

Graphical technique for identifying a monotonic variance stabilizing transformation for absolute gene intensity signals

Kellie J Archer*^{1,2}, Catherine I Dumur³ and Viswanathan Ramakrishnan¹

Address: ¹Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298, USA, ²Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23298, USA and ³Department of Pathology, Virginia Commonwealth University, Richmond, VA 23298, USA

Email: Kellie J Archer* - kjarcher@vcu.edu; Catherine I Dumur - cidumur@hsc.vcu.edu; Viswanathan Ramakrishnan - vramesh@hsc.vcu.edu

* Corresponding author

Published: 17 May 2004

Received: 19 February 2004

BMC Bioinformatics 2004, **5**:60

Accepted: 17 May 2004

This article is available from: <http://www.biomedcentral.com/1471-2105/5/60>

© 2004 Archer et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: The usefulness of \log_2 transformation for cDNA microarray data has led to its widespread application to Affymetrix data. For Affymetrix data, where absolute intensities are indicative of number of transcripts, there is a systematic relationship between variance and magnitude of measurements. Application of the \log_2 transformation expands the scale of genes with low intensities while compressing the scale of genes with higher intensities thus reversing the mean by variance relationship. The usefulness of these transformations needs to be examined.

Results: Using an Affymetrix GeneChip[®] dataset, problems associated with applying the \log_2 transformation to absolute intensity data are demonstrated. Use of the spread-versus-level plot to identify an appropriate variance stabilizing transformation is presented. For the data presented, the spread-versus-level plot identified a power transformation that successfully stabilized the variance of probe set summaries.

Conclusion: The spread-versus-level plot is helpful to identify transformations for variance stabilization. This is robust against outliers and avoids assumption of models and maximizations.

Background

Microarrays measure the abundance of thousands of mRNA transcripts in one experiment. Currently, two different microarray technologies dominate gene expression research efforts, namely, custom spotted arrays and Affymetrix GeneChips[®]. Custom spotted arrays are characterized by long single strands of complimentary DNA (cDNA) affixed to a solid substrate in spots to which two different fluorescently labelled samples are hybridized. Affymetrix GeneChips[®] are characterized by the use of several (11–20) short oligonucleotide (25-mers) probes to interrogate for a single gene and to which one fluorescently labelled sample is hybridized. For both technologies, the fluorescence intensity for each probe/spot is

assumed to be indicative of the amount of mRNA transcript in the sample. Since two samples are hybridized to custom spotted arrays, the ratios of the experimental signal relative to a control signal of the resulting intensities are analyzed. On the other hand, for the one sample hybridized to an Affymetrix GeneChip[®], the absolute intensities indicative of the number of transcripts are the resulting gene expression measures.

To satisfy assumptions required for statistical analyses, data from both technologies are often transformed by a suitable function [1]. Specifically, the reasons for applying a transformation to a dataset include to achieve stability in variance, or to achieve linearity, additivity and/or

normality. Sometimes a transformation is applied because it facilitates interpretation and induces symmetry. Such is the case when applying the \log_2 transformation to custom spotted microarray (cDNA) data. This transformation defines the relative abundance of a transcript in an experimental sample in comparison to the control sample as the unit of analysis. For genes where the experimental intensity is greater than the control intensity, the ratio could take values in the range $(1, \infty)$; for genes where the experimental intensity is less than the control intensity, the ratio is compressed to the range $(0, 1)$. The \log_2 transformation of ratio data promotes symmetry by treating under and over expressed genes similarly. A typical example used to elucidate this concept is the one that considers two genes, one with a ratio of 2.0 representing a doubling of intensity and another with a ratio of 0.5 represents a halving of intensity; on a \log_2 scale these values are symmetric about zero (no change) with values 1 and -1 respectively [1]. The application of the \log_2 transformation to cDNA data thus has proven useful for achieving symmetry as well as ease of interpretation.

The usefulness of the \log_2 transformation for cDNA microarray data has led to its widespread application to Affymetrix GeneChip® data as well [2-4]. While this transformation is appealing to the biologists due to the reasons stated above for cDNA arrays, it neither facilitates interpretation nor does it necessarily render true the assumptions required for statistical analysis, such as equal variance, normality, etc., in Affymetrix GeneChip® data. When data are amounts or counts, as with Affymetrix GeneChip® data, where the intensities are assumed to represent amounts of mRNA transcripts, there is often a systematic relationship between the variance of the measurements and magnitude of the measurements. The \log_2 transformation expands the scale of genes with low absolute intensities while compressing the scale of genes with higher intensities; it essentially reverses the direction of the relationship between the variance and the mean expression level. That is, after the transformation lowly expressed genes have a higher variance and highly expressed genes have a lower variance [5].

Recently, other variance stabilizing transformation methods for microarray gene expression data have been introduced [6-8]. The gene expression data have been modelled as

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

where y represents the measured intensity, α represents the average background, μ represents the true gene expression level, with normally distributed error terms η and ε with zero mean and differing non-zero variances [6]. For this model, when μ is large, y is distributed approximately

as a lognormal random variable. Therefore, a log transformation of y results in observations with constant variance when μ is large. In this case the generalized logarithm transformation is

$$g(y) = \ln (y - \alpha + \text{sqrt} [(y - \alpha)^2 + c]), \quad (2)$$

where $c = \sigma_{\varepsilon}^2 / S_{\eta}$, stabilizes the asymptotic variance [6] for large samples. Rocke and Durbin [7] compared three logarithmic-based transformations (the generalized logarithm, the started logarithm, and the log-linear hybrid) using a simulation study and in application to an existing dataset. They found the generalized logarithm resulted in better overall performance in achieving variance stabilization.

In this paper, an Affymetrix GeneChip® HG-U133A dataset consisting of 16 technical replicates (QAQC Dataset), where the Microarray Suite Software (version 5.0) was used to derive the expression summaries for all probe sets, is used to demonstrate some of the problems associated with applying the \log_2 transformation to absolute intensity data. Another approach to identify an appropriate variance stabilizing transformation using the spread-versus-level plot [9] is proposed. The spread-versus-level plot plots the log of the median on the x-axis ($\log(M_X)$) against the log of the fourth-spread on the y-axis. The slope of the spread-versus-level plot (b) can be used to suggest a power transformation that will most appropriately stabilize the variance. When using a spread-vs-level plot, the suggested power of the transformation is $p = 1 - b$.

The effectiveness of using the spread-versus-level plot to identify a suitable monotone transformation that appropriately stabilizes the variance for probe set expression summaries is demonstrated using the QAQC dataset.

Results

To demonstrate the dependence of the variance on level of gene expression, a plot of the mean for each probe set across the 16 technical replicates by the associated variance was constructed (Figure 1). It is apparent from the plot that the variance increases with increasing gene expression. Figure 2 represents the plot of the mean versus the corresponding variance for the \log_2 transformation. This plot demonstrates, as previous authors have found [4,5], that the log transformation increases the variability on the lower end of the range while compressing the variability among the highly expressed genes. In other words, clearly, the log transformation has not stabilized the variance in this case.

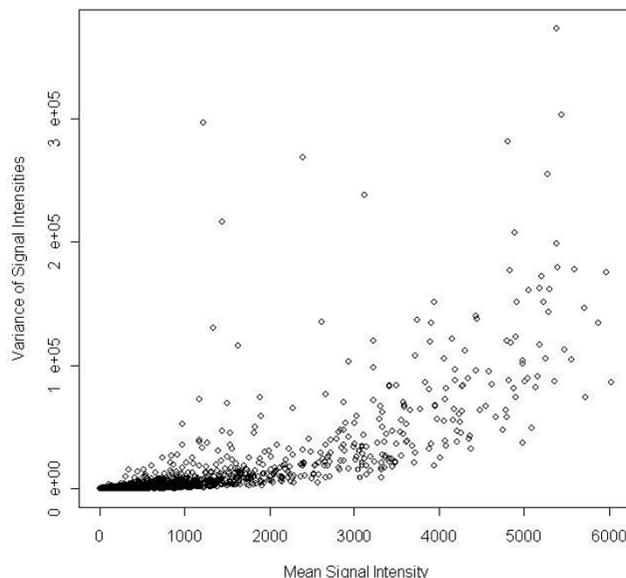


Figure 1
Mean versus variance plot. Plot of the mean probe set MAS 5.0 signal intensities by the corresponding variance across the 16 HG-UI33A GeneChips®.

In Figure 3 the spread-versus-level plot for the same data is presented. The estimated slope of the linear regression model fit to the spread-vs-level plot was $\hat{\beta}_1 = 0.569$. Therefore, the estimate of the power transformation turned out to be, $\hat{p} = 1 - 0.569 = 0.431$. Typically, for simplicity the parameter estimate \hat{p} is rounded to the nearest half-integer, which in this case would lead to $\hat{p} = 0.5$. The family of power transformations is defined as

$$\Phi_p(x) = \begin{cases} \frac{x^p - 1}{p} & \text{for } p \neq 0, \\ \ln(x) & \text{for } p = 0, \end{cases} \quad (3)$$

where, x is the observed data, p is a number to be determined and Φ is the monotonic transformation. Substituting the p obtained from the plot yields,

$$\Phi_{0.5}(x) = (x^{0.5} - 1)/0.5 = 2\sqrt{x} - 2.$$

Applying this transformation to the signal intensities in the QAQC dataset and plotting mean versus the variance as before (Figure 4) shows that stabilization of the variance is achieved. There are a few outlying probe sets identifiable in Figure 4. One of the advantages of using the estimated slope from the spread-versus-level plot to identify a power transformation is that this plot uses robust measures of location (median) and spread (fourth-spread). For this dataset, the "outlying" probe sets were defined as those with a variance greater than 50. The estimated slope after removing the outliers was 0.568; as expected, the slope of the linear regression fit to the spread-versus-level plot is fairly robust against outliers. Thus, the transformation identified by the spread-versus-level plot is not affected by these few outlying probe sets.

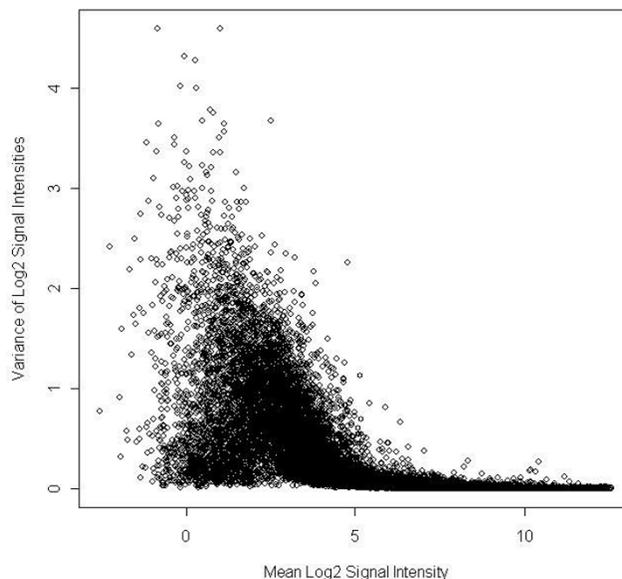


Figure 2

Mean versus variance plot of \log_2 transformed data. Plot of the mean \log_2 transformed MAS 5.0 probe set signal intensities by their associated variance for the 16 HG-U133A GeneChips®.

Figure 4, however, is not on the same scale in comparison to Figures 1 and 2. Therefore, a plot of the rank of the median probe set signal intensities versus the associated fourth-spread across the 16 HG-U133A GeneChips® (Figure 5) is considered [10]. This plot more clearly demonstrates the improvement in variance stabilization achieved by this transformation. The fitted lowess regression curve is overlaid, where the span is $1/3$, is roughly horizontal, indicating constant variance.

As noted in this figure, there are still a few probe sets at the highest intensity level that suffer from increased variability. These probe sets are among those expression summary values that are greater than the expression summary values for the most concentrated hybridization spike-in control (that is, the AFFX-CreX and AFFX-r2-P1-cre probe sets, with a mean of 2796 on MAS 5.0 scale, 103.7 on the transformed scale). From our quality assessment of the hybridizations, there is linearity over the range of concentrations covered by all the hybridization spike-in controls.

However, for signal intensities outside of this range, linearity cannot be determined. In a statistical analysis, these may be deemed outliers and may need to be examined separately from the remaining probe sets.

After applying the generalized logarithm transformation [6,8] to the QAQC Dataset using the Bioconductor `vsn()` library [8] and plotting the rank of the median probe set signal intensities by the associated fourth-spread (Figure 6) reveals the retention of increased variability among genes with smaller signal intensities, though vastly improved in comparison to the \log_2 transformation. The fitted lowess regression curve is overlaid where the span is $1/3$. The coefficient of variation for the fitted values from the lowess regression for the $2\sqrt{x} - 2$ transformed data was 0.15 while the coefficient of variation of the fitted values from the lowess regression for the generalized logarithm transformation was 0.22. Bootstrap percentile-based confidence intervals for the coefficient of variation were estimated by resampling the paired median and

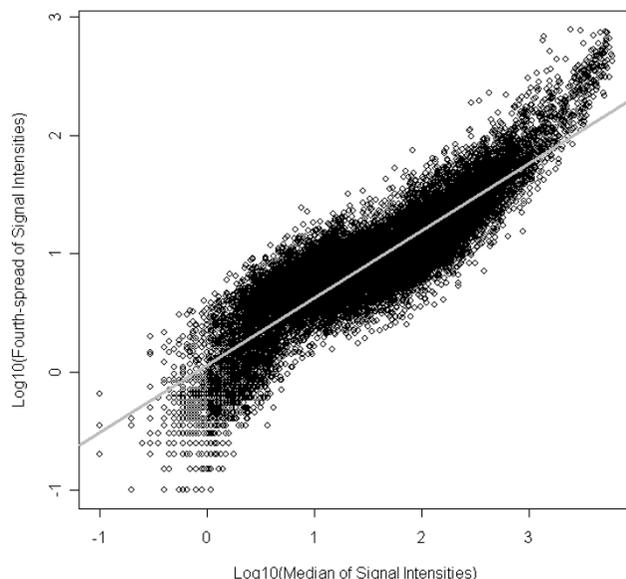


Figure 3
Spread versus level plot. Spread-versus-level plot for 16 HG-U133A GeneChips® using MAS 5.0 probe set expression summaries; parameter estimates from least squares regression: $\hat{\beta}_0 = 0.052$, $\hat{\beta}_1 = 0.57$.

spread observations, fitting a lowess regression to the bootstrapped data, then estimating the coefficient of variation using the lowess fitted values. The 99% confidence interval for the $vsn()$ transformed data was (0.2109, 0.2265); the 99% confidence interval for the $2\sqrt{x} - 2$ transformed data was (0.1431, 0.1644). Thus, it may be concluded that the variance was better stabilized by the $2\sqrt{x} - 2$ transformation.

Discussion

Reasons for transforming data often include the desire to find a scaling of the data that would simplify the analysis, for example, by making the data meet assumptions such as symmetry, equal variance, straight line relationship, and/or additivity of effect. One common assumption of many traditional statistical methods is that the variance is constant over the levels of gene expression. Since the absolute intensities are positive, this assumption is generally not valid because the variance increases with increasing intensities. The \log_2 transformation, which is routinely

applied to absolute intensities that result from Affymetrix GeneChip® experiments [2,3], unfortunately does not achieve the desired results. As shown (Figure 2) the transformation only reverses the mean-variance relationship. To overcome this drawback of the \log_2 transformation, methods such as the Local Pooled Error test have been suggested [4] to more accurately estimate the variance for lowly expressed \log_2 signal intensities. In this article, as one alternative, the spread-versus-level plot to identify more appropriate monotonic transformations that stabilize the variance was proposed. The spread-vs-level plot proved useful for identifying a transformation useful for simplifying an analysis by meeting assumptions such as symmetry and equal variance.

For the QAQC Dataset, the spread-versus-level plot led to a square root transformation, rather than the \log_2 transformation. This dataset consisted of 16 technical replicates; since the spread-vs-level plot is used to empirically identify a power transformation, the most appropriate power transformation to stabilize the variance is completely data

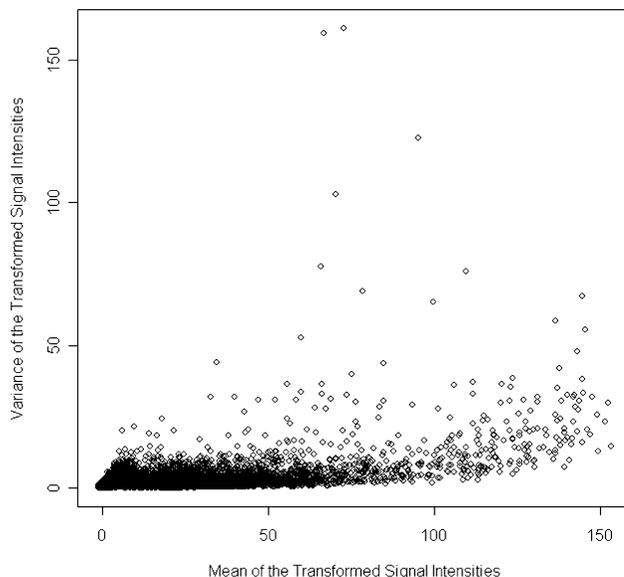


Figure 4
Mean versus variance plot of power transformed data. Plot of the mean of the probe set signal intensities after applying the $2\sqrt{x} - 2$ transformation by the associated variance for the I6 HG-U133A GeneChips®.

specific. Therefore, the $2\sqrt{x} - 2$ transformation suggested should be treated as an illustration and is not necessarily the most appropriate one for all data sets. In fact, there may be situations in which our method may lead to the more commonly used log transformation itself, while for other datasets it may not lead to any power transformation that could stabilize the variance. We note, however, that in a serial dilution experiment conducted to assess linearity of signal [11], the authors found the square root transformation by trial and error to be an appropriate transformation. Thus this empirically derived transformation obtained through the use of the spread-versus-level plot confirmed a similar finding.

Conclusions

The estimated slope from the spread-versus-level plot could be used to identify a power transformation that will appropriately transform gene expression data to stabilize the variance. It is computationally easy to implement as it avoids the requirement of a statistical model and maximization of a likelihood. Moreover, since this plot uses

robust measures of location and spread, the slope of the linear regression fit to the spread-versus-level plot is fairly robust against outliers.

Methods

Transformation

One of the most commonly used families of variance stabilizing transformations is the Box-Cox transformation [12]. This family of transformations is defined by equation (3). Box and Cox selected p as the value that maximized the likelihood under any given model. A graphical technique based on the spread-versus-level plot may also be used to empirically estimate p [9]. This method is robust against outliers and avoids the need for maximization of the likelihood. The spread-versus-level plot plots the log of the median (x -axis) by the log of the fourth-spread (y -axis). The slope b of the spread-versus-level plot identifies the approximate value of the exponent for the power transformation, where $p = 1 - b$. When the slope b is close to 1, p is close to 0 and therefore the transformation would be $\ln(x)$.

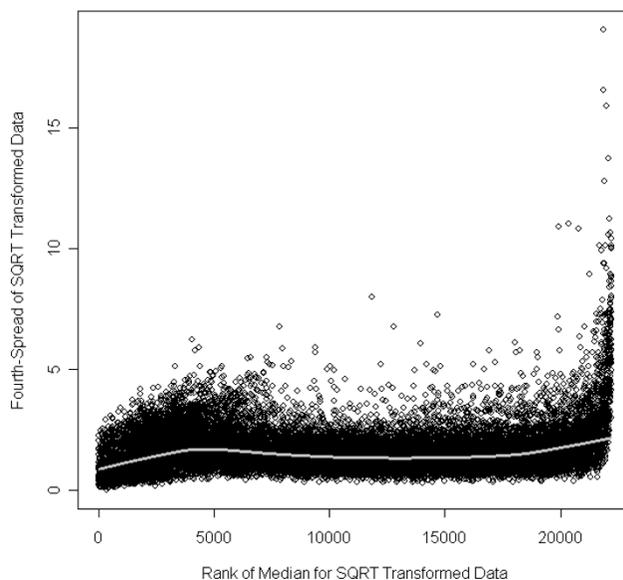


Figure 5
Rank versus spread plot of power transformed data. Plot of the rank of the median probe set signal intensities after applying the $2\sqrt{x} - 2$ transformation by the associated fourth-spread across the 16 HG-U133A GeneChips®. The fitted lowess regression curve is overlaid where the span is 1/3.

Spread-versus-level plot

To construct a spread-versus-level plot, three order statistics, namely the median (M) the upper fourth (F_U) and the lower fourth (F_L) must be identified. These order statistics may be found using the following algorithm. First, sort the data in ascending order. On the basis of ordering, the rank of an observation can be defined as an upward rank (determined by counting up from smallest to largest) or a downward rank (determined by counting down from largest to smallest). Considering both the upward and the downward ranks together, the depth of a data value in a sample is defined as the smaller of its upward rank and its downward rank. For example, the depth of the median is $(n + 1)/2$. (When the number of observations n is even, the median is interpolated to lie between the two middle values.) The fourths are those values with a depth equal to $(\lfloor \text{depth of median} \rfloor + 1)/2$. Therefore, each fourth is halfway between the median and the corresponding extreme [13]. The fourth-spread (d_F) is a robust measure of spread and is estimated by (upper fourth) - (lower fourth),

denoted by $d_F = F_U - F_L$. The interquartile range is nearly the same as the fourth-spread but differs slightly when the number of observations is even.

To understand why a spread-versus-level plot is useful in identifying an appropriate variance stabilizing transformation [9], let X be a real random variable from a population with median M_X , lower fourth F_L , upper fourth F_U , and fourth spread d_F . The spread-versus-level plot identifies a transformation $y = \Phi(x)$ for which the fourth-spread of Y is constant. Assuming the second derivative $\Phi''(x)$ and higher order derivatives exist for all x , a Taylor series expansion of $\Phi(x)$ around M_X is

$$\Phi(x) = \Phi(M_X) + \Phi'(M_X)(x - M_X) + \Phi''(M_X) \frac{(x - M_X)^2}{2!} + \dots \tag{4}$$

Let $\lambda(M_X)$ represent the distance from M_X to F_U as a proportion of the distance d_F , or $\lambda(M_X) = (F_U - M_X)/d_F$; then $[1 - \lambda(M_X)]$ would represent the proportionate distance from

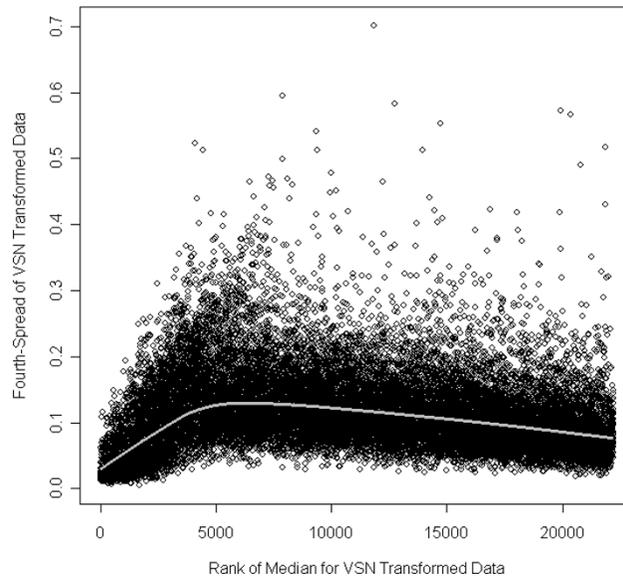


Figure 6

Rank versus spread plot of generalized logarithm transformed data. Plot of the rank of the median probe set signal intensities after applying the generalized logarithm transformation by the associated fourth-spread across the 16 HG-U133A GeneChips®. The fitted lowess regression curve is overlaid where the span is 1/3.

M_X to F_L . Therefore, the median and upper and lower fourths for the data transformed by $\gamma = \Phi(x)$ are

$$M_Y = \Phi(M_X), \quad (5)$$

$$(F_U)_Y = \Phi(F_U) = \Phi [M_X + \lambda(M_X)d_F], \quad (6)$$

$$(F_L)_Y = \Phi(F_L) = \Phi \{M_X - [1 - \lambda(M_X)]d_F\}. \quad (7)$$

Substituting these expressions into equation (4) to get $(F_U)_Y$ and $(F_L)_Y$, then taking the difference to get the fourth-spread for the transformed data yields,

$$(d_F)_Y = d_F \cdot \Phi'(M_X) + \frac{[2 \cdot \lambda(M_X) - 1]d_F^2}{2!} \Phi''(M_X) + \dots \quad (8)$$

Recall that $\lambda(M_X)$ represent the distance from M_X to F_U as a proportion of the distance d_F ; thus the term $[2 \cdot \lambda(M_X) - 1]$ will range from $[0,1]$. Moreover, if the upper and lower fourths are roughly equal distant from the median (i.e., x is roughly symmetric), this quantity will be zero. Addi-

tionally, since $\Phi''(M_X)$ is a measure of concavity of the tangent line at the median, it will generally be small. Therefore, the quadratic term of $(d_F)_Y$ is generally small enough to be negligible compared to the leading term, thus the approximation

$$(d_F)_Y \approx d_F \cdot \Phi'(M_X) \quad (9)$$

is reasonable. Now, if the right-hand side of this equation is held constant, it leads to a transformed variable Y with constant variance. Therefore, supposing that the relationship between spread d_F and level X is a power transformation of the form $d_F(x) = k \cdot x^b$ for some constants k and b ,

$$\Phi(x) = \frac{1}{k} \int \frac{1}{x^b} dx = \begin{cases} c_1 x^{1-b} + c_2 & \text{for } b \neq 1 \\ c_3 \ln(x) + c_4 & \text{for } b = 1. \end{cases} \quad (10)$$

The spread-versus-level plot plots the log of the median on the x-axis ($\log(M_X)$) against the log of the fourth-spread on the y-axis. (Recall, $d_F(x) = k \cdot M_X^b$ so that $\log(d_F)$

$= \log(k) + b \cdot \log(M_X)$). Therefore, the slope of the spread-versus-level plot estimates b . Hence the suggested power transformation would be $x^p = x^{1-b}$.

Data

An experiment using 16 technical replicates in which data were obtained using the Affymetrix GeneChip® technology is considered to illustrate the transformation method suggested in the previous section. Five identical aliquots of 40 µg of the Universal Human Reference RNA (Stratagene, La Jolla, CA), isolated from 10 human cell lines, were used as template for cDNA synthesis, cRNA *in vitro* transcription, and labelling reactions. The final fragmented cRNA aliquots were pooled together. The pooled Reference RNA was hybridized to 16 HG-U133A GeneChips® (Dumur et al, personal communication) to study GeneChip® reproducibility. Hybridization was performed using the same lot number of the HG-U133A arrays, which contains 22,283 probe sets. Immediately after the scanning of every chip, a visual quality control of the image was performed in order to ensure a successful hybridization. Moreover, several parameters produced by the Affymetrix Microarray Suite Software (version 5.0) were monitored for quality control purposes, including: the scaling factor, which is used to scale all probes sets to a predetermined target value (in this study the target value was set at 100); the percentage of probe sets called 'Present' using the Affymetrix Detection Call algorithm [14]; and the 3'/5' ratios of signal intensity values for two housekeeping genes, GAPDH and β-actin. The Microarray Suite Software (version 5.0) was also used to derive the expression summaries for all probe sets. The present study was performed on the 22,215 probe sets designed to match human transcripts; we eliminated those probe sets that query for bacterial spike-in controls from our analyses.

Authors' contributions

KA conceived the transformation study, performed the bioinformatical analysis and drafted the manuscript. CD participated in the experimental design of the microarray study and performed the GeneChip hybridizations. VR participated in manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by the Commonwealth Technology Research Fund (CTRF #SE2002_02). The authors would like to thank the two anonymous reviewers for their helpful insights and recommendations.

References

1. Quackenbush J: **Microarray data normalization and transformation.** *Nature Genetics* 2002, **32(Supplement)**:496-501.
2. Hubbell E, Liu W-M, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18(12)**:1585-1592.
3. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Research* 2003, **31**:e15.
4. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, **19**:1945-1951.
5. Kohane IS, Kho AT, Butte AJ: **Microarrays for an integrative genomics.** Cambridge, MA: MIT Press; 2003.
6. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18(Suppl. 1)**:S105-S110.
7. Rocke D, Durbin B: **Approximate variance-stabilizing transformations for gene-expression microarray data.** *Bioinformatics* 2003, **19(8)**:966-972.
8. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl. 1)**:S96-S104.
9. Emerson JD: **Mathematical Aspects of Transformation.** *Understanding Robust and Exploratory Data Analysis*. Edited by Hoaglin DC, Mosteller F, Tukey JW. New York: John Wiley & Sons; 1987:247-282.
10. Geller SC, Gregg JP, Hagerman P, Rocke D: **Transformation and normalization of oligonucleotide microarray data.** *Bioinformatics* 2003, **19(14)**:1817-1823.
11. Ramdas L, Coombes KR, Baggerly K, Abruzzo L, Highsmith WE, Krogmann T, Hamilton SR, Zhang W: **Sources of nonlinearity in cDNA microarray expression measurements.** *Genome Biology* 2001, **2(11)**:RESEARCH0047.
12. Box GEP, Cox DR: **An analysis of transformations.** *Journal of the Royal Statistical Society, Series B* 1964, **26(2)**:211-252.
13. Hoaglin DC: **Letter Values: A Set of Selected Order Statistics.** *Understanding Robust and Exploratory Data Analysis* Edited by: Hoaglin DC, Mosteller F and Tukey JW. New York, John Wiley & Sons; 1987:33-57.
14. Liu W-M, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP: **Analysis of high density expression microarrays with signed-rank call algorithms.** *Bioinformatics* 2002, **18(12)**:1593-1599.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

