BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S105–S110

# A variance-stabilizing transformation for gene-expression microarray data

B. P. Durbin[1], J. S. Hardin[2], D. M. Hawkins[3] and D. M. Rocke[4]

[1]Department of Statistics, UC Davis, Davis, CA 95616, USA, [2]Department of Mathematics, Pomona College, Claremont, CA 91711, USA, [3]School of Statistics, U. of Minnesota, Minneaplis, MN 55455, USA and [4]Department of Applied Science, UC Davis, Davis, CA 95616, USA

Received on January 24, 2002; revised and accepted on March 31, 2002

## ABSTRACT

**Motivation:** Standard statistical techniques often assume that data are normally distributed, with constant variance not depending on the mean of the data. Data that violate these assumptions can often be brought in line with the assumptions by application of a transformation. Gene-expression microarray data have a complicated error structure, with a variance that changes with the mean in a non-linear fashion. Log transformations, which are often applied to microarray data, can inflate the variance of observations near background.

**Results:** We introduce a transformation that stabilizes the variance of microarray data across the full range of expression. Simulation studies also suggest that this transformation approximately symmetrizes microarray data.

**Contact:** bpdurbin@wald.ucdavis.edu

**Keywords:** cDNA array; microarray; statistical analysis; transformation; normalization.

## INTRODUCTION

Many traditional statistical methodologies, such as regression or ANOVA, 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 modelled as

$$y = \alpha + \mu e^{\eta} + \varepsilon, \tag{1}$$

where $y$ is the measured raw expression level for a single color, $\alpha$ is the mean background noise, $\mu$ is the true expression level, and $\eta$ and $\varepsilon$ are normally-distributed error terms with mean 0 and variance $\sigma_\eta^2$ and $\sigma_\varepsilon^2$, respectively.

At low expression levels (i.e., $\mu$ close to 0) the measured expression can therefore be written as

$$y \approx \alpha + \varepsilon, \tag{2}$$

implying that the measured expression is approximately normally distributed with mean $\alpha$ and constant variance $\sigma_\varepsilon^2$.

This phenomenon is demonstrated using the example data, which were collected from an experiment in which male Swiss Webster mice were injected with a toxin (Bartosiewicz *et al.*, 2000). The treated mouse received 0.15 mg/kg of $\beta$-napthoflavone and the control mouse received an equal amount of the corn-oil carrier. Each cDNA clone on the slide was replicated, usually 8 times.

Figure 1 shows robust estimates of the replicate mean and standard deviation for low level data. (As in the remaining plots, the robust estimates location.m and scale.a from the S-Plus statistical software package were used to estimate the mean and variance in order to minimize the impact of outliers.) Notice that when the robustly-estimated mean is close to 0, the standard deviation remains essentially constant, although it begins to increase for larger values of the mean.

When $\mu$ is large, the middle term in (1) dominates the others and the measured expression may be modelled as

$$y \approx \mu e^{\eta}. \tag{3}$$

Here the variance of $y$ is approximately $\mu^2 S_\eta^2$, where $S_\eta^2 = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$. The measured expression $y$ is

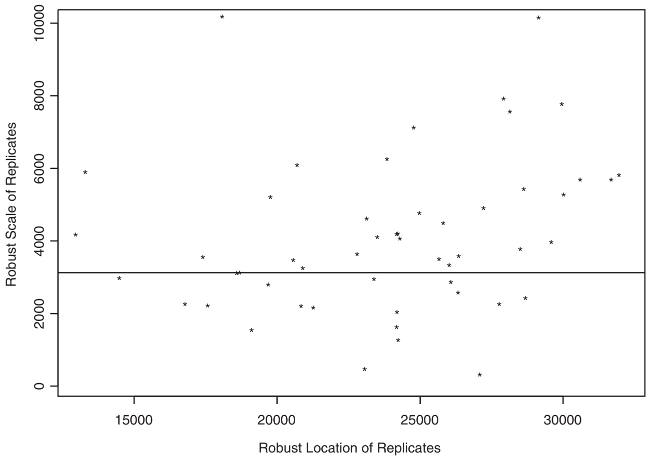

**Fig. 1.** Robust location of replicates versus robust scale of the replicates for raw (untransformed) data for low levels of expression. The line is the theoretical scale from the two-component model.

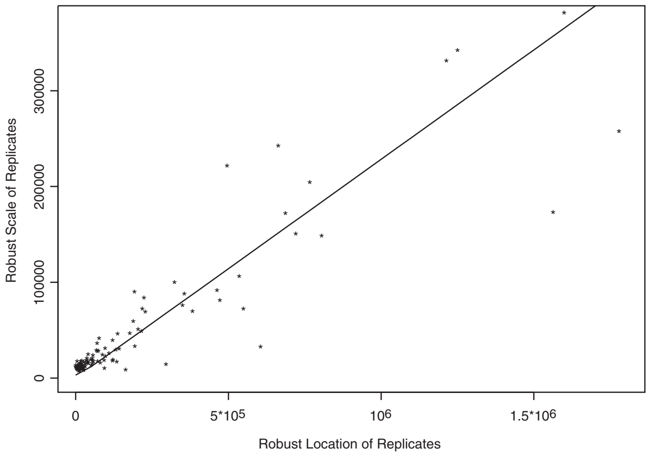

**Fig. 2.** Robust location of replicates versus robust scale of the replicates for raw (untransformed) data. The line is the theoretical scale from the two-component model.

here distributed approximately as a lognormal($\ln(\mu)$, $\sigma_\eta^2$) random variable. Notice that the standard deviation of $y$, $\mu S_\eta$, varies linearly with $\mu$. This can be seen in Figure 2, which shows robust estimates of the replicate mean and standard deviation of high level data. The linear asymptotic standard deviation is shown by the regression line through the plot.

On the log scale, (3) can be written as

$$\log(y) = \log(\mu) + \eta, \tag{4}$$

which implies that $\log(y)$ has constant variance for $\mu$ sufficiently large. This behaviour can be observed in Figure 3, which shows robust estimates of the replicate mean and standard deviation of the background-corrected log-transformed data. For values of the robustly-estimated mean greater than about 13, the estimated replicate standard deviation remains constant.

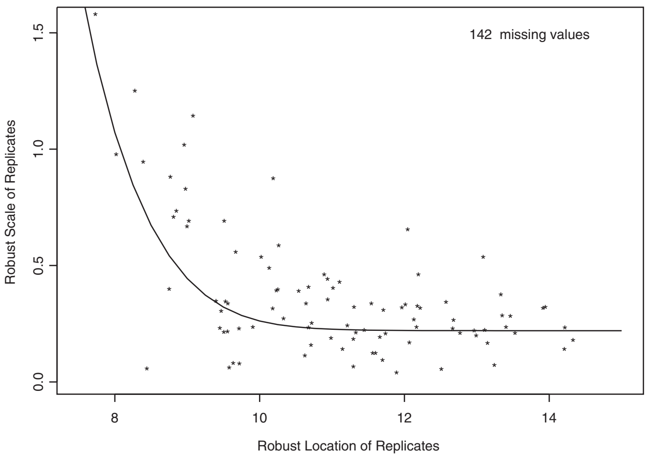

**Fig. 3.** Robust location of replicates versus robust scale of the replicates for data that has been mean centered (background correction) and log transformed. The line is the theoretical scale from the two-component model.

When $\mu$ falls in between these two extremes, all terms in (1) play a significant role. The measured expression $y$ is distributed as a linear combination of a normal and a lognormal random variable and has variance

$$\mathrm{Var}(y) = \mu^2 S_\eta^2 + \sigma_\varepsilon^2. \tag{5}$$

This also depends on $\mu$, but in a more complicated fashion than high-level data. In essence, the distribution of the measurement error changes depending on $\mu$, making the error structure of microarray data quite complicated. Microarray data therefore require transformation before standard statistical methodologies may be applied.

Chen *et al.* (1997), Ideker *et al.* (2000), and Newton *et al.* (2001) have proposed alternative models for the measurement error in microarray data. Chen *et al.* (1997) suggest that the measurement error is normally distributed with constant coefficient of variation (CV). The constant CV assumption is in accord with much experience, but of course cannot be correct for zero or near-zero expression as it would imply negligible measurement error. Ideker *et al.* (2000) introduce a model similar to (1), but with a multiplicative error component that is normally distributed, rather than lognormal. However, plots of the skewness coefficient of replicated observations show that while the replicates are symmetrically distributed about their mean for low expression levels, they display positive skewness above a cutoff point. If the multiplicative error were normally distributed, one would expect to see replicates symmetrically distributed about their mean over the full range of expression. Finally, Newton *et al.* (2001) propose a gamma model for measurement error. We have not compared the performance of this model with (1).

## TRANSFORMING MICROARRAY DATA

### Log transformations

Although microarray data may clearly benefit from transformation, it is not immediately apparent which transformation should be used. Speed (2000) recommends the use of log transformations, but this approach is subject to a number of problems.

First, suppose that the data have been background corrected so that analysis is performed on $\hat{\mu} = y - \hat{\alpha}$, where $\hat{\alpha}$ is an estimate of $\alpha$, the mean background level. Then $\ln(\hat{\mu}) = \ln(y - \hat{\alpha})$ is not defined for $y \leq \hat{\alpha}$, and observations where $y - \hat{\alpha} < 0$ must be removed from the data prior to transformation. However, since $\alpha$ is the mean of the unexpressed genes, $y - \hat{\alpha}$ may be negative for as many as half of the unexpressed genes, which in turn may constitute a large part of the data. While the large quantity of data generated by a single array often means that one may discard half of the data with seeming impunity, this approach is clearly not optimal.

Furthermore, a straightforward delta-method approach shows that the asymptotic variance of $\ln(\hat{\mu})$, $\text{AV}(\ln(\hat{\mu})$, is

$$\text{AV}(\ln(\hat{\mu})) = \sigma_\eta^2 + \frac{\sigma_\varepsilon^2}{\mu^2} \tag{6}$$

which is approximately constant for large $\mu$ but which approaches infinity as $\mu \to 0$ (Rocke and Durbin, 2001). The asymptotic variance of log-transformed background-corrected observations is thus greatly inflated for small $\mu$.

Figure 3 shows robust estimates of the replicate mean and standard deviation of log background-corrected data. The solid line on the plot shows the asymptotic standard deviation, which is the square root of (6). The estimated standard deviation is essentially constant for large mean values, but increases dramatically as the mean approaches 0. Note that the 142 of 276 observations where $y < \hat{\alpha}$ were excluded from this plot.

An alternative approach would be to consider transformations of the form $\ln(\hat{\mu} + c)$, where $c$ is some positive constant. If we let $c = \hat{\alpha}$ we arrive at the transformation $\ln(\hat{\mu} + \hat{\alpha}) = \ln(y - \hat{\alpha} + \hat{\alpha}) = \ln(y)$, which has the advantage of being defined over the full range of the data. However, delta-method calculations show that the asymptotic variance of $\ln(y)$ is

$$\text{AV}(\ln(y)) = \frac{\mu^2\sigma_\eta^2 + \sigma_\varepsilon^2}{(\mu + \alpha)^2} \tag{7}$$

$$\approx \frac{\mu^2\sigma_\eta^2 + \sigma_\varepsilon^2}{y^2}, \tag{8}$$

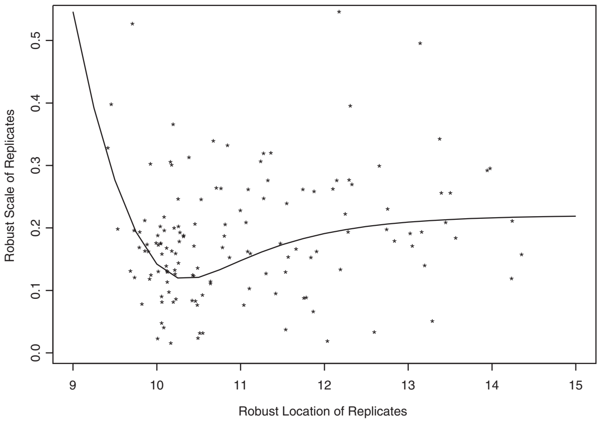

**Fig. 4.** Robust location of replicates versus robust scale of the replicates for data that has been log transformed. The line is the theoretical scale from the two-component model.

which is again approximately constant at high levels but increases as $\mu \to 0$, although not as dramatically as $\text{AV}(\ln(\hat{\mu}))$ (Rocke and Durbin, 2001).

Figure 4 shows robust estimates of the replicate mean and standard deviation of log-transformed uncorrected data. The solid line on the plot shows the asymptotic standard deviation, which is the square root of (8). As mentioned above, the standard deviation is approximately constant for large mean values, but as the mean approaches 0, the standard deviation dips and then increases substantially. The variance, however, appears to be much more constant than for $\log(\hat{\mu})$. Should one insist on using a log transformation, transformation of data that have not been background-corrected appears to give better results.

Though log transformations approximately stabilize the variance of data expressed at high levels, their performance on data at or near background leaves much to be desired. A common approach is to eliminate observations at or near background, but a transformation that would allow all of the data to be used in an analysis would certainly be preferable.

### A variance-stabilizing transformation

Given the shortcomings of the log transformation, one might wish to find a transformation for microarray data which stabilizes the asymptotic variance over the full range of the data. Delta-method calculations allow us to derive such a transformation:

Let $g(\cdot)$ be a smooth function. Then the asymptotic variance of $g(y)$ as $y \xrightarrow{p} \theta$ is

$$\text{AV}(g(y)) = \dot{g}(\theta)^2\text{Var}(y),$$

where $\dot{g}(\theta) = \left.\frac{\partial g}{\partial t}\right|_{t=\theta}$.

**Table 1.** Variance and skewness of simulated data transformed using the variance-stabilizing transformation

| $\mu$ | Variance | 95% CI for Variance | Skewness | 95% CI for Skewness |
|---|---|---|---|---|
| 0 | 0.0530 | (0.0487, 0.0574) | 0.0011 | (−0.1382, 0.1404) |
| 5 000 | 0.0533 | (0.0487, 0.0579) | −0.1308 | (−0.2651, 0.0035) |
| 10 000 | 0.0538 | (0.0491, 0.0584) | −0.2110 | (−0.3493, −0.0726) |
| 15 000 | 0.0541 | (0.0493, 0.0588) | −0.2449 | (−0.4013, −0.0884) |
| 20 000 | 0.0540 | (0.0493, 0.0588) | −0.2401 | (−0.4001, −0.0802) |
| 25 000 | 0.0540 | (0.0490, 0.0590) | −0.2159 | (−0.3767, −0.0550) |
| 30 000 | 0.0537 | (0.0488, 0.0586) | −0.2040 | (−0.3745, −0.0335) |
| 40 000 | 0.0533 | (0.0483, 0.0583) | −0.1489 | (−0.3112, 0.0133) |
| 50 000 | 0.0529 | (0.0483, 0.0574) | −0.1121 | (−0.2755, 0.0512) |
| 60 000 | 0.0524 | (0.0476, 0.0572) | −0.0842 | (−0.2475, 0.0791) |
| 65 000 | 0.0525 | (0.0479, 0.0571) | −0.0764 | (−0.2359, 0.0831) |
| 70 000 | 0.0523 | (0.0475, 0.0571) | −0.0637 | (−0.2205, 0.0931) |
| 80 000 | 0.0521 | (0.0474, 0.0568) | −0.0522 | (−0.2091, 0.1048) |
| 100 000 | 0.0519 | (0.0474, 0.0565) | −0.0347 | (−0.1908, 0.1215) |
| 200 000 | 0.0517 | (0.0471, 0.0562) | −0.0112 | (−0.1732, 0.1508) |
| 300 000 | 0.0515 | (0.0470, 0.0561) | −0.0028 | (−0.1548, 0.1493) |
| 500 000 | 0.0516 | (0.0470, 0.0562) | −0.0008 | (−0.1586, 0.1571) |
| 1 000 000 | 0.0515 | (0.0470, 0.0560) | 0.0027 | (−0.1490, 0.1545) |

Suppose we wish to find a transformation $g(\cdot)$ for $y = \alpha + \mu e^{\eta} + \varepsilon$ such that $\mathrm{AV}(g(y))$ is constant. Setting

$$\mathrm{AV}(g(y)) = \dot{g}(\mu + \alpha)^2 \mathrm{Var}(y) = k,$$

where $k$ is some constant, and solving for $g(\cdot)$ we find

$$\begin{aligned} \dot{g}(\mu+\alpha)^2 &= \frac{k}{\mathrm{Var}(y)} \\ &= \frac{k}{\mu^2 S_\eta^2 + \sigma_\varepsilon^2} \\ \Longleftrightarrow \quad \dot{g}(\mu+\alpha) &= \frac{k}{\sqrt{\mu^2 S_\eta^2 + \sigma_\varepsilon^2}} \\ \Longleftrightarrow \quad \dot{g}(y) &= \frac{k}{\sqrt{(y-\alpha)^2 S_\eta^2 + \sigma_\varepsilon^2}} \\ \Longleftrightarrow \quad \int \dot{g}(y)dy &= \int \frac{k}{\sqrt{(y-\alpha)^2 S_\eta^2 + \sigma_\varepsilon^2}} dy. \end{aligned}$$

One solution is

$$g(y) = \ln(y - \alpha + \sqrt{(y-\alpha)^2 + c}), \tag{9}$$

where $c = \frac{\sigma_\varepsilon^2}{S_\eta^2}$. This transformation, which was first introduced by Hawkins (2001) in the context of another application, exactly stabilizes the asymptotic variance of data distributed according to (1), making the asymptotic variance of the transformed data equal to $S_\eta^2$.

This transformation is defined and monotonically increasing for all values of $y$, positive or negative. It is approximately the natural logarithm for large values of $y$ and is approximately linear at $y = 0$.

*Performance on simulated data* The variance-stabilizing transformation was tested on data simulated from (1), for values of $\mu$ ranging from 0 to 1 000 000. For each value of the true expression level $\mu$, 1000 samples of size 1000 were simulated. The parameters used were $\alpha = 24\,800$, $\sigma_\eta = .227$, and $\sigma_\varepsilon = 4800$, which were the values estimated from the control example data. The data were then transformed using (9), where $c = \frac{\sigma_\varepsilon^2}{S_\eta^2} = 413\,822\,950$. For each transformed sample, the sample variance and sample skewness were calculated, and the mean and standard deviation of these quantities over all 1000 samples were then used to create asymptotically-normal 95% confidence intervals.

Table 1 shows the variance and skewness of the transformed simulated data for various values of $\mu$, along with confidence intervals for these quantities. Tables 2 and 3 show, for the purposes of comparison, estimates of these statistics for the same simulated data following transformation using $\log(\hat{\mu})$ and $\log(y)$, respectively. (Negative observations were removed from the data prior to log transformation).

For the data transformed using the variance-stabilizing transformation, the confidence intervals for the variance all include $S_\eta^2 = 0.0557$, indicating that the variance has been stabilized across the full range of the data. In contrast, the variance of log background-corrected data

**Table 2.** Variance and skewness of $\log(\hat{\mu})$, simulated data

| $\mu$ | Variance | 95% CI for Variance | Skewness | 95% CI for Skewness |
|---:|---:|---:|---:|---:|
| 0 | 1.2348 | (0.9730, 1.4967) | −1.4774 | (−1.9993, −0.9555) |
| 5 000 | 0.8526 | (0.6879, 1.0174) | −1.8583 | (−2.4329, −1.2837) |
| 10 000 | 0.4648 | (0.3613, 0.5683) | −2.2024 | (−3.1295, −1.2753) |
| 15 000 | 0.2394 | (0.1839, 0.2948) | −2.0297 | (−3.2916, −0.7679) |
| 20 000 | 0.1430 | (0.1175, 0.1685) | −1.3945 | (−2.5418, −0.2472) |
| 25 000 | 0.1048 | (0.0884, 0.1212) | −0.9218 | (−1.8641, 0.0205) |
| 30 000 | 0.0865 | (0.0761, 0.0968) | −0.6551 | (−1.1323, −0.1779) |
| 40 000 | 0.0703 | (0.0630, 0.0775) | −0.3748 | (−0.5970, −0.1525) |
| 50 000 | 0.0632 | (0.0575, 0.0690) | −0.2512 | (−0.4429, −0.0595) |
| 60 000 | 0.0594 | (0.0539, 0.0650) | −0.1784 | (−0.3582, 0.0013) |
| 65 000 | 0.0584 | (0.0532, 0.0636) | −0.1561 | (−0.3302, 0.0180) |
| 70 000 | 0.0574 | (0.0521, 0.0627) | −0.1313 | (−0.2966, 0.0340) |
| 80 000 | 0.0559 | (0.0508, 0.0611) | −0.1028 | (−0.2660, 0.0603) |
| 100 000 | 0.0544 | (0.0495, 0.0592) | −0.0669 | (−0.2263, 0.0926) |
| 200 000 | 0.0523 | (0.0476, 0.0569) | −0.0190 | (−0.1816, 0.1436) |
| 300 000 | 0.0518 | (0.0472, 0.0564) | −0.0062 | (−0.1586, 0.1461) |
| 500 000 | 0.0517 | (0.0471, 0.0563) | −0.0020 | (−0.1599, 0.1559) |
| 1 000 000 | 0.0515 | (0.0470, 0.0560) | 0.0024 | (−0.1493, 0.1542) |

**Table 3.** Variance and skewness of $\log(y)$, simulated data

| $\mu$ | Variance | 95% CI for Variance | Skewness | 95% CI for Skewness |
|---:|---:|---:|---:|---:|
| 0 | 0.0418 | (0.0372, 0.0464) | −0.6867 | (−0.9944, −0.3791) |
| 5 000 | 0.0294 | (0.0263, 0.0324) | −0.5400 | (−0.7631, −0.3170) |
| 10 000 | 0.0246 | (0.0221, 0.0270) | −0.4335 | (−0.6316, −0.2353) |
| 15 000 | 0.0228 | (0.0207, 0.0250) | −0.3324 | (−0.5251, −0.1398) |
| 20 000 | 0.0225 | (0.0205, 0.0245) | −0.2350 | (−0.4160, −0.0540) |
| 25 000 | 0.0230 | (0.0208, 0.0251) | −0.1515 | (−0.3229, 0.0199) |
| 30 000 | 0.0237 | (0.0215, 0.0259) | −0.1039 | (−0.2787, 0.0709) |
| 40 000 | 0.0256 | (0.0232, 0.0280) | −0.0161 | (−0.1779, 0.1456) |
| 50 000 | 0.0275 | (0.0251, 0.0298) | 0.0275 | (−0.1331, 0.1881) |
| 60 000 | 0.0292 | (0.0265, 0.0318) | 0.0529 | (−0.1086, 0.2145) |
| 65 000 | 0.0301 | (0.0275, 0.0327) | 0.0584 | (−0.0972, 0.2140) |
| 70 000 | 0.0308 | (0.0280, 0.0336) | 0.0677 | (−0.0883, 0.2236) |
| 80 000 | 0.0322 | (0.0293, 0.0351) | 0.0716 | (−0.0832, 0.2264) |
| 100 000 | 0.0346 | (0.0316, 0.0376) | 0.0766 | (−0.0778, 0.2310) |
| 200 000 | 0.0412 | (0.0375, 0.0448) | 0.0578 | (−0.1029, 0.2185) |
| 300 000 | 0.0440 | (0.0402, 0.0479) | 0.0466 | (−0.1040, 0.1972) |
| 500 000 | 0.0468 | (0.0427, 0.0509) | 0.0307 | (−0.1264, 0.1877) |
| 1 000 000 | 0.0490 | (0.0447, 0.0533) | 0.0191 | (−0.1322, 0.1705) |

decreases steadily as $\mu$ increases, while the variance of the log uncorrected data decreases initially and then increases again before stabilizing at around $\mu = 200\,000$. It should be noted that the log uncorrected data appear to perform nearly as well as the variance-stabilized data, and that these data have a much more constant variance than the log background-corrected data. However, the variance-stabilizing transformation produces transformed observations with the most constant error variance, as would be expected. Furthermore, the intensity values have no natural zero scale, and may have already been 'corrected' for pixel background, so that the good behaviour of the log intensity may be dependent on the zero scaling of the intensity. If the log intensity transformation is to be used, the behaviour over the range of the data needs to be examined by producing a fitted variance curve like the one in Figure 3.

Confidence intervals for skewness from Table 1 indicate that the variance-stabilizing transformation symmetrizes the data as well, except for values of $\mu$ between 10 000

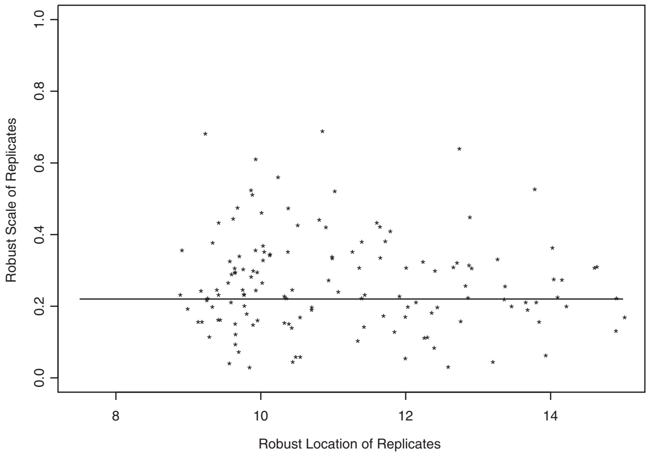

**Fig. 5.** Robust location of replicates versus robust scale of the replicates for data that has been transformed using the new procedure outlined in this paper. The line is the theoretical scale from the two-component model.

and 30 000, where the data are slightly skewed to the left. (Outside of this range, the confidence intervals include 0, which is the skewness of data distributed symmetrically about its mean.) This result appears on first examination to be equivalent to that for the log uncorrected data, which are symmetric except for values of $\mu$ between 0 and 20 000, where they are left-skewed. However, the degree of skewness is rather less for the variance-stabilized data. Both transformations exhibit much greater symmetry than the log background-corrected data, which is left-skewed for $\mu < 50\,000$.

*Performance on microarray data* Figure 5 shows robust estimates of the replicate mean and standard deviation of data transformed using the variance-stabilizing transformation. The parameters $\sigma_\varepsilon^2$ and $\sigma_\eta^2$, which are needed to calculate the shift constant *c*, were estimated using the procedure described in Rocke and Durbin (2001). (It should be noted that, although $\sigma_\varepsilon^2$ may be estimated from unreplicated data, replicated observations from several genes expressed well above background are needed to estimate $\sigma_\eta^2$.) Notice that the standard deviation remains constant across the full range of the data. The asymptotic standard deviation, which is the square root of (9), is shown on the plot as a solid horizontal line. The variance having been approximately stabilized, further analysis may now be performed on the these data without needing to cull any of the observations.

## CONCLUSION

Microarray data, with their complicated error structure, need to be transformed prior to analysis using standard statistical methods. Log transformations provide good variance stabilization at high levels, but inflate the variance of near-background observations, particularly in data that have been background-corrected. We have introduced a transformation which stabilizes the asymptotic variance of microarray data across the full range of the data, as well as making the data more symmetric. This allows further analysis to be performed on these data without violation of assumptions and without needing to remove low-level observations.

## ACKNOWLEDGEMENTS

The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P43 ES04699).

## REFERENCES

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

Bartosiewcz,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.

Box,G.E.P. and Cox,D.R. (1964) An analysis of transformations. *J. Roy. Stat. Soc.* Series B *(Methodological)*, **26**, 211–252.

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

Hawkins,D. (2001) Diagnostics for conformity of paired quantitative measurements. *Stat. Med.*, in press

Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2001) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Newton,M.A., Kendziorski,C.M., Richmond,C.S., Blattner,F.R. and Tsui,K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

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

Speed,T. (2001) Always log spot intensities and ratios. *Speed Group Microarray Page*,http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html