adequate job of stabilizing the variance of the data. For comparison, Figure 4 shows the robustly estimated SD of the log ratios of the data plotted against the robustly-estimated mean expression. We removed 180 negative numbers (out of a total sample size of 2304) before taking the log transformation. As the lowess smooth shows, the SD increases as the mean expression decreases.

7.1 Minimizing the average deviation from constancy

One alternative to a transformation that minimizes the theoretical maximum deviation from constancy is one that minimizes the mean deviation from constancy of the actual data. The minimum-mean started log transformation was found for the data of Bartosiewicz *et al*. (2000) by taking the replicate SD of differences of transformed observations for each gene and looking at the mean absolute deviation from $\hat{r}$, the estimated theoretical limiting SD. This procedure was repeated for a number of different values of the shift constant c until the minimum was found.

For these data, the minimum-mean shift constant is $\tilde{c} = 2.42$, compared with $\hat{c} = 2.12$ for the minimax transformation. As the vast majority of observations are close to the expression background, in the same region where the theoretical maximum deviation from constancy occurs, these two procedures are likely to yield similar transformation parameters.

8 CONCLUSIONS

We have presented three variance-stabilizing transformations for gene-expression microarray data from two-color arrays, one that exactly stabilizes the delta-method variance of differences of transformed observations, and two other transformations, the started-log and log-linear hybrid transformations,

that provide approximate stabilization of the delta-method variance. When applied to actual data, each of these transformations appears to stabilize adequately the variance of differences of transformed observations, and all these transformations provide better variance stabilization than does the log transformation.

It should be mentioned that the 'exactness' of the variance-stabilization performed by the generalized-log transformation refers to its theoretical performance based on an approximation to the variance of the transformed data. Therefore, the other transformations in question, which are further approximations to an initial approximation, may not be less 'exact' in any meaningful sense. This can be seen in the equivalent performance of the three transformations compared. As with any theoretical result, the proof remains in the application.

ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for their insightful comments. The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020 and DMS 98-70712) and the National Institute of Environmental Health Sciences, National Institutes of Health (P43ES04699).

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.

Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18** (Suppl. 1), S105–S110.

Ferguson,T.S. (1996) *A Course in Large Sample Theory*. Chapman and Hall, London.

Hawkins,D.M. (2001) Diagnostics for conformity of paired quantitative measurements. *Stat. Med.*, **21**, 1913–1935.

Holder,D., Raubertas,R.F., Pikounis,V.B., Svetnik,V. and Soper,K. (2001) Statistical analysis of high density oligonucleotide arrars: a SAFER approach. *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*. Nov. 19, 2001, Bethesda, Maryland.

Huber,W., Von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Munson,P. (2001) A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*. Nov. 19, 2001, Bethesda, Maryland.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Rocke,D.M. and Durbin,B.P. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.

Searle,S., Casella,G., McCulloch,C.E. (1992) *Variance Components*, Wiley, Hoboken, NJ.

Tukey,J.W. (1964) On the comparative anatomy of transformations. *Ann. Math. Stat.*, **28**, 602–632.

Tukey,J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.