Vol. 19 no. 8 2003, pages 966–972
DOI: 10.1093/bioinformatics/btg107

Approximate variance-stabilizing transformations for gene-expression microarray data

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

[1]Department of Applied Science, University of California, Davis, Davis, CA 95616 and [2]Department of Statistics, University of California, Davis, Davis, CA 95616, USA

Received on August 20, 2002; revised on October 20, 2002; accepted on November 1, 2000

ABSTRACT

Motivation: A variance stabilizing transformation for microarray data was recently introduced independently by several research groups. This transformation has sometimes been called the generalized logarithm or glog transformation. In this paper, we derive several alternative approximate variance stabilizing transformations that may be easier to use in some applications.

Results: We demonstrate that the started-log and the log-linear-hybrid transformation families can produce approximate variance stabilizing transformations for microarray data that are nearly as good as the generalized logarithm (glog) transformation. These transformations may be more convenient in some applications.

Contact: dmrocke@ucdavis.edu

1 INTRODUCTION

Many traditional statistical methodologies, such as regression or the analysis of variance, are based on the assumptions that the data are normally distributed (or at least symmetrically distributed), with constant variance not depending on the mean of the data. If these assumptions are violated, the statistician may choose either to develop some new statistical technique which accounts for the specific ways in which the data fail to comply with the assumptions, or to transform the data. Where possible, data transformation is generally the easier of these two options (see Box and Cox, 1964; Atkinson, 1985).

Data from gene-expression microarrays, which allow measurement of the expression of thousands of genes simultaneously, can yield invaluable information about biology through statistical analysis. However, microarray data fail rather dramatically to conform to the canonical assumptions required for analysis by standard techniques. Rocke and Durbin (2001) demonstrate that the measured expression levels from microarray data can be modeled as

$$y = \alpha + \mu e^{\eta} + \varepsilon \qquad (1)$$

where y is the measured raw expression level for a single color, α is the mean background noise, μ is the true expression level, and η and ε are normally-distributed error terms with mean 0 and variance σ_η^2 and σ_ε^2, respectively. This model also works well for Affymetrix GeneChip arrays either applied to the PM-MM data or to individual oligos.

The variance of y under this model is

$$\mathrm{Var}(y) = \mu^2 S_\eta^2 + \sigma_\varepsilon^2, \qquad (2)$$

where $S_\eta^2 = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$. In Durbin *et al.* (2002); Huber *et al.* (2002) and Munson (2001), it was shown that for a random variable z satisfying $V(z) = a^2 + b^2\mu^2$, with $E(y) = \mu$, there is a transformation that stabilizes the variance to the first order, meaning that the variance is almost constant no matter what the mean might be. There are several equivalent ways of writing this transformation, but we will use

$$\left[\frac{(z + \sqrt{z^2 + c^2})}{2}\right] \qquad (3)$$

where $c = a/b$. This transformation converges to $\ln(z)$ for large z, and is approximately linear at 0 (Durbin *et al.*, 2002). Since this is exactly the natural logarithm when $c = 0$, it was called the generalized logarithm or glog transformation by Munson (2001), a terminology that we adopt. The inverse transformation is

$$f_c^{-1}(w) = e^w - c^2 e^{-w}/4.$$

Both f_c and its inverse are monotonic functions, defined for all values of z and w, with derivatives of all orders. For array data, we use $z = y - \alpha$ or $z = y - \hat{\alpha}$ so that the random variable satisfies (exactly or approximately) $V(z) = a^2 + b^2 E(z)^2$.

2 THE STARTED LOGARITHM

In some situations, it may not be convenient to use the glog transformation (3). In particular, the supposed ease of

*To whom correspondence should be addressed.

interpretation of log ratios has provided a major justification for use of the log transformation on microarray data. However, for a random variable z satisfying $E(z) = \mu$ and $V(z) = a^2 + b^2\mu^2$, the logarithmic transformation $\ln(z)$ has certain disadvantages. The delta method (i.e. propogation of errors) shows that $V(\ln(z)) \approx b^2 + a^2/\mu^2$, which goes to infinity as $\mu \to 0$. Furthermore, when $\mu = 0$, z will be frequently non-positive, for which the transformation is not defined.

A common modification of the logarithmic transformation, designed at a minimum to avoid negative arguments, is to add a constant to all of the values before taking the logarithm. Following Tukey (1964, 1977) we call this the 'started logarithm;' its form is

$$g_c(z) = \ln(z + c)$$

with $c > 0$. This transformation can, given the appropriate constant c, mitigate some of the problems with negative observations that plague the log transformation. A transformed observation $g_c(z)$ has approximate variance function

$$V(g_c(z)) = \frac{a^2 + b^2\mu^2}{(\mu + c)^2}. \quad (4)$$

This will not completely stabilize the variance of z if the variance function is (2), but we can ask for the choice of constant c which minimizes the maximum deviation from constancy. An examination of the function (4) shows that it takes the value a^2/c^2 at $\mu = 0$ and has an asymptote at b^2 as $\mu \to \infty$. We will focus on the deviation of the variance from the limiting value b^2.

The derivative of (4) with respect to μ is

$$\frac{2b^2\mu(\mu + c)^2 - 2(a^2 + b^2\mu^2)(\mu + c)}{(\mu + c)^4}. \quad (5)$$

The denominator of (5) is never zero for $\mu \geq 0$, so any change in sign of the derivative will occur where

$$2b^2\mu(\mu + c)^2 - 2(a^2 + b^2\mu^2)(\mu + c) = 0 \quad \text{or}$$
$$\mu = \frac{a^2}{b^2 c}.$$

Note also that the derivative of the variance function at $\mu = 0$ is $-2a^2/c^3 < 0$ (so long as $c > 0$), indicating that the variance decreases initially, before increasing again at $\mu = a^2/(b^2 c)$. It is clear that the value of c that minimizes the maximum deviation of (4) from b^2 is where the variance at 0 (a^2/c^2) is as much above b^2 as the variance at the minimum is below b^2 (see Figure 1). Since the minimum is at $\mu = a^2/b^2 c$, the variance at the minimum is

$$\frac{a^2 + b^2 a^4/(b^4 c^2)}{(a^2/b^2 c + c)^2} = \frac{a^2 b^2}{a^2 + b^2 c^2}$$

The condition to minimize the maximum deviation from constant variance is

$$\frac{a^2}{c^2} - b^2 = b^2 - \frac{a^2 b^2}{a^2 + b^2 c^2} \quad \text{or}$$
$$c = \frac{a}{2^{1/4} b}$$

The achieved minimum deviation is $b^2\sqrt{2} - b^2$, and the ratio of the standard deviation at 0 to the asymptotic standard deviation b is about 1.2.

We illustrate this transformation with a case from Durbin *et al.* (2002) in which $\alpha = 24\,800$, $a = 4\,800$ and $b = 0.227$. Figure 1 shows the standard deviation function for the optimal started-log transformation with $c = a/(2^{1/4} b) = 17\,781$, as well as two other values of c. The dashed line shows the value b, which is the value that all of the transformations tend to as the expression gets large. The upper (dotted) curve is for $c = 0$, corresponding to the logarithm of the background corrected data. The standard deviation approaches infinity as the estimated expression approaches 0. The lower curve (dot-dash) is for $c = 24\,800$, corresponding to the log uncorrected intensity. Here the variance at zero and at the minimum is too low. The optimal choice of $c = 17\,781$ (middle curve, solid line) has the correct balance between the two. In this case, the logarithm of the raw intensity data is not too bad. There is no guarantee that this would be true in general, since the zero of the intensity scale is rather arbitrary.

3 LOG-LINEAR HYBRID

According to the two-component model (1), the untransformed data have approximately constant variance for μ close to 0 and approximately constant coefficient of variation for μ large. This suggests that we might use a linear transformation for small z and a log transformation for large z. Keeping this in mind, another variant of the logarithm that may be appropriate for microarray data is the log-linear hybrid transformation (Holder *et al.*, 2001). Here we take the transformation to be $\ln(z)$ for z greater than some cutoff k, and a linear function $c + dx$ below that cutoff. This eliminates the singularity at zero. We choose c and d so that the transformation is continuous with continuous derivative at k.

The last requirements give the two equations

$$ck + d = \ln(k)$$
$$c = 1/k$$

and thus $d = \ln(k) -$ Thus, our transformation family is

$$h_k(z) = z/k + \ln(k) - 1, \quad z \leq k$$
$$= \ln(z), \quad z > k \quad (6)$$

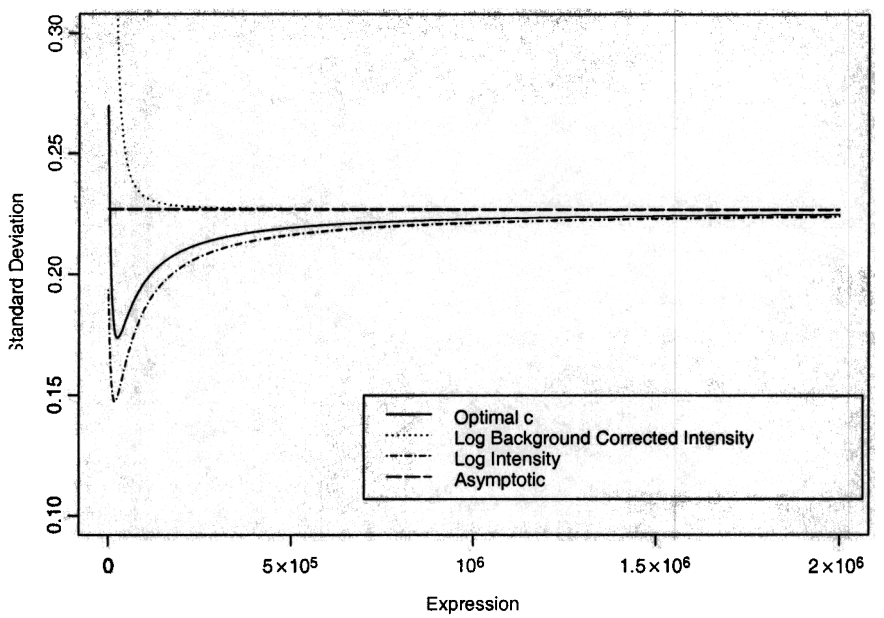

Fig. 1. Standard deviation of the started-log for three values of the constant.

The asymptotic delta-method variance function is given by

$$V(h_k(z)) = (a^2 + b^2\mu^2)/k^2, \quad z \leq k$$
$$= b^2 + a^2/\mu^2, \quad z > k. \tag{7}$$

Note that the two expressions agree at the splice point, due to the choice of c and d to make the derivative continuous at k.

It is easy to see that the choice of k that leads to the minimum deviation from constant variance is the one in which the variance at 0 is as much below b^2 as the variance at the splice point is above b^2. Thus

$$b^2 - a^2/k^2 = (b^2 + a^2/k^2) - b^2 \quad \text{or}$$
$$k = \sqrt{2}a/b \tag{8}$$

Figure 2 shows the optimal log-linear hybrid (solid line), the optimal started log (dotted line) and the optimal glog transformation (dot-dash line). In this case, the started log has a smaller maximum deviation from constant variance, but this is dependent on the parameter values and this can be reversed. Any of these transformations may be sufficient to stabilize the variance for practical purposes.

One can further reduce the maximum deviation from constant variance by employing both a linear segment and a started log, so that the transformation would be linear below a cutoff k and above that point be $\ln(z + c)$. However, the extra complexity that this would entail would make this choice an unlikely alternative to the glog transformation of Durbin *et al.* (2002); Huber *et al.* (2002); Munson (2001).

It should also be noted that the started log and log-linear hybrid each correspond to a variance function. The started log will be the optimal variance stabilizing transformation if $V(z) = (E(z) + c)^2$ and the log linear hybrid will be optimal if the variance is constant at $V(z) = k^2$ when $z < k$ and $V(z) = E(z)^2$ for $z \geq k$. These functions will be difficult to distinguish from the variance function (2) generated by the two-component model (1), although it may be possible with large data sets. We prefer the transformation (3) corresponding to the variance function (2) because it is generated by the physically plausible model (1), but the results are likely to be similar if the parameters are chosen carefully.

4 SIMULATION STUDIES

The relative performance of each of the three transformations was tested on data simulated from the two-component model of Rocke and Durbin (2001). The parameters used were $\sigma_\eta = 0.227$ and $\sigma_\varepsilon = 4800$. We use the value $b = \sigma_\eta = 0.227$ rather than $S_\eta = 0.236$ since the logartithms of data distributed according the the two-component model have a standard deviation that tends exactly to σ_η for large μ. To the order we are working, these quantities are the same, and make no practical difference for data analysis, but the difference can show up in large simulations. Data were simulated for values of μ ranging from 0 to 1 000 000 at increments

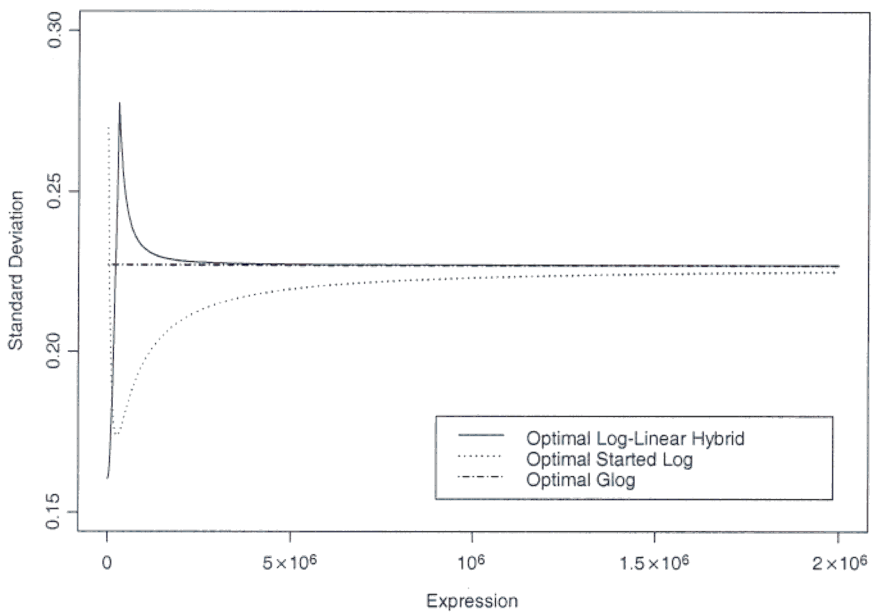

Fig. 2. Standard deviation of the optimal log-linear hybrid and the optimal started log.

of 5000. For each value of μ, 1000 samples of size 1000 were simulated from

$$z = \mu e^{\eta} + \varepsilon,$$

where $\eta \sim N(0, \sigma_\eta^2)$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The simulated data sets were transformed using each of the three transformations and used to calculate confidence intervals for the standard deviation and skewness of the transformed data. The optimal transformation within each family was used in all cases.

Figure 3 shows the standard deviation of the transformed simulated data, averaged over 1000 samples, for all three transformations. As would be expected, the glog transformation shows the most nearly constant standard deviation. The standard deviation of the data transformed using the log-linear-hybrid transformation stabilizes somewhat sooner than that using the started-log transformation, but otherwise these two transformations appear of similar quality.

Graphs (not shown) of the actual and model-predicted standard deviation of simulated data transformed using each of the three transformations, averaged over 1000 samples, show that the simulated data conform closely to the theoretical values, supporting the use of the delta-method theory in this analysis.

Upon examining the standard deviation of simulated data for each of the three transformations, it appears that the glog transformation provides the most nearly constant variance of transformed data, followed by the log-linear hybrid transformation. However, the skewness of the simulated data can also be informative, as symmetry of data is also important when applying standard statistical methodologies. Figure 4 shows the skewness of simulated data from each of the three transformations, averaged over 1000 samples. For a dataset of size 1000, the skewness differs significantly from 0 at the 95% level if it is greater than 0.1518 in absolute value. The glog transformation shows significant skewness between $\mu = 10\,000$ and $\mu = 35\,000$, with a maximum skewness of -0.2475 occuring at $\mu = 15\,000$. The started-log transformation shows significant skewness for values of $\mu < 30\,000$, with a maximum skewness of -1.2254 occuring at $\mu = 0$. Finally, the log-linear-hybrid transformation shows significant skewness for values of μ between 35 000 and 65 000, with a maximum skewness of -0.227 occuring at $\mu = 45\,000$. The glog and log-linear-hybrid transformations appear to perform equivalently at symmetrizing the simulated data, and both do far better than the started log transformation. Taking both variance-stabilization and symmetry into account, the glog transformation appears to perform best on the simulated data, followed by the log-linear hybrid.

5 EXAMPLE

Figures 5–7 show the results of applying the three transformations to the data from Durbin *et al.* (2002). All are much improved from the raw data or the logarithms of the background corrected data. Of these, the glog

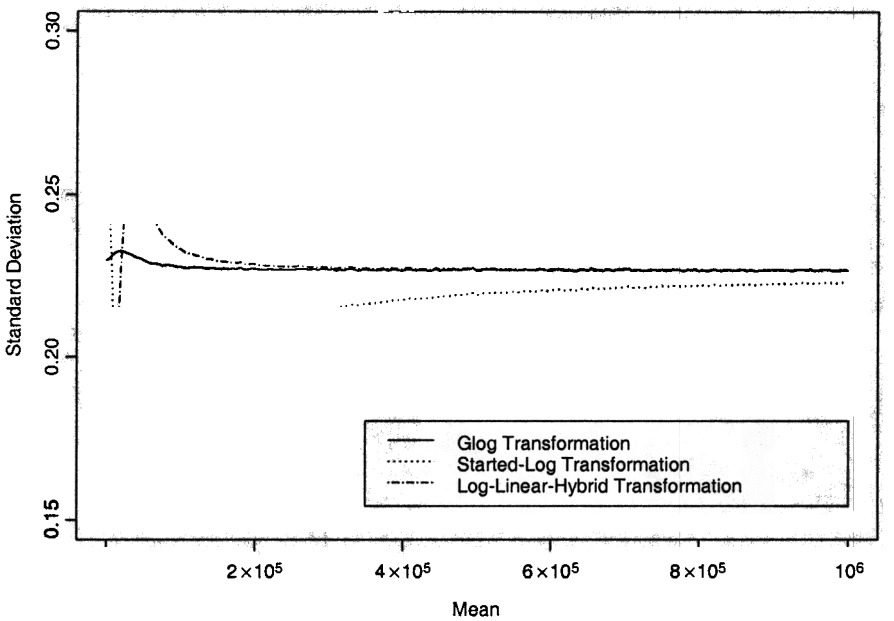

Fig. 3. Standard deviation of simulated data for three transformations.

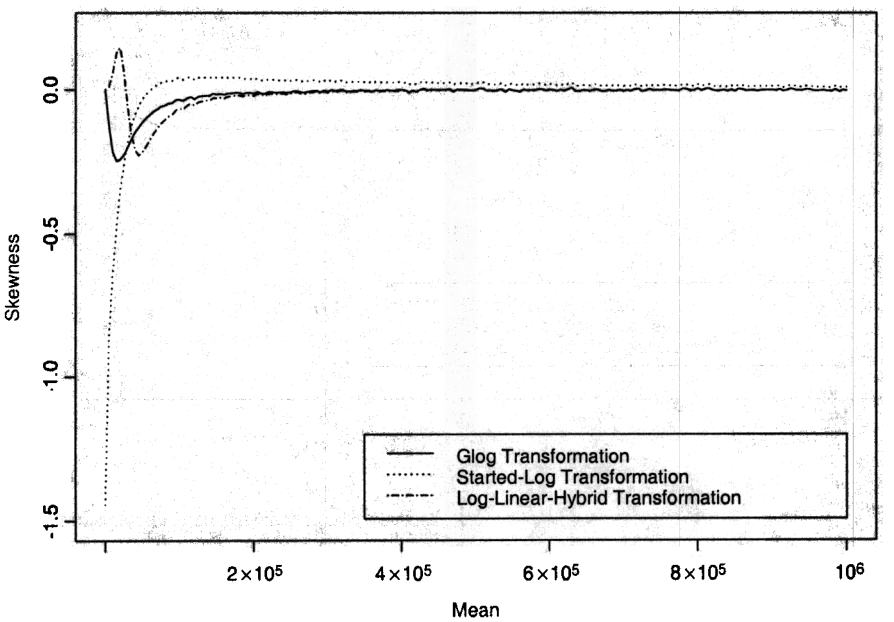

Fig. 4. Skewness of simulated data for three transformations.

transformation (Fig. 5) appears to have done the best job. The started log (Fig. 6) has several high-variance genes at the low end that deviate more from constancy than is the case with the glog transformation. The log-linear hybrid (Fig. 7) appears to have more low-variance genes near the low end (thus departing more from constancy of variance) than is the case with the variance-stabilizing transformation.

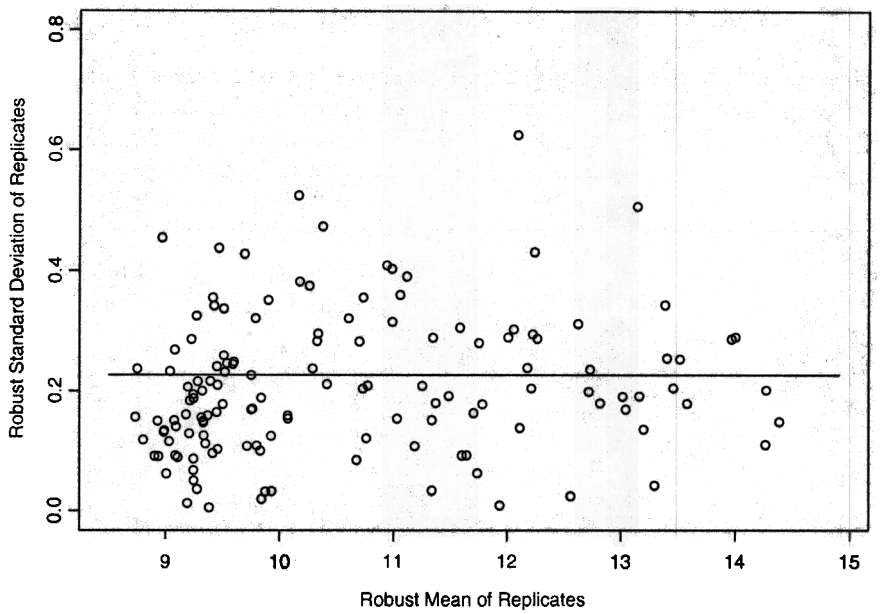

Fig. 5. Spread vs. location for the generalized log transformation.

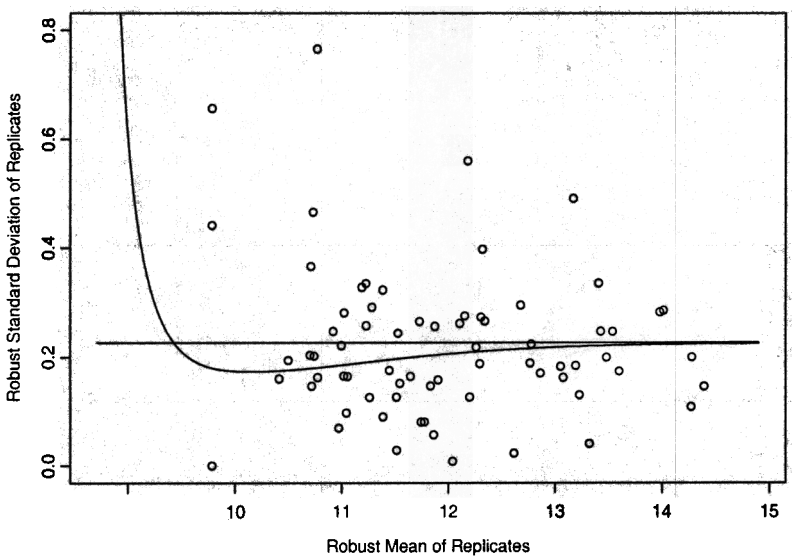

Fig. 6. Spread vs. location for the started-log transformation.

6 CONCLUSIONS

We have compared three transformation families, each optimized for stability of variance, for use with microarray data. Any of these could be usefully employed in this application, although evidence from theory and from an application suggest that the glog transformation of Durbin *et al.* (2002); Huber *et al.* (2002) and Munson (2001) is probably the best choice when it is convenient to use it.

ACKNOWLEDGEMENTS

The research reported in this paper was supported by grants from the National Science Foundation (ACI

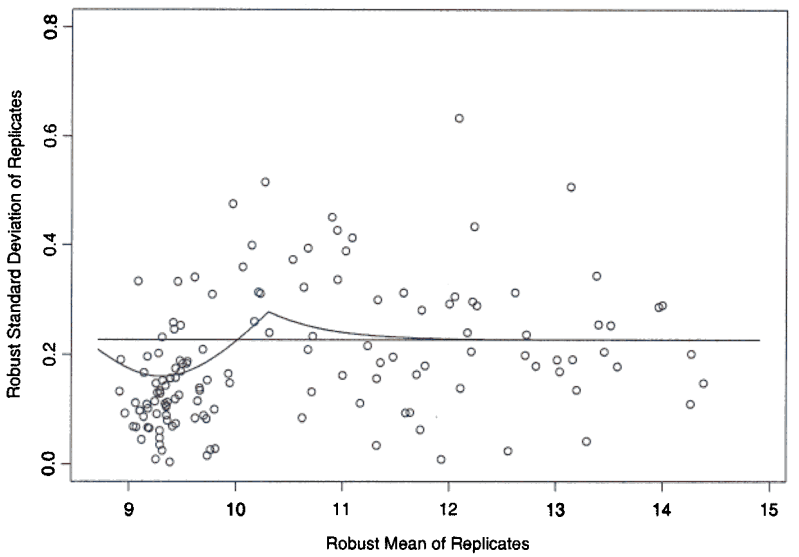

Fig. 7. Spread vs. location for the log-linear-hybrid transformation.

96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P43 ES04699). Tha authors are grateful for helpful suggestions from three referees that improved the presentation of the paper.

REFERENCES

Atkinson,A.C. (1985) *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford.

Bartosiewicz,M., Trounstine,M., Barker,D., Johnston,R. and Buckpitt,A. (2000) Development of a toxicological gene array and quantitative assessment of this technology. *Archives of Biochemistry and Biophysics*, **376**, 66–73.

Box,G.E.P. and Cox,D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, **26**, 211–252.

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**, S105–S110.

Hawkins,D.M. (2002) Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine*, **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.

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**, 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.

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

Tukey,J.W. (1964) On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, **28**, 602–632.

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