BIOINFORMATICS

Vol. 19 no. 14 2003, pages 1817–1823
DOI: 10.1093/bioinformatics/btg245

Transformation and normalization of oligonucleotide microarray data

Sue C. Geller[1,*], Jeff P. Gregg[2], Paul Hagerman[3] and David M. Rocke[4]

[1]Department of Mathematics, Texas A&M University, College Station, TX 77843–3368, USA, [2]Department of Pathology, [3]Department of Biological Chemistry, School of Medicine, and [4]Department of Applied Science and Division of Biostatistics, University of California, Davis, CA 95616, USA

Received on July 23, 2002; revised on February 17, 2003; accepted on April 9, 2003

ABSTRACT

Motivation: Most methods of analyzing microarray data or doing power calculations have an underlying assumption of constant variance across all levels of gene expression. The most common transformation, the logarithm, results in data that have constant variance at high levels but not at low levels. Rocke and Durbin showed that data from spotted arrays fit a two-component model and Durbin, Hardin, Hawkins, and Rocke, Huber *et al*. and Munson provided a transformation that stabilizes the variance as well as symmetrizes and normalizes the error structure. We wish to evaluate the applicability of this transformation to the error structure of GeneChip microarrays.

Results: We demonstrate in an example study a simple way to use the two-component model of Rocke and Durbin and the data transformation of Durbin, Hardin, Hawkins and Rocke, Huber *et al*. and Munson on Affymetrix GeneChip data. In addition we provide a method for normalization of Affymetrix GeneChips simultaneous with the determination of the transformation, producing a data set without chip or slide effects but with constant variance and with symmetric errors. This transformation/normalization process can be thought of as a machine calibration in that it requires a few biologically constant replicates of one sample to determine the constant needed to specify the transformation and normalize. It is hypothesized that this constant needs to be found only once for a given technology in a lab, perhaps with periodic updates. It does not require extensive replication in each study. Furthermore, the variance of the transformed pilot data can be used to do power calculations using standard power analysis programs.

Availability: SPLUS code for the transformation/normalization for four replicates is available from the first author upon request. A program written in C is available from the last author.

Contact: geller@math.tamu.edu

1 INTRODUCTION

Appropriate ways of preprocessing microarray data before analysis is a topic of continuing interest and discussion. Among the reasons for such preprocessing are normalization, i.e. removing slide/chip effects, background intensities, and other sources of systematic error, and transforming the data so that assumptions needed for the analysis are met. The order in which normalization and transformation are performed varies as well, but the two topics are usually treated separately.

It was shown in Rocke and Durbin (2001) that the data from spotted microarrays conformed to a two-component model of error, i.e. $y = \alpha + \mu e^{\eta} + \epsilon$, where y is the measured intensity, μ is the (unknown) expression level in arbitrary units, and η and ϵ are normally distributed error terms with mean 0 and variance σ_η^2 and σ_ϵ^2, respectively. Note that the additive error dominates when μ is small, and the proportional error dominates when μ is large. They described a procedure by which to estimate σ_η and σ_ϵ, the former being approximately the standard deviation of the high level log-transformed data and whose estimation requires biologically constant replication of at least one sample. Durbin *et al*. (2002), Huber *et al.* (2002), and Munson (2001) independently developed a data transformation to stabilize the variance for data that fit the two-component model, namely, $z = \ln\left[(y-\alpha) + \sqrt{(y-\alpha)^2 + c}\right]$, where c is a constant to be determined from the data, specifically $c = \hat{\sigma}_\epsilon^2/\hat{\sigma}_\eta^2$. They noted that the commonly used log transform stabilizes the variance for highly expressed genes but not for genes with low or zero expression, and showed that the transformation was effective for spotted microarray data. The transformed data not only has constant variance but the errors and transformed data are approximately symmetric.

Many methods of normalization, such as subtraction of an estimated background intensity and the commonly used average difference score for Affymetrix GeneChips, can result in negative measured intensities. This can also happen with other normalization methods such as simple ANOVA or

[*] To whom correspondence should be addressed.

regression techniques, or the program dChip, which uses an iterative procedure with regression and outlier detection (`http://www.dchip.org;` Li and Wong, 2001). The negative values for intensities (e.g. when MM $>$ PM) cause problems when the commonly used log transform is applied to the normalized data. Also, if the variance is not constant across expression levels, then the mean chip/slide intensity is not an optimal measure of the overall level of a given chip/slide.

We will demonstrate that data from Affymetrix GeneChips conform at least approximately to the two-component model of error. Then we shall provide a procedure by which to find the transformation constant c and normalize the data at the same time. This simultaneous procedure is necessary under the assumptions of our model. This method avoids many of the problems of alternative normalization and transformation methods. In particular there is no problem with negative observed intensities.

2 DATA

In order to investigate replicability and the appropriateness of the two-component model of error for data from Affymetrix chips, a single lymphoblastoid cell line from one autistic child was grown in four separate T75 flasks. The cells were split from the same parent flask and grown in the same incubator, same shelf, at the same time. Four separate RNA extractions were performed when the cells were near confluence. cDNA synthesis and *in vitro* transcription (IVT) labeling were performed on each sample. Each sample was then hybridized to an Affymetrix Human Genome U95Av2 oligonucleotide GeneChip array, which contains 12 625 probe sets, with each probe set designed to represent a single human gene. Note that these are true replicates of the whole measurement process, not just machine replicates in which the same sample after processing is divided and hybridized to several different chips.[†]

This study is part of a larger project to investigate the genetic basis of autism, a behavioral diagnosis that reflects complex and ill-defined patterns of inheritance and possible environmental factors. Although autism is highly heritable, with over 90% concordance for autism spectrum disorders, genome-wide linkage studies have been unable to define specific chromosomal loci for the preponderance of autistic individuals, excluding known genetic disorders [e.g. Rett syndrome, fragile X syndrome, tuberous sclerosis, 15q(dup)] that have autistic features as a variable component of their phenotypes. An alternative approach to the identification of genetic components of autism, including those conferring differential susceptibility to putative environmental factors, is the identification of patterns of altered gene expression through microarray methods. This approach entails the analysis of blood and tissue samples from autistic patients and unaffected family members with the objective of identifying patterns of gene expression that associate with the autism spectrum phenotype. Information derived from such investigations would be of critical importance for both predictive efforts (which children are at increased risk to develop autism) and efforts to define the various biochemical/genetic mechanisms that lead to autism.

All graphs and computed statistics in this paper are from the data set of Perfect Match (PM)/ Mismatch (MM) differences for 12 625 genes for four replicates or flasks. We used the average difference (PM–MM) data as they are reported by Microarray Suite 4.0 (MAS 4.0; Affymetrix Inc.) without using their normalization program. Version 4 reports negative values when the MM is greater than the PM.[§]

Since the standard deviation of the low expression level data is approximately 8, we considered genes for which the median value of the four replicates was less than or equal to -25 to be unusable outliers,[¶] and eliminated those 655 genes. In general the data can be thought of as an $n \times m$ array of intensity levels where n is the number of genes, m the number of slides or chips analyzed, and $n \gg m$, $11\,970 \times 4$ in our example data. Note that in general the estimates would not be changed had we left these data in, but the plots would have been more difficult to interpret.

3 TWO-COMPONENT MODEL

Figure 1a–c show that the data fit the two-component model. Figure 1a is a graph of the median of the four replicates (one for each of the four chips) for each of the 11 970 genes by the interquartile range of the four replicates for that gene. Figure 1b is the same plot except that the medians are restricted to those under 1000, eliminating 372 points. Figure 1c is the same plot as 1a but with the ordinate being the ranks of the medians so that the low level data are more visible. The solid lines are the loess smooths of the scatter plot, whose slope in Figure 1a for large positive intensities shows that the scale of the error is linear in the location, as would be expected. Figure 1b expands out the region containing over 96% of the

[†] There are a number of differing types of replicates depending on where in the process one starts. A fundamental assumption of our model is that the replicates used to estimate it come from the same biological sample for all replicates in a group. Thus, we are considering measurement error, not variability within and between organisms. The samples we are using for this paper all come from the same cell line. We are measuring, in this case, all of the sources of error beginning with the collection of the cells, through hybridizing to the chip and reading the results.

[§] Version 5 avoids negative average difference scores by artificially constraining the results to be positive. It is not yet clear whether this makes analysis more or less difficult.

[¶] Under the usual assumptions about PM–MM, the expected value should never be less than zero. Thus, very few data points should be less than three standard deviations below zero, and even fewer medians of four should be less than this cutoff. By eliminating data in which the median PM–MM is less than -25, we are eliminating data that do not conform the expected behavior for expressed or unexpressed genes.

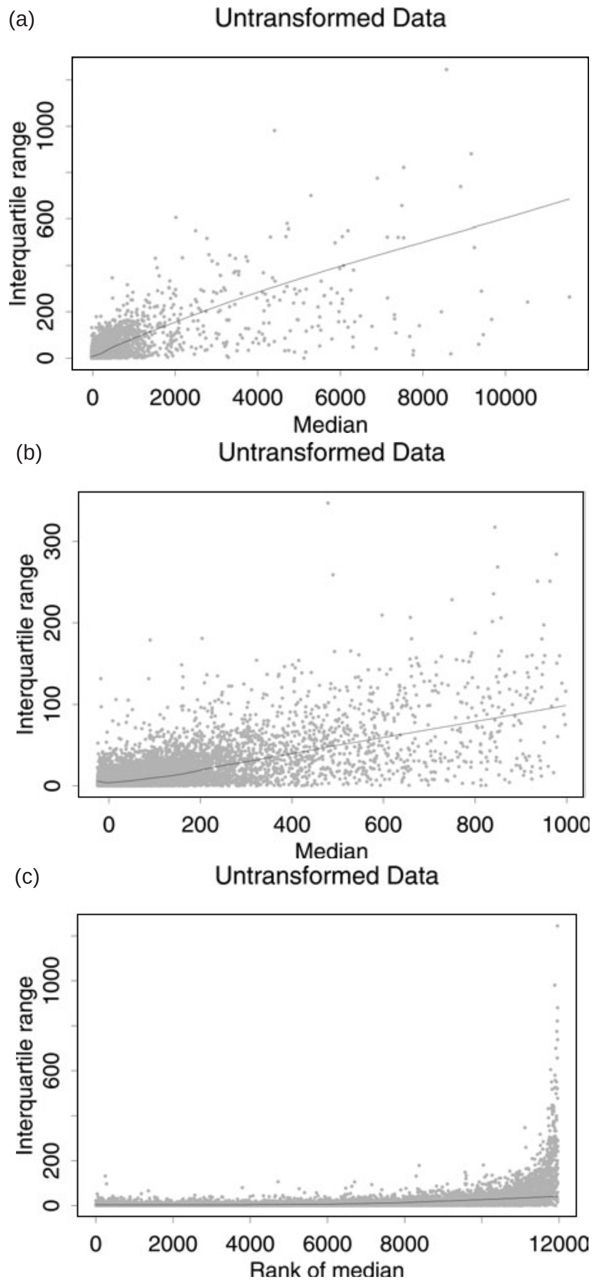

Fig. 1. **(a)** Median measured intensity versus interquartile range for the raw (or untransformed) data. The line is a loess smooth with degree 1 and span 0.1. **(b)** Median measured intensity versus interquartile range for the raw (or untransformed) data for medians less than 1000. The line is a loess smooth with degree 1 and span 0.1. **(c)** The rank of the median measured intensity versus IQR for the raw (or untransformed) data. The line is a loess smooth with degree 1 and span 2/3.

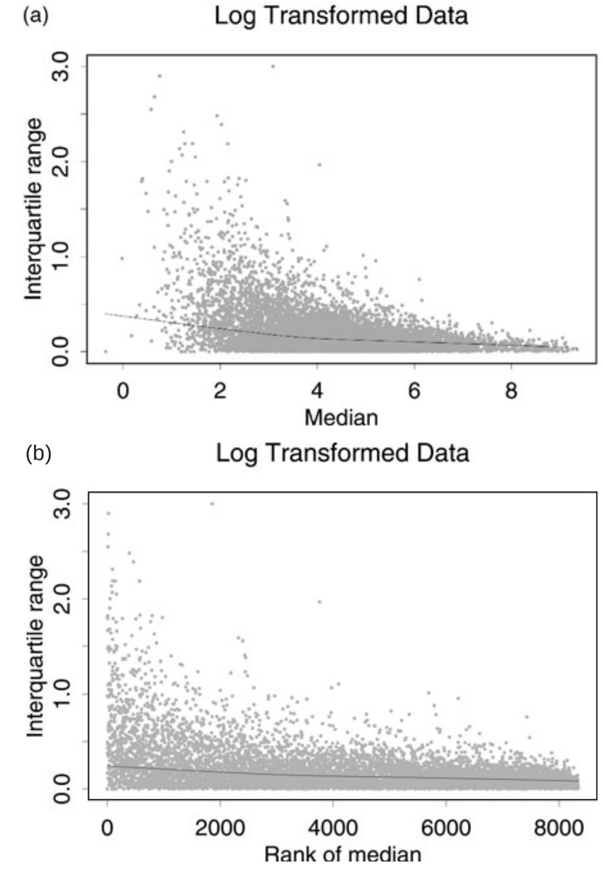

Fig. 2. **(a)** Median measured intensity versus interquartile range for the data transformed using the natural log on genes for which all four intensities are positive. The line is a loess smooth with degree 1 and span 2/3. NOTE: Only 8339 of 12 625 genes (or the 11 970 in the cleaned data set) could be so transformed. **(b)** The rank of the median measured intensity versus interquartile range for the data transformed using the natural log on genes for which all four intensities are positive. The line is a loess smooth with degree 1 and span 2/3. NOTE: Only 8339 of 12 625 genes (or 11 970 in the cleaned data set) could be so transformed.

data to see that the two component model is valid in that region. Figure 1c shows the constant variance for intensities near 0. Degree 1 and span $= 0.1$ were used for Figure 1a and b because of the small number of points on the right-hand portion of the graph. Degree 1 and span $= 2/3$, the default values in SPLUS, were used for all other graphs, but there is little difference in the smoothing lines using degree 2 or a smaller span. Note that we use median and interquartile range (IQR) instead of mean and standard deviation for robustness against outliers, which are not uncommon in this type of data. With four observations, we use the difference of the middle two numbers as the IQR. The Appendix contains definitions of the finite-sample IQR for sample sizes up to 20 and the required constants for estimating the standard deviation from the IQR.

Figure 2a and b are similar scatter plots for the natural log transformed data. Note that the variance is approximately constant only for large expression levels and that there are many outliers or high variance data at lower levels. Furthermore, only 8339 of the 12 625 genes (or of the 11 970 genes of the data subset) were usable as the other approximately 30% of the data had one or more of the replicates with non-positive values. Omission of such a large amount of data is not optimal. Even if it is considered that these data are unimportant in this sample, since the genes are not expressed at a high level, comparison of different conditions is made much more difficult if a gene is expressed at a low level in one sample and a high level in another. Simply replacing negative estimated expression values by some arbitrary positive value such as 10 or 20, or by an estimated mean, is also non-optimal since it distorts the variability pattern of the data.

4 TRANSFORMATION AND NORMALIZATION

The approach we pursue is to formulate a model that fits the variance patterns in the data, and that contains normalization constants, and then use a procedure that can simultaneously determine the transformation parameter and the normalization. The statistical model used is that, for gene i and chip j, the measured (average difference) value y_{ij} satisfies

$$f_c(y_{ij}) = \mu_i + n_j + \epsilon_{ij},$$

where μ_i is the true expression of gene i in the sample, n_j is an additive chip normalization, and ϵ_{ij} is an additive symmetric measurement error.

The back-transformed, normalized measurements

$$\tilde{y}_{ij} = f_c^{-1}[f_c(y_{ij}) - n_j] + \alpha$$

are assumed to fit the two-component model, so that

$$V(\tilde{y}_{ij}) = \sigma_\epsilon^2 + S_\eta^2 \mu_i^2.$$

In this model, $S_\eta^2 = \mathrm{e}^{\sigma_\eta^2}(\mathrm{e}^{\sigma_\eta^2} - 1)$, which is to the first order the same as σ_η^2. For example, if $\sigma_\eta = 0.1$, corresponding to a 10% coefficient of variation, then $S_\eta = 0.1008$.

The transformation used is

$$f_c(y) = \ln\left[y - \alpha + \sqrt{(y-\alpha)^2 + c}\right],$$

where $c = \sigma_\epsilon^2 / S_\eta^2$, or equivalently to the first order $c = \sigma_\epsilon^2/\sigma_\eta^2$. Note that we could treat α as negligible given that the primary data are PM–MM, which can be presumed to have a mean near zero for unexpressed genes, but actually estimated all the values, including α. We will briefly examine the result of keeping $\alpha = 0$. We can estimate σ_ϵ^2 from the low level data and σ_η^2 from the high level data, as in Rocke and Durbin (2001), but the values of $\hat{\sigma_\eta}$, $\hat{\sigma_\epsilon}$, and c are interdependent in the sense that we cannot determine $\hat{\sigma_\eta}$ and $\hat{\sigma_\epsilon}$ without knowing c, and vice versa. We solve this problem using the following iteration:

1. Starting with a trial value of c, transform the data using

$$f_c(y_{ij}) = \ln\left[y_{ij} - \hat{\alpha} + \sqrt{(y_{ij} - \hat{\alpha})^2 + c}\right].$$

(We could use $\hat{\alpha} = 0$ for these average difference data or estimate it.)

2. Determine the normalization constants n_j by taking the median of all genes over each slide minus the median of all the genes on all the slides. Then, the transformed, normalized data are

$$f_c(y_{ij}) - n_j.$$

3. Back-transform the normalized transformed values using the inverse transformation

$$f_c^{-1}(z) = \frac{\mathrm{e}^z - c\mathrm{e}^{-z}}{2} + \hat{\alpha}.$$

4. Determine the parameters σ_ϵ and σ_η of the two-component model by pooling low and high levels of expression as given in the Appendix and re-estimate α, a step not necessary when assuming that $\alpha = 0$. We use a fixed number of genes at the high end, and at the low end use all genes whose medians are sufficiently close to α.
5. Determine a new transformation parameter $c = \sigma_\epsilon^2/\sigma_\eta^2$, and return to step 1 (using the original data not the back-transformed data).
6. Stop the process when the parameter values stop changing.

For this application, we did not assume that the average difference scores of unexpressed genes average out to near zero, so that α is estimated, rather than assumed to be zero. We check out the effects of assuming $\alpha = 0$ below. We estimated σ_η by pooling IQRs of the log transformed data corresponding to the highest h of the medians of the log transformed data (see appendix). Also note that we cannot determine the transformation accurately without the iterative normalization. In our data the flask effects (flask median minus overall median) on the transformed scale ranged from 1/4 to 3/4 of a robust estimate of the standard deviation of the errors, so would have affected our computation of c.

The above procedure was applied to the autism data. It converged after 25 iterations to $c = 4300$ (from $\sigma_\epsilon = 8.20$ and $\sigma_\eta = 0.125$ with $\alpha = 2.74$ and $h = 125$). Figure 3 is a graph of the transformation, $f_{4300}(y) = \ln\left[y - 2.74 + \sqrt{(y - 2.74)^2 + 4300}\right]$, over the range of the data with a graph (dotted line) of $z = \ln[2(y - 2.74)]$, the equivalent log transform, for comparison. Figure 4a and b

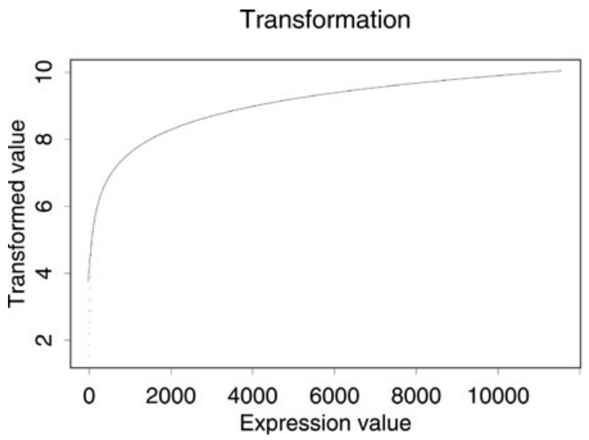

Fig. 3. The transformation $y = \ln\left[y - 2.74 + \sqrt{(y-2.74)^2 + 4300}\right]$ over the range of the untransformed data. The dotted line is $\ln[2(y - 2.74)]$, the equivalent log transformation.

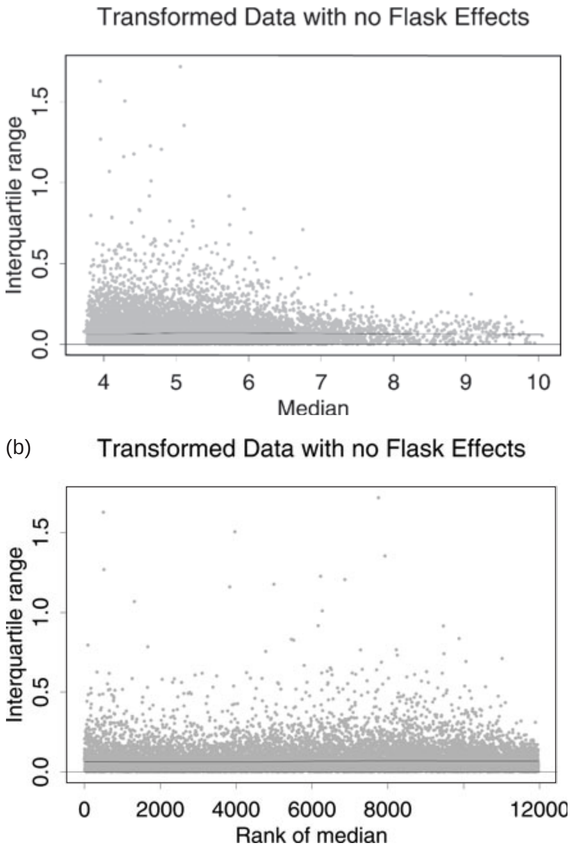

Fig. 4. **(a)** Median measured intensity versus IQR for the data transformed using the procedure outlined in this paper, including the removal of the chip effect. The line is a loess smooth with degree 1 and span 2/3. **(b)** The rank of the median measured intensity versus IQR range for the data transformed using the procedure outlined in this paper, including the removal of the chip effect. The line is a loess smooth with degree 1 and span 2/3.

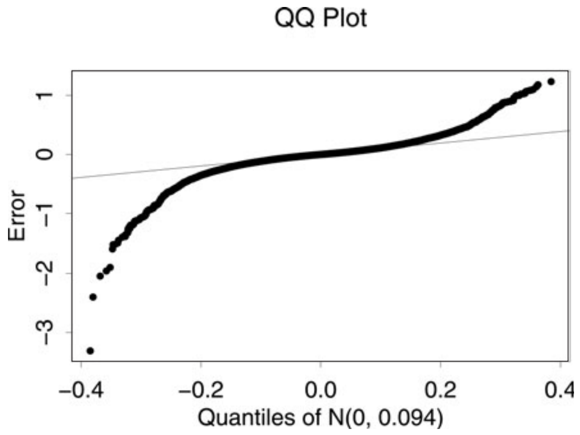

Fig. 5. QQ plot of $N(0, 0.094)$ and the errors of the data transformed by the procedure outlined above. The standard deviation (0.094) is the robust estimate of the standard deviation of the errors (IQR of the errors/1.349).

are scatter plots of the median and IQR of the transformed data ($c = 4300$) from which the flask effects had first been subtracted. Note that the loess lines in both plots are very close to constant. The errors are symmetric but long-tailed compared to a normal distribution, as is shown in the QQ plot in Figure 5.

We also examined the stability of the solutions to starting points and parameters of the process. Whether α is estimated or assumed to be zero, given the granularity of the data and the nature of the median and the IQR, there are often two solutions that differ by a negligible amount as long as the same value of h is used and depending on whether the starting value for α is lower or higher than the final estimated value. The procedure is not sensitive to the initial values of c and ks. The only factor that has a significant effect is h, probably due to the rapid fall in measured intensity at the high levels. As noted above, when $h = 125$, we found $c = 4300$ (from $\sigma_\epsilon = 8.20$ and $\sigma_\eta = 0.125$ with $\alpha = 2.74$). When $h = 200$, we found $c = 4900$ (from $\sigma_\epsilon = 8.22$, and $\sigma_\eta = 0.117$ with $\alpha = 2.73$). If α is assumed to be zero, then the algorithm converges more quickly, taking only seven or eight iterations, but to essentially the same values ($c = 4400$ when $h = 125$ and $c = 5000$ when $h = 200$). It is important to note here that precise estimation of the transformation parameter is not necessary, since transformations with similar parameters can be indistinguishable from each other. Since the transformation constant is not of direct interest in itself, but just a way to stabilize the variance, there is little need for confidence intervals. [See Durbin and Rocke (2003) for confidence intervals for the transformation constant computed using maximum likelihood estimation.]

The linearity of the chip effect, η_j, on the transformed data was checked using Tukey's ODOFFNA (one degree of freedom for non-additivity) method [Tukey (1949)], which corresponds to adding a term to the model $y_{ij} = \mu + \mu_i + n_j + e_{ij}$

which is the product of the row and column deviations from the mean. Then a regression of the residuals of the additive model on $(\mu_i)(n_j)$ is computed. The resulting regression R^2, an estimate of the fraction of variation of the residuals accounted for by non-linearity of the chip effect, is 0.02, demonstrating that the chip effect is linear on the transformed data for all practical purposes.

5 POWER CALCULATIONS

Since the transformed data are approximately symmetric, with approximately constant variance across intensity levels, one can use the variance of the transformed data and information on variability between biological specimens (e.g. from pilot data) to do power calculations using standard software such as that at `www.swogstat.org/stat/public/default.htm` or using packages such as nQuery. Note that there could be other (biological) factors affecting the variance. Rocke (2003, See http://www.cipic.ucdavis.edu/∼dmrocke/preprints.html for a preprint, submitted for publication) contains a method of analyzing the data taking any remaining non-systematic variance heterogeneity into account.

6 CONCLUSION

The iterative technique produces a transformed data set with the following nice properties:

- constant variance of replicates for different genes,
- symmetric errors, and
- no systematic differences from slide to slide.

The normalization and transformation process adds 1–3 chips to an experiment, since biologically identical cells must be used as replicates for the calibration. However, our process removes the requirement of duplicating every sample, a time and money saver. In the absence of appropriate replicate data, it is likely that a value of c near 5000 will provide a substantial improvement of the constancy of variance of any experiment using the same chips and the same probe summary method. The transformed data are appropriate to use in analysis employing both non-parametric and parametric methods. Furthermore, standard software can then be used for power calculations.

It is possible that one can do even better beginning with the probe-level data, however the present method seems satisfactory and can be applied to one of the standard summary methods for Affymetrix data.

ACKNOWLEDGEMENTS

The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172), the National Institute of Environmental Health Sciences, National Institutes of Health (5-P42 ES04699-16 with funding provided by EPA and 1-P01-ES11269), and by the UC Davis MIND Institute Biomarkers Initiative. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH, EPA, or NSF. We are grateful for helpful comments from three referees that improved the substance and presentation of the paper.

REFERENCES

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.

Durbin,B.P. and Rocke,D.M. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **19**, 966–972.

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.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

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.M. (1992) $\bar{X}_Q$ and R_Q charts: robust control charts. *The Statistician*, **41**, 97–104.

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

Tukey,J.W. (1949) One degree of freedom for nonadditivity. *Biometrics*, **5**, 232–242.

APPENDIX: FITTING THE TWO-COMPONENT MODEL

In this paper, we use a variant of the procedure of Rocke and Durbin (2001), the main differences being the use of the median and IQR instead of the mean and standard deviation. We use a fixed number of genes to pool at the high and use a number of genes at the low end that should encompass most of the unexpressed genes. The following is description of the procedure to estimate $\hat{\alpha}$, σ_ϵ, and σ_η after an initial estimate of c has been used to transform the data and remove flask effects.

Let $\tilde{y}_{ij}$ be the normalized, back-transformed value for gene i $(1 \le i \le n)$ on chip j $(1 \le j \le m)$ as input for step 4 of the algorithm given in Section 4. Let M_i be the median over the m replicates of the values for gene i; let $\tilde{M}_i$ be the median of the natural logarithms of the replicates [which may differ from $\ln(M_i)$ if m is even] when all replicates are positive; and let s_i be an estimate of the scale of the replicates for gene i based on the IQR and defined as follows:

- Let $z_{(1)}, z_{(2)}, \ldots, z_{(m)}$ be the replicates for a gene sorted from smallest to largest.

Table A.1. Constants for determination of estimates of scale based on the IQR

m	a	b	d_2^Q
2	1	2	1.1284
3	1	3	1.6926
4	2	3	0.5940
5	2	4	0.9900
6	2	5	1.2835
7	2	6	1.5147
8	3	6	0.9456
9	3	7	1.1439
10	3	8	1.3121
11	3	9	1.4577
12	4	9	1.0737
13	4	10	1.2057
14	4	11	1.3235
15	4	12	1.4298
16	5	12	1.1400
17	5	13	1.2389
18	5	14	1.3296
19	5	15	1.4132
20	6	15	1.1806

- Let $s_i = (z_{(b)} - z_{(a)})/d_2^Q$, where a, b, and d_2^Q are given in Table A.1. If $m > 20$, use $a = \lceil m/4 \rceil$, $b = m - a + 1$, and $d_2^Q = 1.35$. Alternatively, more exact values for d_2^Q for $m > 20$ can be derived from tables of normal order statistics as described in Rocke (1992).

Similarly, let $\tilde{s}_i$ be an estimate of the scale of the logarithms of the replicates defined as for s_i. If there are differing numbers of replicates due to missing values or other considerations, one uses the actual number of replicates for each group. This allows the method to cope with missing values.

Let $h \ll n/2$ and $k > 0$. We chose k for this exercise to be 1.645 times the standard error of the median of four observations from a normal distribution; let H be the set of indices i such that $\{\tilde{M}_i | i \in H\}$ are the h largest medians of the logarithms of the data, and let L be the set of indices i such that $\{|M_i - \hat{\alpha}| < k\hat{\sigma_\epsilon} | i \in L\}$, where $\hat{\sigma_\epsilon}$ is the value from the previous iteration. Our estimates of the two-component model parameters are

$$\hat{\alpha} = \ell^{-1} \sum_{i \in L} M_i, \tag{A.1}$$

$$\hat{\sigma}_\epsilon = \ell^{-1} \sum_{i \in L} s_i, \tag{A.2}$$

$$\hat{\sigma}_\eta = h^{-1} \sum_{i \in H} \tilde{s}_i. \tag{A.3}$$

In most of the computations reported in this paper, we estimated α at each step rather than assuming that $\alpha = 0$.