Statistical Applications in Genetics and Molecular Biology

Volume 2, *Issue* 1 2003 *Article* 4

Transformations for cDNA Microarray Data

Xiangqin Cui* M. Kathleen Kerr†

Gary A. Churchill‡

*The Jackson Laboratory, xcui@uab.org

†University of Washington, katiek@u.washington.edu

‡The Jackson Laboratory, garyc@jax.org

Copyright ©2003 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress, which has been given certain exclusive rights by the author. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). http://www.bepress.com/sagmb

Transformations for cDNA Microarray Data

Xiangqin Cui, M. Kathleen Kerr, and Gary A. Churchill

Abstract

Two channel microarray data often contain systematic variations that can be minimized by data transformation prior to further analysis. The most commonly observed effects are revealed by viewing scatter plots of the logarithm of the ratio by the average logarithmic intensity of the two color channels (RI plots). In this paper we present a general model for signal intensity data with multiple error sources. We demonstrate how these sources of error influence the shape of an RI plot. We then compare some currently available transformation strategies in terms of their mechanism and performance on both simulated and real microarray data. A linlog transformation is proposed to stabilize the variance of the log ratios. We also propose a regional smoothing method to remove variation in log ratios due to spatial heterogeneity on the microarray surface. The discussed transformations represent an important initial step in microarray data analysis for both ratio-based and ANOVA methods.

Introduction

In a spotted cDNA microarray experiment, a mixture of two cDNA samples (targets) that are differentially labeled with fluorescent dyes is hybridized to DNA sequences (probes) immobilized on a glass slide (Schena et al., 1995). Sequences from the two targets hybridize to the complementary probe sequences. The observed fluorescent signals at each spot are, therefore, correlated with the mRNA concentrations in the RNA samples from which the cDNA targets were reverse-transcribed. The ratio of the two fluorescent signals at each spot is commonly used to infer the ratio of the mRNA concentrations in the two RNA samples; however, the ratio of the fluorescent signals is influenced by systematic effects from non-biological sources that can introduce biases and should be removed before drawing conclusions about the relative levels of gene expression.

The process of removing systematic effects is often referred to as normalization. We consider this process in three steps: background subtraction, data transformation, and data normalization. Background subtraction is the step of subtracting background from the fluorescent signals at each spot. The spot background is usually estimated locally by gridding softwares using various algorithms. Many algorithms can overestimate the background (Yang et al., 2002a); we leave this choice to the discretion of users but suggest that it is worthwhile to critically examine the effect that background subtraction has on results. Data transformation is applied to data from one microarray at a time to remove systematic effects from log ratios. Normalization is a calibration of the signals from different microarrays to put them all on a comparable scale. A variety of data normalization approaches have been proposed. The total signal method (Quackenbush, 2001; Bilban et al., 2002) adjusts signals based on the assumption that the total hybridization strength is the same between the two channels for each array. The ratio function method (Chen et al., 1997) assumes that the coefficients of variation of the two channels are the same. The ANOVA methods (Kerr et al., 2000; Jin et al., 2001; Wolfinger et al., 2001) adjust for overall effects of array and dye across genes. Since the background subtraction and normalization steps have been addressed elsewhere (Yang et al., 2002a; Quackenbush, 2001; Quackenbush, 2002), this paper will focus on the transformation step.

A commonly observed feature of microarray data is the dependence of mean log ratio on fluorescent intensity. This dependence can be diagnosed by viewing a scatter plot of log ratio versus average of log intensity as suggested by Dudoit *et al*. (2000). This graphical representation is referred to here as the ratio by intensity (RI) plot, but it has also been referred to as the MA plot by some researchers. Under the assumption that most genes are not differentially expressed and/or that any differential expression is symmetric with respect to up- and down-regulation, most points on an RI plot should fall along a horizontal line. In practice, RI plots often show various kinds of curvature. There are several strategies to transform the data in order to remove the curvature. Shift methods adjust the signals of the two channels using an additive constant prior to taking logarithms (Kerr et al., 2002; Newton et al., 2001). Curve-fitting strategies use local (on intensity axis) regression to estimate a standard curve and then recenter the data. Both lowess (Yang et al., 2002b) and centralization (Zien et al., 2001) methods belong to this category.

It has been demonstrated that the variance of log ratios also depends on signal intensity (Rocke and Durbin, 2001). When raw data are considered, variation increases as intensity increases. When log-transformed data are considered, the variance is usually stable above a certain intensity but the low intensity spots can be highly variable. This effect is often exaggerated in background-subtracted microarray data. Arsinh transformations were proposed to stabilize the variance at the low intensity end (Huber et al., 2002; Durbin et al., 2002). Here we propose an alternative transformation called linlog, which combines linear and log transformations, to achieve the same goal (Holder et al., 2001).

Spatial heterogeneity on the microarray surface is another source of systematic variation of log ratios. Because cDNA microarrays are typically spotted with a set of probe DNAs in an arbitrary fashion (with

the possible exception of control spots) one does not expect to see an association between over- or under-expression and spatial regions of a slide. However, microarray users sometimes observe spatial patterns of log ratios that are obviously not due to effects of differential expression. Colantuoni *et al*. (2002) proposed a local mean normalization approach to remove the spatial bias of signal intensities in filter arrays. We describe a regional smoothing method in the same spirit to reduce the spatial bias of log ratios for cDNA microarrays and a joint smoothing method to remove both intensity and spatial biases in one step.

In this paper, we attempt to understand microarray data using a linear model with multiple sources of measurement error. We review and compare some of the available data transformation methods for cDNA microarray data. Due to space considerations, we will not discuss other methods such as third root (Tusher et al., 2001), conditional expectation (Bolstad et al., 2003), and non-linear regression (Kepler et al., 2002). Some alternative approaches are provided and general recommendations are given on how to choose transformation methods for various situations.

1 Modeling Signal Intensity

The basic assumption for cDNA microarray technology is that the fluorescent signal intensities measured at each spot are correlated with mRNA concentrations of the corresponding gene in the samples. However, the signals are also related to spot characteristics. Acknowledging the lack of control over spot characteristics, cDNA microarrays use a two-dye system to compare signals from the same spot. The two dyes have different properties, such as brightness and incorporation efficiency. These differences result in systematic biases that we would like to remove before drawing biological conclusions from the data. In this section, we introduce a model starting from the simplest situation to explain some of the features we have observed in microarray data.

1.1 Linear Model

The simplest and most ideal situation occurs when fluorescence intensity (Y) detected from each channel at each spot on a microarray is linear to the corresponding mRNA concentration in the target. That is

$$Y_{ik} = a_i + b_i X_{ik} \tag{1}$$

For $i = r$ or g (channels) and $k = 1, 2, ...K$ (spots), the fluorescent signal at channel i and gene k, Y_{ik}, is the sum of channel mean background signal, a_i, and the signal from the hybridization of the corresponding target, $b_i X_{ik}$. X_{ik} is the mRNA concentration of gene k in the corresponding target sample of channel i. b_i is the channel-specific slope of the linear relationship. Due to differences between the two dyes and the setting of the two scanning channels, a_i and b_i may have different values for the two channels (i), but in this ideal model they are considered constant across all of the spots. In the absence of internal standards for calibration we cannot know the functional relationship between RNA concentration and fluorescent signal intensity, but it appears that the linear relationship is approximately valid over a wide range of intensities (Dudley et al., 2002).

In practice, fluorescent intensity cannot exceed the saturation value of the scanner, such as 65535 for the GenPix 4000B (Axon Instrument, Inc., CA). Figure 1 shows the relationship between mRNA concentration and fluorescent signal intensity on the original scale and on the logarithmic scale. Saturation of the scanner results in the flattening of the curves at high RNA concentration. In real data, the transition to saturation is always more smooth (Ramdas et al., 2001). This could result because the saturation of individual pixels within a spot is not simultaneous due to irregularities in the distribution of DNA probe. In cases of extensive saturation, we recommend rescanning the slide at a lower intensity.

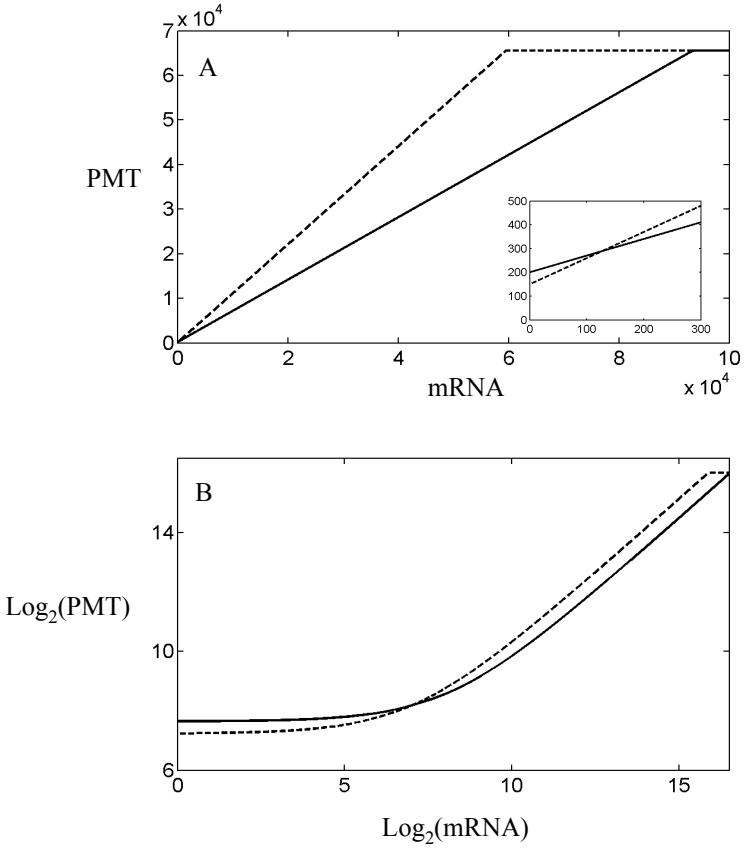

Figure 1: Idealized relationship between mRNA concentration and fluorescent signal intensity for raw (A) and log-transformed (B) data. The solid and dash lines represent Cy3 and Cy5 channels, respectively. Each channel has its own characteristic slope and non-negative intercept (background). Truncation of high signals represents scanner saturation. The mean background difference between the two channels in panel A is shown in the inserted panel with lower scales. PMT, raw fluorescent signal intensity.

Analysis methods that make adjustments for saturation could be developed (Dudley et al., 2002), but are likely to depend strongly on modeling assumptions.

Microarray data are complicated by stochastic errors from a variety of sources. Here we decompose the measurement error into multiple components. Our model is similar to the one proposed by Rocke and Durbin (2001). Inserting error terms into model (1), we have:

$$Y_{ik} = a_i + b_i X_{ik} e^{\eta_k + \zeta_{ik}} + \varepsilon_k + \delta_{ik} \tag{2}$$

The error associated with the fluorescent signal at spot k in channel i is decomposed into multiplicative and additive components. The multiplicative component ($\eta_k + \zeta_{ik}$) is related to RNA labeling, scanning, and spot features. The additive component ($\varepsilon_k + \delta_{ik}$) is related to local background. Each of these two components is in turn decomposed into a common component (η_k or ε_k) that is shared by both channels and a channel-specific component (ζ_{ik} or δ_{ik}). We assume that the distributions of all error components are symmetric with mean zero.

Model (2) can be generalized to make any of the parameters dependent on the spatial location of the spot on the microarray surface. In particular, the parameters a_i and b_i could be spatially indexed by s to represent local mean ($a_i(s)$)and local slope ($b_i(s)$).

1.2 Ratio by Intensity Plots

The scatter plot of log ratios versus the average of log intensities (RI plot) has become an important diagnostic plot for detecting intensity dependent biases in two channel cDNA microarrays (Dudoit et al., 2000). It is essentially the same as the Tukey Mean-difference plot (Bland and Altman, 1986) and is equivalent to the scatter plot of log intensities in the two channels (Newton et al., 2001; Sapir and Churchill, 2000) with a 45 degree clockwise rotation, but has the advantage that the log ratio becomes one of the axes. Under the assumption that most genes are not differentially expressed, most points in an RI plot should fall along a horizontal line. In practice, RI plots reveal systematic dependence of the ratios on fluorescent intensity. Figure 2 shows some of the commonly-observed features.

In order to better understand the features observed in RI plots, we attempted to simulate these features by varying the parameters of model (2). To simulate the raw signal intensity for each channel at each spot on one array, X_{ik} was randomly drawn from a lognormal distribution (Hoyle et al., 2002) with mean 7 and standard deviation 1.1. We used the same X_{ik} for the two channels at each spot throughout the paper, therefore, no gene is differentially expressed in our simulation unless specially denoted and all features revealed by RI plots come from non-biological sources. Multiplicative errors, η_k and ζ_{ik}, were drawn from normal distributions N(0, σ_η^2) and N(0, $\sigma_{\zeta_i}^2$), respectively, and the additive errors, ε_k and δ_{ik}, were drawn from normal distributions, N(0, σ_ε^2) and N(0, $\sigma_{\delta_i}^2$), respectively. Finally, mean background a_i was added to complete the simulation. Values of b_i and a_i were varied in different simulations.

The RI plots from simulated data are shown beside some real RI plots in Figure 2 for comparison. Truncation occurs when a moderate number of spots reach the saturation of the scanner (Figure 2A, 2B). Excess variation at the low intensity end can be simulated using large channel-specific additive errors (δ_{ik}) (Figures 2C, 2D). If the errors for the two channels are large and roughly equal, the low intensity end of the RI plot will show large variance and will distribute symmetrically around a horizontal line. If additive errors in the two channels have different variances, the low intensity end of the RI plot will tilt toward the channel with smaller variance to form a hockey stick shape. The common component of the additive error only affects the distribution along the intensity axis and does not affect the appearance of the RI plot (not shown). Variation of the log ratios at the high intensity end is controlled by the channel-specific multiplicative error (ζ_{ik}) (Figure 2E, 2F). Variation of the common multiplicative error (η_i) only affects variation along the intensity axis at the high end (not shown).

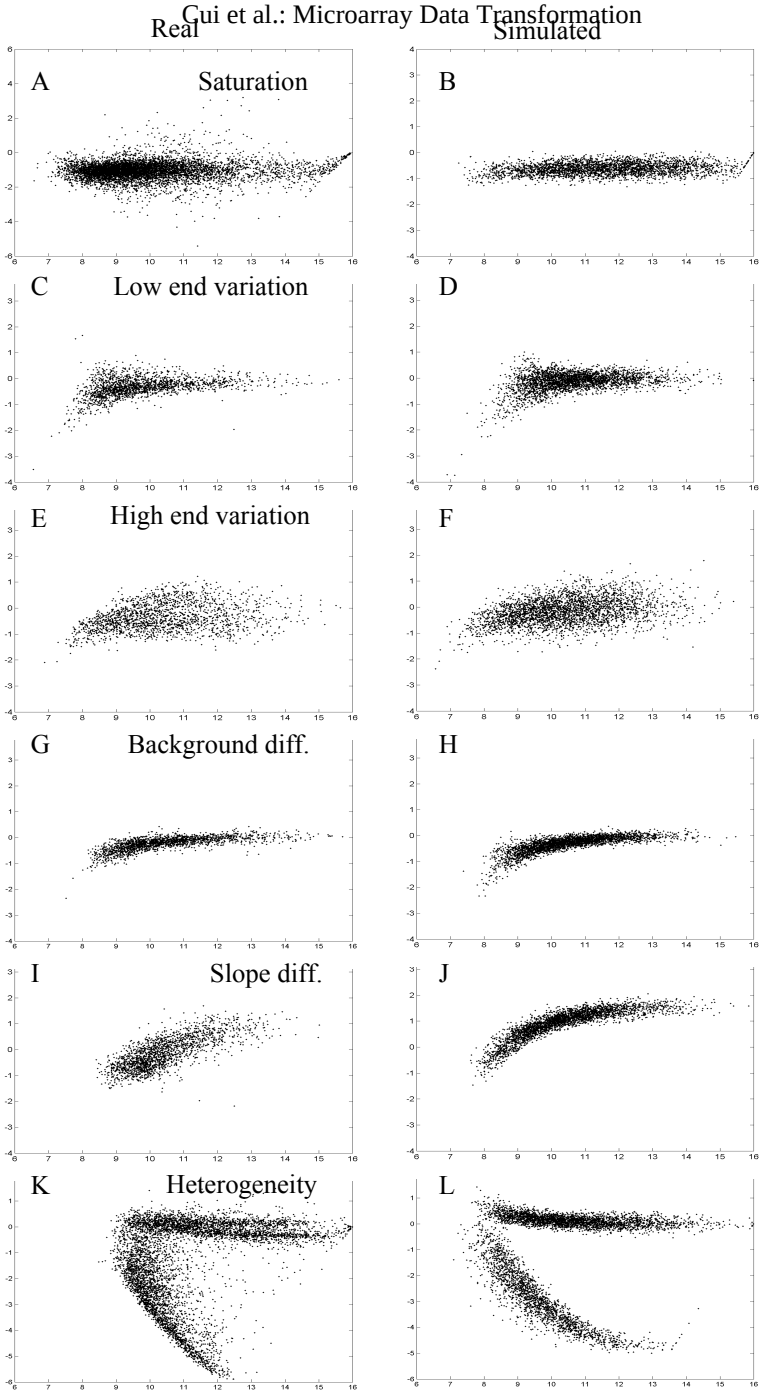

Figure 2: Comparison between RI plots from real and simulated data. Panels A, C, E, G, I, and K are from real data and panels B, D, F, H, J, and L are from simulated data. Truncation at the high intensity end (A, B) is caused by scanner saturation. High variation at the low intensity end (C, D) is from larger channel specific additive error. High variation at the high intensity end (E, F) is from larger channel specific multiplicative error. The curvatures in G and H are from channel mean background differences. The curvatures in I and J are from slope differences. The split RI plots in K and L come from spatial heterogeneity.

Curvature of RI plots can be generated by unequal mean backgrounds (a_i) and/or unequal slopes (b_i) between the two channels. The curvature generated by the unequal mean backgrounds is usually obvious only at the low intensity end and it straightens as intensity increases (Figure 2G, 2H). The curvature generated by unequal slopes usually spans across the whole intensity range (Figure 2I, 2J). A combination of these two types of curvatures will result when we have both unequal mean backgrounds and unequal slopes.

Spatial heterogeneity within an array can generate complicated features in an RI plot. Extreme examples are shown in Figures 2K and 2L. These features can be simulated by using spatial dependent slope difference.

2 Data Transformations

Raw microarray data are usually transformed, e.g., by taking the logarithm, before further statistical analysis. Data transformation can be used to achieve favorable statistical properties, for example stabilizing the variance and obtaining additivity (Snedecor and Cochran, 1989). In this section, some of the commonly used transformations for microarray data are discussed and alternative transformations are proposed. We use the notation Y_{ik} to denote intensity data on the original scale and Z_{ik} to denote transformed data. Subscripts i and k indicate channels ($i = g$ or r) and spots ($k = 1, 2, ...K$).

2.1 Logarithm transformation

A logarithmic transformation (usually base 2)

$$Z_{ik} = \log_2(Y_{ik}) \tag{3}$$

is often applied to microarray data to facilitate the calculation of fold changes from the original fluorescent signals. The logarithmic transformation not only converts ratios into differences between the two channels at each spot but also stabilizes the variance of high intensity spots. For purposes of statistical analysis, the logarithmic transformation converts multiplicative errors into additive errors (Snedecor and Cochran, 1989). If errors are proportional to the signal intensity on the original scale, they will be constant across the range of signal intensity on the logarithmic scale. On the other hand, the presence of substantial additive error on the original scale is problematic when a logarithmic transformation is applied.

Ratios It is convenient and natural to use ratios to describe the relationship between the two samples hybridized on one array; however, there are some assumptions implicit in the use of ratios which should be considered. Considering the error free linear model (1) in a context with no differential expression ($X_{rk} = X_{gk}$), we can express Y_r as a function of Y_g,

$$Y_r = a_r - (b_r/b_g)a_g + (b_r/b_g)Y_g. \tag{4}$$

The relationship is a straight line with slope b_r/b_g and intercept $a_r - (b_r/b_g)a_g$. As elaborated by Tanner (1949), This line has to pass the origin in order for a ratio-based estimation to make sense. Otherwise, the ratio of Y_r/Y_g is a biased estimator of X_r/X_g except at the mean of each variable. The farther the values are from the mean, the more biased the ratio estimator is. Thus, if we use the ratio of signal intensities to estimate the ratio of mRNA concentrations, the regression line of Y_r versus Y_g has to pass through the origin (i.e. $a_r - (b_r/b_g)a_g = 0$) in order to have unbiased estimation. This condition is unlikely to hold without some manipulation of the data. Next, we discuss transformation methods that address this problem.

2.2 Shift Transformations

The shift-log method proposed by Kerr *et al.* (2002) adjusts log ratios by adding a constant to the signal values of one channel and subtracting the same constant from signals in the other channel prior to the logarithmic transformation:

$$\begin{cases} Z_{rk} = \log_2(Y_{rk} + C) \\ Z_{gk} = \log_2(Y_{gk} - C). \end{cases} \tag{5}$$

The constant C is estimated by minimizing the absolute deviation of each log ratio ($Z_{rk} - Z_{gk}$) from the median log ratio of the array. In terms of meeting the unbiased ratio condition, the shift-log transformation moves the origin on a scatter plot of Y_r versus Y_g along the line $Y_r = -Y_g$ line to approach the regression line of Y_r versus Y_g. The curvature-causing background difference is, therefore, minimized. The original attempt to shift microarray data with two unrelated constants for the two channels turned out to be an ill-defined problem because there are an infinite number of solutions (Sapir and Churchill, 2000). Shift-log does not specifically adjust the slope of the regression line of Y_r versus Y_g; therefore, it should be less effective on curvatures resulting from slope differences. In addition, the variance of the log ratios is not dramatically affected because the signals from the two channels are shifted in opposite directions.

Newton *et al.* (2001) proposed a similar shift transformation in the context of shrinkage estimation. Their method moves the origin along the line $Y_r = Y_g$ by adding the same positive constant to both channels.

$$\begin{cases} Z_{rk} = \log_2(Y_{rk} + C) \\ Z_{gk} = \log_2(Y_{gk} + C) \end{cases} \tag{6}$$

Although this was not the intended purpose of this transformation, it can decrease the curvature in an RI plot. However, when the slopes of the two channels are the same ($b_r = b_g$), moving the origin along the $Y_r = Y_g$ line cannot bring the origin closer to the regression line. The major effect of this method is at shrinking the variance of log ratios at the low intensity end, which is appropriate when the variance at the low intensity end is high. It is not difficult to generalize this method by expanding the range of C to include negative numbers. This allows one to increase the variance at the low intensity end, which could be useful for stabilizing the variance in cases where the variance is too low at the low intensity end.

2.3 Curve Fitting Transformations

Another curvature-correcting transformation method that is commonly used in microarray analysis works by fitting a local regression line to the RI plot via locally weighted least square methods and then recentering the data along this line representing genes not differentially expressed (Yang et al., 2002b). The transformed data are

$$\begin{cases} Z_{rk} = \log_2(Y_{rk}) + C_k/2 \\ Z_{gk} = \log_2(Y_{gk}) - C_k/2, \end{cases} \tag{7}$$

where C_k is a spot-specific constant determined by the local regression line. The most common curve fitting procedure is the lowess method (Note: If a linear function is used for the local regression, it is called lowess with a "w". If a quadratic function is used for the local regression, it is called loess without the "w"). Since differentially expressed genes may appear as outliers and may have a large influence on the local fit, robust fitting procedures are preferred. Lowess fitting requires a choice of the "span" that determines which data are local. If the span is too big, the curvature cannot be removed effectively. If the span is too small, the data will be over-fit. Choice of span is subjective (usually 20% is chosen; Yang et al., 2002b). In theory, the largest span that removes the obvious intensity-dependence of the log

ratios is ideal, but this may be difficult to assess. The lowess data fitting procedures are straightforward yet a bit perilous. First, we run the risk of overfitting our data and introducing errors larger than those we remove. Second, the loess curve is flexible enough to capture many kinds of shapes in the RI plots. We must acknowledge that we may be forcing the data to meet our expectations.

2.4 Variance Stabilizing Transformations

The variance of log ratios is often seen to increase as intensity decreases (Rocke and Durbin, 2001). This can be seen in RI plots as a wider spread at the low intensity end. It can also be detected using a plot of the interquartile range versus the median intensity of each bin containing a fixed number of spots with similar intensities. This plot is referred to as an IQR (interquartile range) plot. Similar to the RI plot, the points in the IQR plot should fall along a horizontal line if there is no dependency between the variance of log ratios and spot intensities. However, larger interquartile ranges at low intensities are often observed with background-subtracted data. Some researchers have suggested removing data that fall below a threshold for intensity (Yang et al., 2001). However, this approach could sacrifice important information. In classical statistical applications where the assumption of uniform variance is needed it is a common practice to seek a transformation of the data that stabilizes the variance (Snedecor and Cochran, 1989).

Arsinh Transformation Huber *et al.* (2002) proposed an arsinh transformation to stabilize the variance of microarray data based on the assumption of a quadratic relationship between variance and intensity of microarray signals at the original scale. The transformation is

$$Z_{ik} = \log(b_i Y_{ik} + C_i + \sqrt{(b_i Y_{ik} + C_i)^2 + 1}). \tag{8}$$

The parameters, b_i, and C_i, are estimated through a robust variant of maximum likelihood estimation (Huber et al., 2003). The function adjusts the slope and mean background of each channel in equation (1) using these parameters before taking arsinh transformation. Therefore, it has the potential to correct the intensity-dependent bias as well. Because the curvature correction of this transformation only introduces four parameters for each array, it is not a strong manipulation. At the high intensity end, the arsinh transformation resembles the logarithmic transformation. At the low intensity end, it contracts to zero. The variance stabilizing property of this transformation relies on the assumption of a quadratic relationship between the variance and intensity of the original microarray signals and the effectiveness of this transformation may also depend on parameter estimation.

Linlog Transformation Here we propose another variance stabilizing transformation, linlog, which is similar but less complicated than the arsinh transformation. Model (2) suggests that additive error should be dominant at low intensities and multiplicative error should be dominant at high intensities in microarray data. Thus a linear transformation of raw data is more appropriate for low intensity spots and the logarithmic transformation is more appropriate for high intensity spots. The linlog transformation combines the linear and logarithmic transformations through a smooth transition to take advantage of both. Above a certain channel-specific signal intensity, d_i, the linlog transformation is logarithmic, while below d_i, the linlog transformation is linear. The function

$$Z_{ik} = \begin{cases} \log_2(d_i) - 1/\ln 2 + Y_{ik}/(d_i \times \ln 2) & Y_{ik} < d_i \\ \log_2(Y_{ik}) & Y_{ik} \geq d_i \end{cases} \tag{9}$$

is continuous and monotone with a continuous first derivative. The values of d_i can be estimated by minimizing the absolute deviation of the interquartile range of log ratios in each bin from the median

	truth	simulation	shift-log	lowess	arsinh
truth	1.00	0.49	1.00	0.98	0.94
simulation	0.53	1.00	0.58	0.62	0.64
shift-log	0.82	0.65	1.00	0.99	0.95
lowess	0.86	0.57	0.95	1.00	0.96
arsinh	0.83	0.53	0.88	0.93	1.00

Table 1: Pair-wise correlation coefficient among the log ratios of the truth without curvature (truth), simulated data with curvature (simulation), simulation transformed by shift-log (shift-log) and by lowess (lowess). The upper right half and the lower left half of the table are simulated using mean-background and slope differences for generating curvature, respectively.

interquartile range of the array in an IQR plot. In practice we find that values of d_i that place 25-30% of data in the linear range work quite well. A similar transformation was independently proposed by Holder et al. (2001).

Like arsinh, linlog is a transformation intended for achieving a scale at which the variance of observations is stabilized, therefore, it does not introduce any parameter that changes signal values, nor does it correct the curvatures in RI plots. However, combining linlog with either shift-log or lowess transformation will stabilize the variance and minimize the curvature. Linglogshift is the combination of linlog and shift-log as

$$\begin{cases} Z_{rk} = \text{linlog}(Y_{rk} + C) \\ Z_{gk} = \text{linlog}(Y_{gk} - C) \end{cases} \tag{10}$$

where C has the same definition here as in (5). The same algorithm for estimating C in shift-log can be implemented for linlogshift using the linlog transformation instead of the logarithmic transformation. It is also possible to apply lowess and linlog transformations in sequence to achieve the same purpose.

2.5 Performance Comparison on Simulated Data

Minimizing Curvatures Three basic transformations discussed above (shift-log, lowess, and arsinh) can correct curvatures in RI plots. In order to compare the performance of these transformations, we simulated microarray data with curvatures from mean background differences ($a_r \neq a_g$) or slope differences ($b_r \neq b_g$) as in section 1.2 and applied these transformations to the simulated data. The arsinh transformation was performed using the vsn package at *http://www.bioconductor.org/* with default input parameter settings. Results are summarized in Figure 3. In general we find that all three transformations can remove curvature caused by either background or slope differences. Minor differences arise at the low intensity end of the RI plots. Table 1 shows the correlation coefficients between the true log ratios, logarithm transformed data with curvature, shift-log transformed, lowess transformed, and arsinh transformed data. For curvature generated by mean background differences, the correlation between the shift-log transformed data and the truth is slightly higher than those for lowess or arsinh transformed data. For curvature generated by slope differences, the lowess transformed data showed slightly better correlations with the truth.

Stabilizing Variance Microarray data were simulated to give larger variances at low intensity spots in addition to curvature from both mean background and slope differences (Figure 4). The linlog and arsinh were applied to these data. For linlog transformation, d_i was set as 30% of the data. The arsinh transformation was performed as above. Both transformations can effectively stabilize the variance as shown by RI plots (Figure 4D, 4S) and IQR plots (Figure 4E, 4T). It is noticeable that the linlog

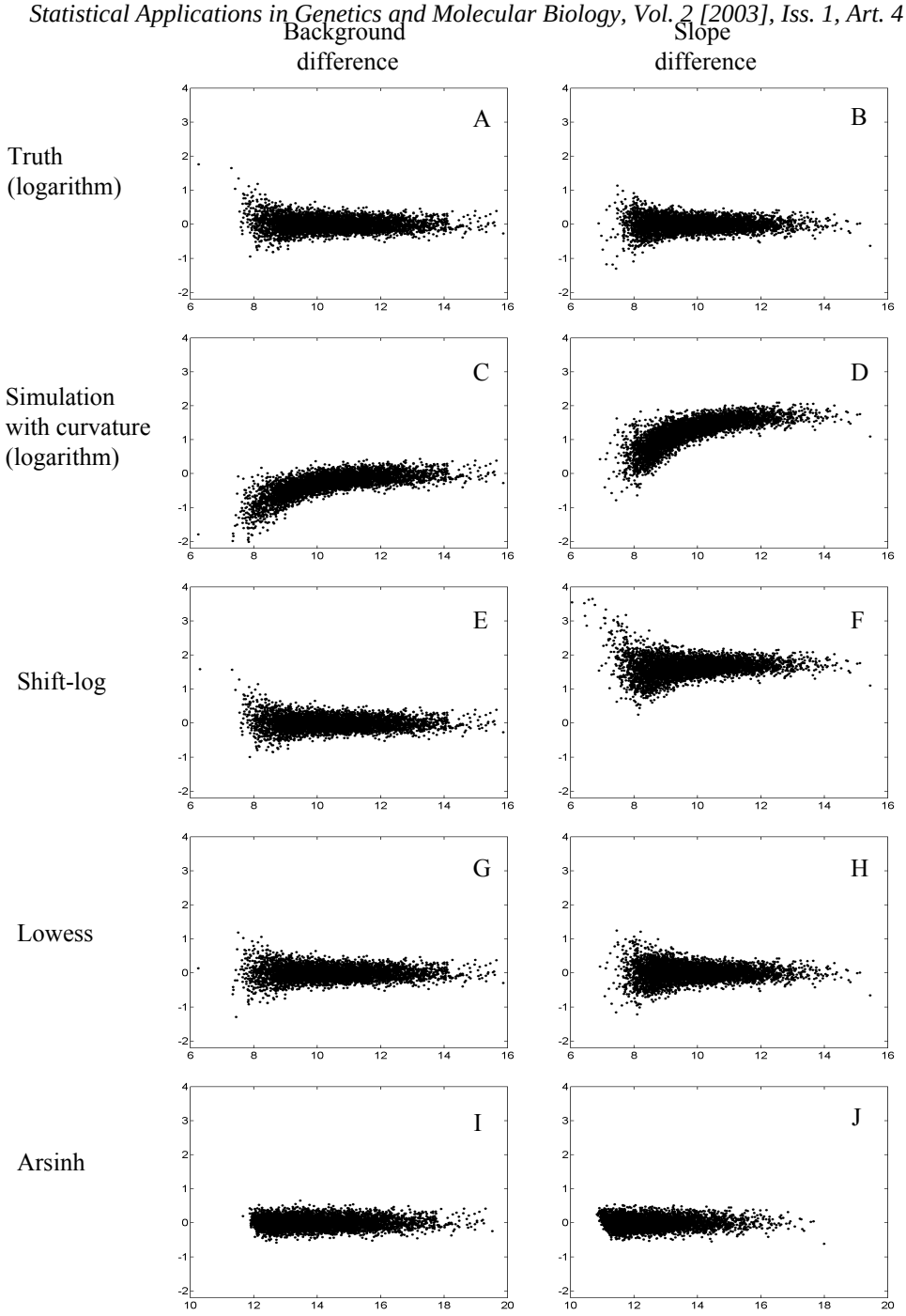

Figure 3: Comparison of shift-log, lowess, and arsinh on straightening curvatures using simulated data. Data are simulated using equation (2) with mean background difference (C, E, G, I) or slope difference (D, F, H, J). Panels A and B are truth without mean background or slope differences. The curvature in the simulated data (C, D) are effectively removed by all three transformations. The RI plot of the transformed data from shift-log, lowess and arsinh are similar to the truth (A, B). The difference is only noticeable at the low intensity end (G vs A and F vs B).

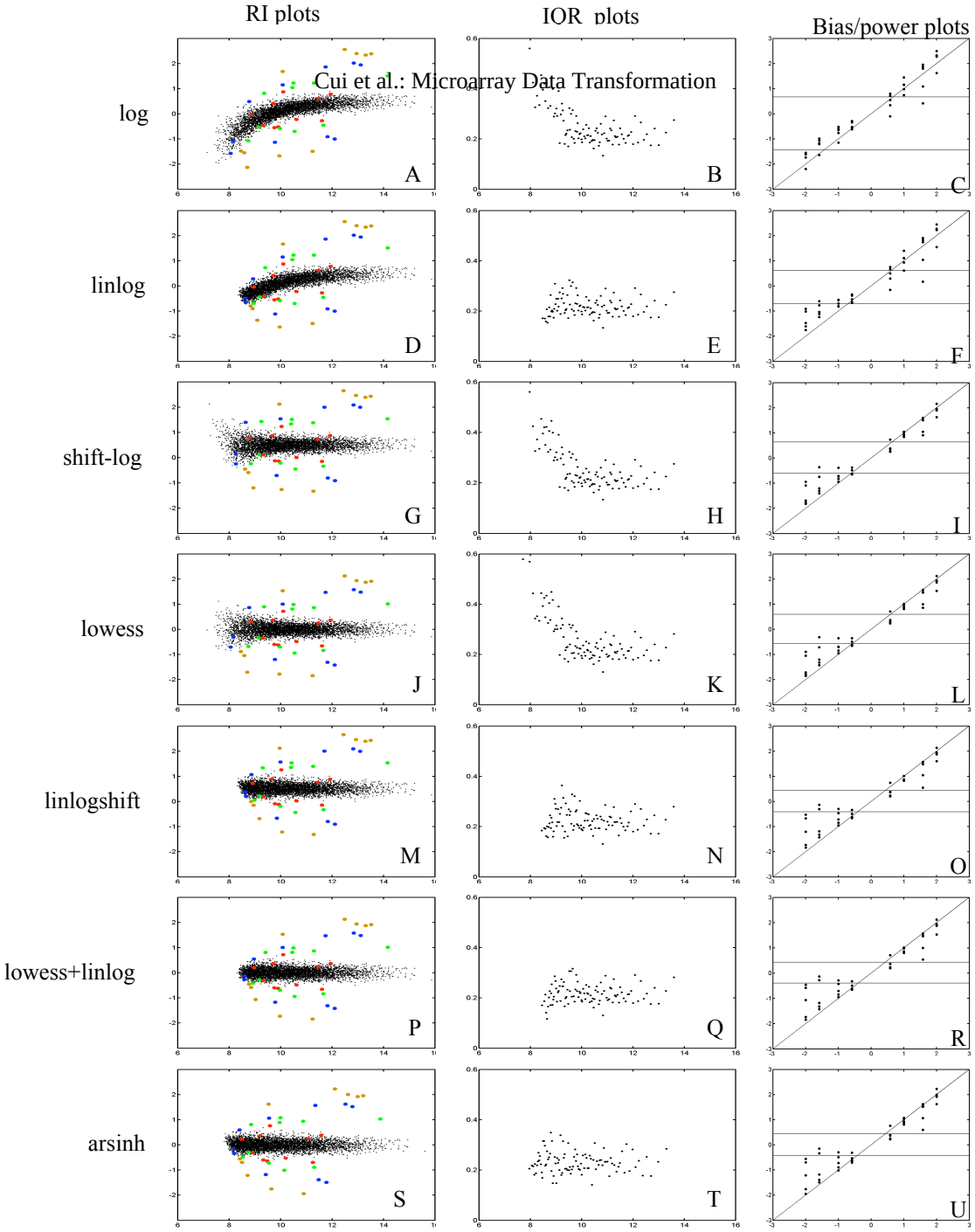

Figure 4: Comparison among transformation methods in reducing curvature, stabilizing variance, and impacting differentially expressed genes. The logarithmic transformation shows both curvature (A) and high variation at low intensity end (B). Linlog only removes high variance at the low intensity end (E, D). Shift-log and lowess only remove curvature (G, H, J, K). Arsinh, linlogshift, and lowess+linlog can remove both curvature (M, P, S) and high variance at low intensity (N, Q, T). Logarithm transformation does not bias the fold changes of differentially expressed genes, but has little power to detect them (C). Other transformations underestimate fold changes but increase the detection power of differential genes (F, I, L, O, R, U). Differentially expressed genes are shown as colored points in the RI plots. Their fold changes are 1.5 (red), 2 (green), 3 (blue), and 4 (brown). IQR plots were generated by dividing all spots into 100 bins with equal number of genes according to intensity. The bias/power plots show the true fold change (x axis) versus estimated fold change after transformation (y axis) for all differentially expressed genes. The two horizontal lines indicate the critical fold changes at type I error of 0.01. The diagonal line is the identity line.

transformation gives a dip at the very low intensity end of the IQR plot (Figure 4E), which indicates slight over-adjustment. In contrast, arsinh transformation gives a more even variance (Figure 4T).

When all the transformations discussed in sections 2.3-2.6 are compared using these simulated data with both curvature and high variation at the low intensity end, it is obvious that the logarithmic transformation shows both curvature and high variance at low intensity spots. Linlog can only stabilize the variance. Shift-log and lowess can only straighten the curvature. Linlogshift, lowess+linlog, and arsinh transformations can straighten the curvature and stabilize the variance. The linlog-related transformations over-adjust the very low intensity end. More examples using real data are shown in supplemental Figure 1 (*http://www.jax.org/staff/churchill/labsite/pubs/index.html*).

Bias and Power To determine the effects of the data transformations on the fold changes of the differentially expressed genes, we included 40 differential genes with five for each of the up and down 1.5, 2, 3, and 4 fold changes among the 5000 simulated genes. After each transformation, their fold changes were recalculated. Results show that the absolute values of the fold changes were slightly decreased by all the transformations except the logarithm (Bias/power plots in Figure 4). To determine the power of each transformation in identifying the true differential genes, we controlled the false positive rate (type I error) at 0.01 by identifying the critical fold change values as the 0.5% and 0.995% percentile fold changes of all the non-differentially expressed genes for each transformation method. The power of each transformation is then represented by the number of true differential genes that show estimated fold changes beyond the critical fold changes (points beyond the two horizontal lines in the Bias/power plots of Figure 4). Compared with the logarithm transformation, the arsinh, linlogshift, and lowess+linlog transformations have the largest power increase followed by shiftlog and lowess. The power increase of linlog transformation is the least due to the presence of large curvature. In summary, these transformation methods greatly increase the power for distinguishing differential genes from the non-differential genes but reduce the fold changes slightly.

2.6 Real Example: Channel Balancing Experiment

When a hybridized slide is scanned, adjusting the laser settings or the photo-multiplier tube amplification is usually recommended in order to obtain overall similar signal strength in the two channels. If the channels are not balanced, the signals in the Cy5 channel are usually weaker than in the Cy3 channel. In order to determine the effects of lack of balance between the two channels, a dye-swap experiment was conducted to compare a mouse mammary tumor RNA sample with a Stratagene universal mouse reference RNA sample (Stratagene Co., CA) using Ontario mouse 15k arrays (UHN Microarray Centre, Toronto, Canada). After hybridization, the two arrays were scanned with balanced and unbalanced photo-multiplier tube amplifications. Data were transformed using logarithm, shift-log, lowess, and arsinh. RI plots are shown in Figure 5. The four transformations give very similar results for the balanced scans, but different results for the unbalanced scans. The large curvatures shown in the RI plots of the logarithmic transformation are corrected by the other three transformations; however, some differences are noticed at the low intensity end.

The transformed data were then normalized and analyzed using ANOVA (Kerr et al., 2000). Correlation coefficients for estimated relative expression differences between the two arrays are shown in Table 2. The correlations among the transformations are much higher for the balanced scans (≥ 0.97) than for the unbalanced scans (0.59 to 0.88). When the transformations are compared for their abilities to remove the effects of unbalanced scanning, lowess, shift-log, and arsinh all give higher correlations between the two scans compared with the logarithmic transformation (0.59), but with different degrees. Lowess improves the correlation to 0.82, followed by shift-log (0.74). Less improvement was obtained from arsinh (0.67). It is obvious that none of these transformations can completely remove the effects of

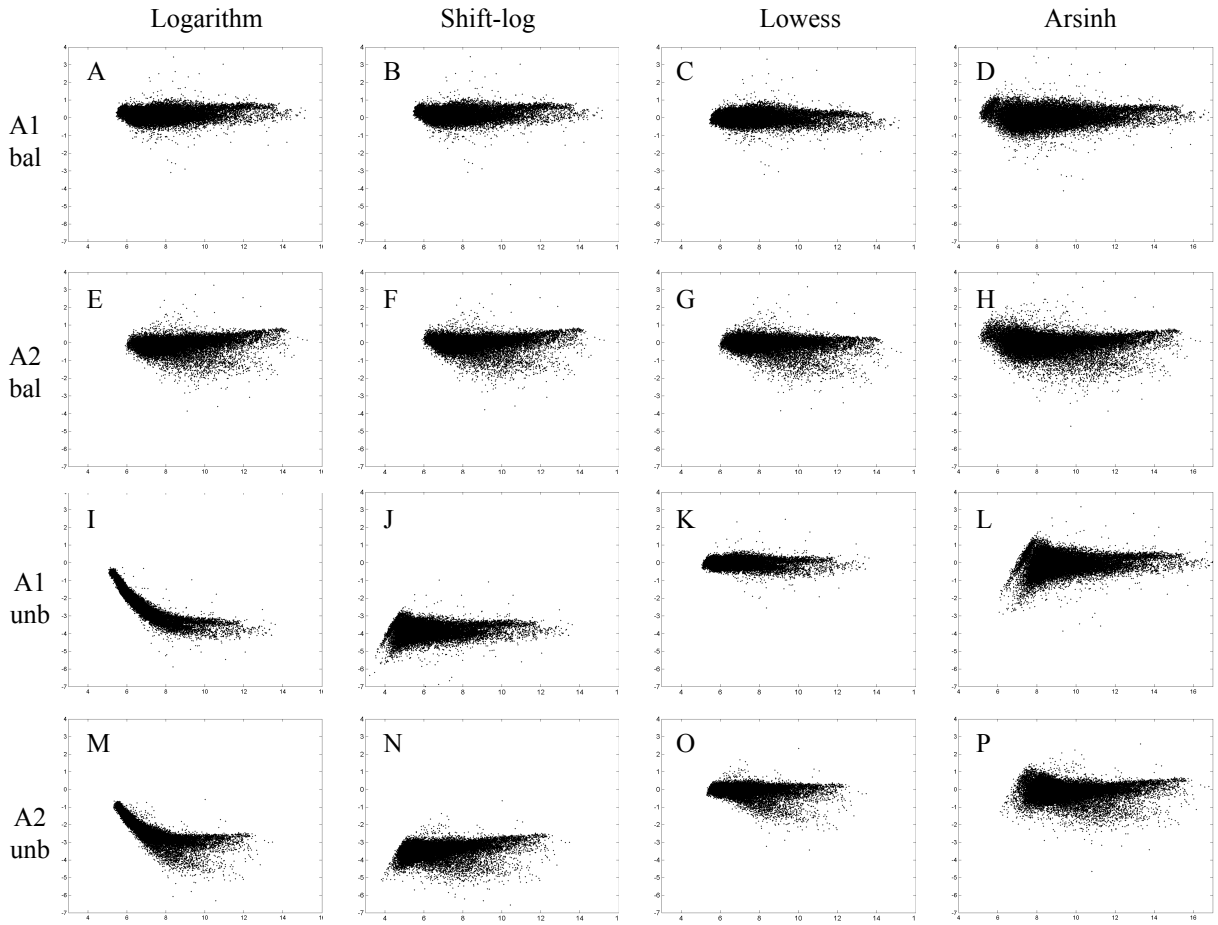

Figure 5: Comparison among logarithm, shift-log, lowess, and arsinh transformations on data from the dye-swap experiment scanned using balanced and unbalanced photo-multiplier tube settings. All transformations give similar results from the balanced scan (A to H), but different results from the unbalanced scan (I to P). For the unbalanced scan, logarithm transformation (I and M) show large curvature. All three other transformations straighten the curvatures. The shift-log transformation (J and N) shows excess variation at low intensity end, so does the arsinh transformation (L and P). The lowess transformation (K and O) gives most balanced-like RI plots. A1, array 1; A2, array 2; bal, balanced scan; unb, unbalanced scan.

	bal-log	bal-L	bal-S	bal-A	unb-log	unb-L	unb-S	unb-A
bal-log	1.00	0.98	0.99	0.97	0.59	0.78	0.75	0.67
bal-L		1.00	0.99	0.97	0.59	0.82	0.72	0.68
bal-S			1.00	0.98	0.60	0.80	0.74	0.69
bal-A				1.00	0.53	0.73	0.67	0.66
unb-log					1.00	0.69	0.87	0.59
unb-L						1.00	0.88	0.90
unb-S							1.00	0.86
unb-A								1.00

Table 2: Pair-wise correlation coefficients among logarithm, shift-log, lowess, and arsinh transformations on data from balanced and unbalanced scans. The different scans were first transformed using logarithm, lowess, shift-log, or arsinh and then normalized using ANOVA. The difference between the transformed data from the two samples were calculated and the correlation coefficients were computed among all the combinations of scans and transformations. bal-log, balanced scan and logarithm transformation; unb-log, unbalanced scan and log transformation; bal-L, balanced scan and lowess transformation; unb-L, unbalanced scan and lowess tranformation; bal-S, balanced scan and shift-log transformation; unb-S, unbalanced scan and shift-log transformation; bal-A, balanced scan and arsinh transformation; unb-A, unbalanced scan and arsinh transformation.

unbalanced scanning. Therefore, we recommend balancing the two channels as much as possible when scanning microarrays and use of one of these transformations to remove the remaining effect.

It is noticeable that the curvatures of the unbalanced scans are downward (Figure 5) while most of the curvatures observed from real data are upward in RI plots with log ratios computed as $\log(\text{Cy5}/\text{Cy3})$. The reason could lie in the fact that the Cy5 channel has smaller slope than the Cy3 channel in equation (2) and the signal difference between the two channels decreases as the spot intensity increases in the unbalanced scan. In the balanced scan, the slope of the Cy5 channel is raised by increasing the laser setting or the photo-multiplier tube amplification to match that of the Cy3 channel. The curvatures in the RI plots are, therefore, reduced. In some cases, the slope of Cy5 is over-adjusted to compensate for the low background of Cy5. This over adjustment would result in upward curvature (Figure 6).

2.7 Removing Spatial Biases in Microarray Data

Up to this point we have focused on intensity-dependent effects on log ratios. As mentioned in section 1.1, any of the parameters in model (2) can depend on the spatial location of the spot on the array surface. Spatial heterogeneity due to lack of uniformity in array printing or hybridization is a common problem. It can be detected by representing the log ratios using a color map and plotting them according to their spot location on the array. We refer to this plot as an *arrayview* plot. A previous study associated spatial variation with the print-tip groups in cDNA array and proposed to apply lowess to each print-tip group (print-tip lowess) (Yang et al., 2002b). However, we find that spatial variation is generally not restricted to the boundaries of print-tip groups.

Because spots are printed on microarrays in a rectangular grid, we can describe the location of a spot on an array by its row and column coordinates. One way to correct for systematic spatial variation is to fit a surface over the array that follows the systematic trend and then re-calibrate the data with respect to this surface. This is directly analogous to the loess approach in correcting intensity-dependent curvature in equation (7) except that the C_k depends on the location of the spot (row, column) instead of the intensity of the spot (I).

Figure 7 demonstrates that intensity-dependent loess does not remove spatial bias while the spatial-dependent loess does not remove intensity-dependent bias. These two steps could be used in sequence

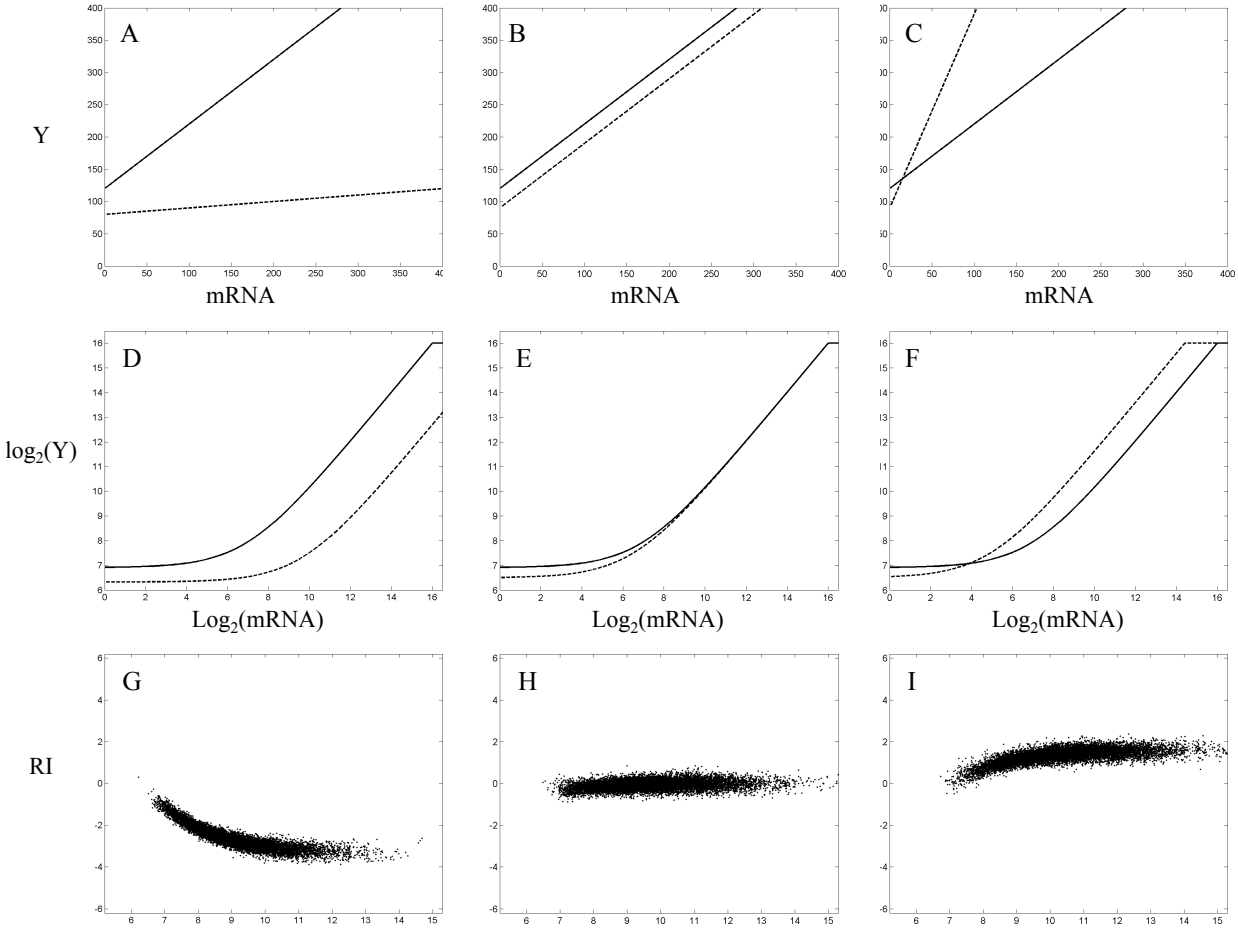

Figure 6: Channel balancing affects curvature direction in RI plots. In an unbalanced scan, the smaller slope of Cy5 channel (A, D) causes the downward curvature in RI plot (G). When the slope of the Cy5 channel is raised to match that of the Cy3 channel (B, E), the curvature of the RI plot is reduced (H). If the slope of the Cy5 channel is over adjusted (C, F), it will result in upward curvature (I). Panels A to C plot the raw fluorescent signal versus mRNA concentration simulated without error using model (1). Only a part of the plot (form origin to 400 at each axis) is shown to illustrate the background difference. The saturation at high intensity spots is not shown in these panels. Panels D to F plot signal intensity versus mRNA concentration at logarithmic scale. Panels G to I are RI plots of the data simulated with error using model (2). Dash lines represent Cy5 channel and solid lines represent Cy3 channel.

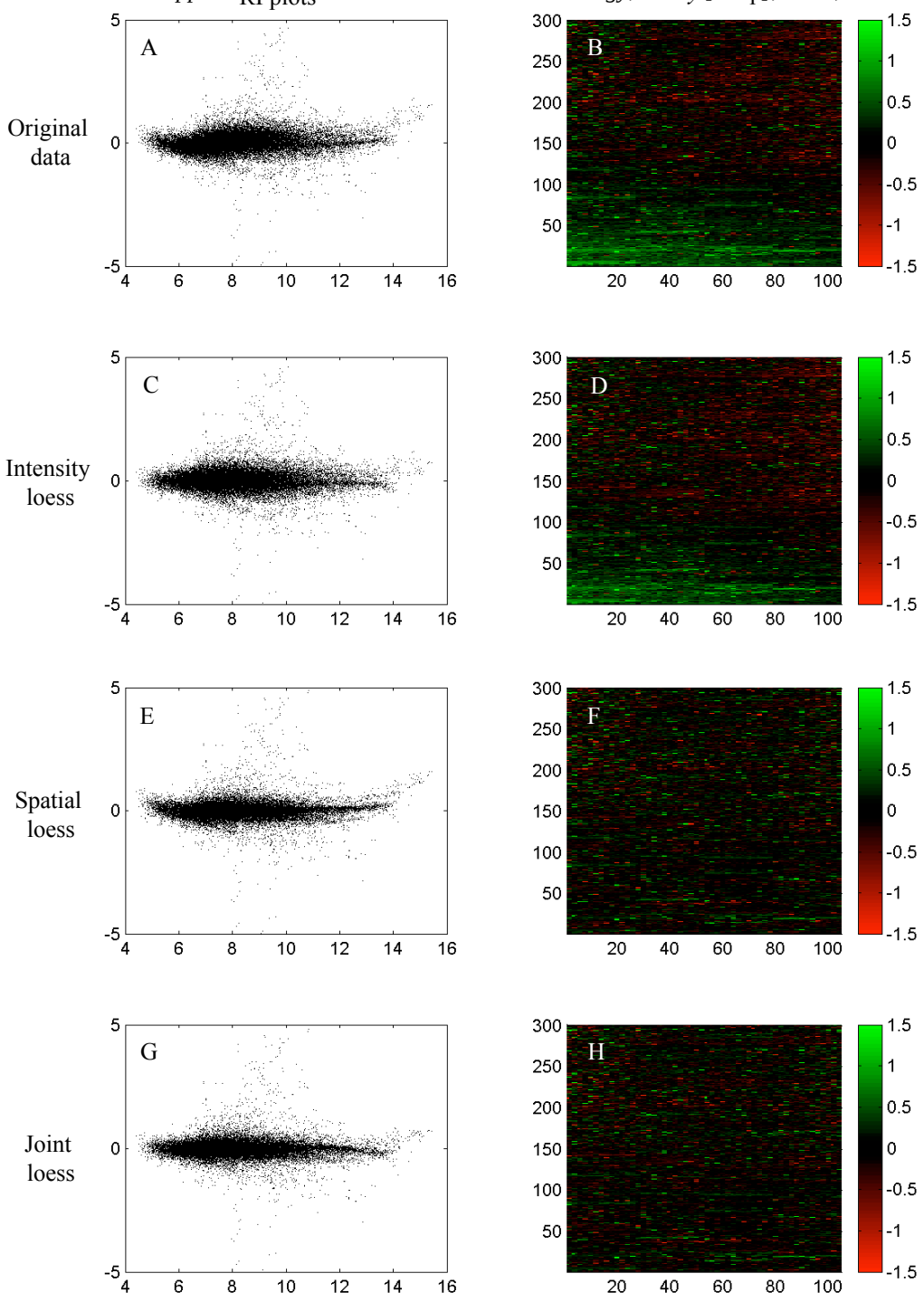

Figure 7: Transformations for removing spatial biases. Intensity loess removes intensity-dependent bias (C), but not spatial bias (D). Spatial loess can remove spatial bias (F) but not intensity-dependent bias (E). Joint loess can simultaneously remove both intensity-dependent and spatial biases (G and H).

Problem	Diagnostic Plot	Symptom	Transformation
Log ratios vary with signal intensity	RI plot	curvature	shift-log lowess arsinh
Variance varies with signal intensity	RI plot IQR plot	wider spread in RI plot, curvature in IQR plot	linlog, arsin
Log ratios and variance vary with signal intensity	RI plot IQR plot	curvatures in both plots	linlogshift lowess+linlog arsinh
Spatial heterogeneity of log ratios	*arrayview* plot	spatial pattern	spatial loess
Spatial heterogeneity and intensity dependence of log ratios	RI plot *arrayview* plot	curvature spatial pattern	joint loess

Table 3: Summary of the recommendations on microarray data transformation.

to remove both. Our approach is to make the two corrections simultaneously by computing the corrected log ratios as a function of both intensity and location. Figures 7G and 7H show that the "joint" loess corrects for both systematic intensity-dependent and spatial variation. More examples are shown in supplemental Figure 2 (*http://www.jax.org/staff/churchill/labsite/pubs/index.html*).

An important assumption of the lowess procedures we have described is that either most genes in a microarray study are not differentially expressed or that the proportion of over- and under-expressed genes between a pair of samples is roughly symmetric. In situations where lots of genes are changing or changes tend to be in one direction, it may be necessary to allocate spots on the slide for a control series where non-differential expression can be assumed. However, this could mean that a substantial proportion of space on the array will have to be dedicated to controls in order to have sufficient coverage of the spatial and intensity range of the array. In addition, all the concerns about using such a nonparametric transformation discussed in the context of the intensity-loess in section 2.3 apply to this procedure.

3 Discussion and Recommendations

For microarray data, correction of spatial and intensity dependent effects on the log ratio are essential to avoid being misled by common artifacts in microarray data. In general it is best to correct biases at the technical level or through clever design rather than to rely on post-hoc data adjustments. Simple precautions such as balancing the photo multiplier tube settings when scanning the arrays can be very effective (Figure 5). Correcting biases at the stage of analysis is undesirable in general because the corrections applied can never be perfectly accurate. The attempted corrections may introduce biases greater than the ones they remove. Nevertheless, we have found that a small arsenal of data transformation tools is essential for reliable microarray data analysis (Table 3). Our advice is to apply the most gentle transformation that corrects the observed problem and to make the judgement call for repeating the experiment when the data are too bad.

The first step toward choosing the right transformation strategy is to diagnose potential problems in

the data. RI plots can reveal intensity dependence of log ratios, which appears as curvature, as well as extra variation at low intensity spots. IQR plots can also be used to diagnose the unstable variance of log ratios. Spatial effects can be detected with an *arrayview* plot.

If there is curvature in the RI plot, either shift-log or lowess can be used to remove it (Figure 3). If the curvature is restricted to the low intensity end, the shift-log transformation will be a good choice. Otherwise, lowess is preferable (Table 1). If extra variation is detected at the low intensity end, you can try linlog or arsinh to remove it. If both curvature and unstable variation exist, you can remove both in one step using linlogshift or arsinh or you can try both lowess and linlog (Figure 4). Stabilizing the variance of log ratios is only important for statistical inferences that assume constant variance across the experiment. For spatial heterogeneity, if it strictly follows print-tip groups, print-tip lowess (Yang et al., 2002b) would be a good choice. Otherwise, spatial loess is recommended (Figure 7). If the log ratios show both intensity dependence and spatial effects, we recommend removing both effects in one step using the joint loess method (Figure 7). No matter which approach is chosen, we recommend applying the same transformation to all of the arrays in an experiment for consistency.

Colophon

We would like to thank Hao Wu for software support; Edward K. Lobenhofer and Cynthia A. Afshari from NIEHS for providing some of the microarray examples. This research is supported by grants CA88327, HL66620, and HL55001 from the National Institute of Health.

The R source code for arsinh transformation (vsn package) can be found at *http://www.bioconductor.org.* All other functions are available in both R and Matlab implementations of the MAANOVA package (Wu et al., 2003) at *http://www.jax.org/staff/churchill/labsite/software/index.html.*

References

Bilban M, L.K. B, Head S, Desoye G, Quaranta V (2002). Normalizing DNA microarray data. *Curr. Issues Mol. Biol.* 4:57–64

Bland M, Altman D (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* i:307–310

Bolstad B, Irizarry R, Astrand M, Speed T (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193

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

Colantuoni C, Henry G, Zegger S, Pevsner J (2002). Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* 32:1316–1320

Dudley AM, Aach J, Steffen MA, Church GM (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *P. Natl. Acad. Sci. USA* 99:7554–7559

Dudoit S, Yang Y, Callow MJ, Speed TP (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html

Durbin B, Hardin J, Hawkins D, Rocke DM (2002). A variance-stabilizing trasformation for gene-expression microarray data. *Bioinformatics* 18:S105–S110

Holder D, Raubertas RF, Pikounis VB, Soper K (2001). Statistical analysis of high density oligonucleotide arrays: a safer approach. *http://128.32.135.2/users/terry/zarray/Affy/GL_Workshop/SAFERv04.pdf*

Hoyle DC, Rattray M, Jupp R, Brass A (2002). Making sense of microarray data distributions. *Bioinformatics* 18:576–584

Huber W, von Heydebreck A, Morgan H, Poustka A, Vingron M (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology* 2:iss 1, art 3

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 1:1–9

Jin W, Riley R, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001). The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nat. Genet.* 29:389–395

Kepler T, Crosby L, Morgan K (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* 3:research0037.1 – 0037.12

Kerr MK, Afshari CA, Bennett L, Bushel B, Martinez J, Walker NJ, Churchill GA (2002). Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sinica* 12:203–217

Kerr MK, Martin M, Churchill GA (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7:819–837

Newton M, Kendziorski C, Richmond C, Blattner F (2001). On differential variablility of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* 8:37–52

Quackenbush J (2001). Computational analysis of microarray data. *Nat. Rev. Genet.* 2:418–427

Quackenbush J (2002). Microarray data normalization and transformation. *Nature Genetics* 32:496–501

Ramdas L, Coombes KR, Baggerly K, Abruzzo L, Highsmith WE, Krogmann T, Hamilton SR, Zhang W (2001). Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology* 2:research0047.1–0047.7

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

Sapir M, Churchill GA (2000). Estimating the posterior probability of differential gene expression from microarray data. *http://www.jax.org/staff/churchill/labsite/pubs/marina.pdf*

Schena M, Shalon D, Davis R, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470

Snedecor GW, Cochran WG (1989). *Statistical Methods*. Ames, Iowa: Iowa State University Press, eighth edition

Tanner JM (1949). Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *J. Appl. Physiol.* 2:1–15

Tusher VG, Tibshirani R, Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98:5116–5121

Wolfinger R, Gibson G, Wolfinger E, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules R (2001). Assessing Gene Significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8:625–637

Wu H, Kerr MK, Cui X, Churchill GA (2003). Maanova: A software package for the analysis of spotted cdna microarray experiments. In: G Parmigiani, ES Garrett, RA Irizarry, SL Zeger (eds.), *The analysis of gene expression data: methods and software*. New York: Springer

Yang M, Ruan Q, Yang J, Eckenrode S, Wu S, McIndoe RA, She JX (2001). A statistical procedure for flagging weak spots greatly improves normalization and ratio estimates in microarray experiments. *Physiol. Genomics* 7:45–53

Yang YH, Buckley MJ, Dudoit S, Speed TP (2002a). Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Stat.* 11:1–29

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002b). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30:e15

Zien A, Aigner T, Zimmer R, Lengauer T (2001). Centralization: a new method for the normalization of gene expression data. *Bioinformatics* 17 Suppl.:S323–S331