NORMALIZATION AND SYSTEMATIC MEASUREMENT ERROR IN cDNA MICROARRAY DATA

David B. Finkelstein[1], Jeremy Gollub[1], J. Michael Cherry[2]
[1]Carnegie Institution of Washington, Department of Plant Biology, Stanford, CA 94305
[2]Stanford University, Department of Genetics, Stanford, CA 94305

KEY WORDS: Gene Expression, Measurement Error, Microarray, Normalization, Shrinkage, Soft-Thresholding

cDNA Microarray Data

Data collected from cDNA Microarrays, like all data from techniques have some measurement error. However cDNA much of the measurement in cDNAs is not random or normally distributed but systematic. Physical parameters inherent in the production of microarrays can be used to approximately measure these systematic effects. Although the causes of these effects are not fully understood their influence can be corrected mathematically. By building a model of gene expression, which is based on a few fundamental assumptions, a method for correction was developed and is presented here. As long as the assumptions are valid, the methods for correcting data should substantially reduce measurement error

Normalization

The term normalization has a meaning that is unique to gene expression-profiling. That is the re-scaling and correction of two data sets prior to comparison. Most often this is between the two mean pixel intensities of the two fluor channels. The term normalization does not necessarily refer to the assumptions of normality.

There are many methods of normalization in gene expression-profiling. Each method implicitly or explicitly makes assumptions about the biology of gene expression ((Finkelstein, Gollub et al. 2001), submitted). Generally, normalization methods calculate a scaling factor or function to correct intensity effects. These factors or functions are then applied to the measure of relative abundance to produce normalized ratios or scaled intensities. Normalization approaches may be based on mean correction (Richmond and Somerville 2000), linear regression ((Finkelstein, Gollub et al. 2000) submitted), nonlinear models (Yang, Dudoit et al. 2000), linear combinations of factors (SVD)(Alter, Brown et al. 2000; Richmond and Somerville 2000) or even Bayesian methods (Newton, Kendziorski et al. 2001).

Most normalization methods effectively reduce the variation of expression measures (Richmond and Somerville 2000). However, what is crucial is not the amount of variance, but the treatment of outliers, which may generate false positives. From a biological viewpoint, the highly probable average case is of no interest. Experimentalists are interested in and will pursue only the exceptional cases. Mean or median correction methods may unintentionally introduce false positives. Since these simple methods best correct for the average or median case and not the biologically crucial outliers. These methods may unwittingly introduce false positives once ratios are calculated. More sophisticated methods that employ non-linear regression or lowess, while designed to predict the average or median case (Yang, Dudoit et al. 2000), frequently do reduce the number of false positives. These methods empirically determine a nonlinear function of intensity for each data set and correct the data to produce a linear function.

A Model of Gene Expression

Given a few assumptions that are consistent with biological observations, a mathematical model of gene expression can be developed. The foremost underlying assumption is that the typical gene in a given experiment is unresponsive. In mathematical terms this means that the ratio of the two measures of gene expression, representing the control and treated samples, should be one for the typical gene. It follows that the mean or median log-transformed expression ratio of a large population of cDNAs is expected to be zero. Except for the few responsive genes, the mathematical relationship of the level of expression under the control state is said to be equal to that under the test condition. Or, in the terms most often used to measure cDNA microarrays, the mean pixel intensity of a fluor labeled cDNA hybridized to a given spot should be equal to the measurement of a second fluor labeled cDNA hybridized to the same spot.

There are two fundamental reasons why this equality may not be observed for a given spot. One, the gene measured by the spot of interest may be authentically differentially expressed. Two, measurement error or systematic effects from technical causes may produce inaccurate data. Normalization is intended to identify, correct and remove these inaccuracies.

Another assumption made here is that genes that are authentically differential are sufficiently distinct from the general population to be readily detected. This assumption may fail for two reasons. One, if the differential expression is modest in comparison to measurement error then it may not be detected. Two, for a highly specialized microarray with

a small population of genes, the typical gene may be responsive to the test conditions. In this rare case the movement of the mass of genes could not be distinguished from technical error without reference to control genes.

It is also assumed here that a rare gene is just as likely to respond to a given response as a highly expressed gene. That is, differential gene expression is independent of transcript abundance and therefore signal intensity. This means that not just the population mean of the genes in the treated sample is expected to be equivalent to that of the control sample. Under this assumption the level of expression under the two conditions are equal along the entire range of abundance. Finally, it is assumed that a negative control gene is expected to have zero signal intensity. A negative measure is an artifact of technique as negative transcripts are biologically infeasible. This assumption is the basis for the additive correction in our procedure.

From these assumptions, we expect that our observed signal should be linear with respect to transcript abundance. If the transcript abundance of gene in a treated RNA sample is y_i and the transcript abundance of the control RNA sample is x_i, the formula is $y_i = mx_i + b$. The expected slope (m) is one and the expected intercept (b) is zero. The range of expected values (x,y) is from zero to some unknown positive value. Any gene that deviates from the linear model is expected to represent a differentially expressed gene. If the population as a whole is nonlinear or the slope and intercept are other than expected, then this is presumed due to measurement error or systematic technical effects.

Two color array measurements

The data from a spotted cDNA microarray is derived from scanned and digitized images. The images are produced after hybridizing two distinct fluorescently-labeled cDNA pools to known DNA fragments spotted and fixed to a specially coated glass slide (Schena, Shalon et al. 1995). In a customary experiment one cDNA pool represents the reference or control and the other represents the treated sample. After hybridization, the fluorescently labeled cDNAs are excited by one or more lasers and detected by a photo-multiplier tube or charged-coupled device camera resulting in two digital images. For two-color arrays these images are aligned and the data extracted with specialized software (Bassett, Eisen et al. 1999). The final data is the mean or median pixel intensity of each spot for each image or "channel". The change in expression for each transcript is commonly reported as a $\log_2$-ratio of the mean channel 2 pixel intensity over the mean channel 1 pixel intensity for each spot, after background corrections and normalization. If background correction is imperfect then additive error may bias the result.

In cDNA microarray experiments uneven printing, scanning or hybridization may result in a spatial bias. Provided that cDNAs were not printed in blocks based on function, there should be no correlation between spot placement and expression. This correlation is termed spatial bias.

In cDNA spotted arrays technical variation can also be correlated to PCR or printing plate. The cDNAs from each 384 or 96- well plate were PCR amplified separately. Any differences in DNA or salt concentration may bias the final expression ratios. For some arrays printing plates were designed to represent functional groups of genes rather than random samples. If this is true then plate bias correction is inappropriate.

Two methods are currently used to detect plate and spatial bias. The first method is an application of ANOVA. The model is that log-ratio is a function of the categorical variable plate or sector (which corresponds to pin). ANOVA provides a measure of the significance of the relation (F statistic) and the strength or degree of the problem (r-squared). These tests presume normality, which is generally valid for log-ratios within an array. The test also presumes equal variances of the log-ratios from plate to plate or sector to sector. This assumption is frequently invalid and therefore ANOVA may under-report bias. If the Kruskal-Wallis test is used the assumption of equal variances can be discarded. This nonparametric F test is based on the Chi distribution and rank differences (Kruskal and Wallis 1952). This test provides a measure of the significance of the effect but not a readily interpreted measure of strength.

In a survey of 199 public AFGC microarrays the spatial and plate bias was measured. For 86% of all arrays had a sector bias that with less than a r squared of 0.18. Of these same 199 microarrays, 81% had a plate bias with an r squared of less than 0.18. It was observed that these biases were not correlated, as expected. However, a high sector bias generally resulted in a low plate bias and vice versa (Figure1). That is whenever one bias was high the other bias was generally low or very low. This result is also expected as the detection plate bias is masked by sector bias and vice versa.

Mechanically, the genes within a plate are evenly distributed among sectors. Thus a strong plate wide effect has the unintended consequence of increasing the correlation of sectors. This explains the observed increase in plate bias in data sets normalized by sector alone (Finkelstein, D. B., J. Gollub, et al. 2000). The effect is visible in Figure 1. where the r squared statistics for the ANOVAs from spatial tests

are plotted against those for ANOVAs used to detect plate bias.

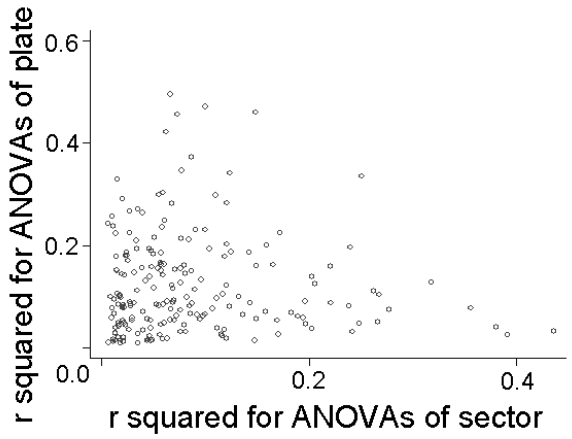

Figure 1. Comparison of plate and sector effects

While sector and plate biases are not correlated the ability to detect high plate biases masked by high sector bias.

METHODS

The methods outlined here are modular. That is, we present several types of correction that may all be used on a single data set or alternatively only one or two of the corrective methods need be applied. In the results and discussion section we compare the modules and make specific recommendations on their comparative use and merits.

Range Correction by Soft-Thresholding

Additive error correction, or range correction, in our procedure is based on the model that each intensity measurement is the sum of the authentic signal function (abundance) plus local additive error (background) and global additive error. Ideally, the function of abundance is linear, although nonlinear cases have been reported (Yang, Dudoit et al. 2000) and a nonlinear example will be discussed here.

$Signal_{observed}$ = f(abundance) + $background_{spot\text{-}specfific}$ + additive $error_{global}$

For cDNA microarrays local additive error is typically removed by using a local spot-specific background correction. The global additive error is corrected by setting the low range of all signals to zero. By adding or subtracting one number to all background corrected intensities the data becomes much more log-normal. This is essentially a soft-thresholding method. The choice of the corrective value or shrinkage factor is critical. A shrinkage factor may be based on the empirical measurements of negative controls or lowest 1st percentile. Alternatively, the shrinkage factor that produces the most log-normal data may be chosen. In our procedure log normality is based on kurtosis alone as skew may to be due to biological causes.

The underlying assumption in global range correction is that any number attributed to a negative control is purely due to technical causes. Clearly, a negative number for transcript abundance is not possible in biological terms. The problems caused by ignoring small positive additive error are less evident. Even small positive values when added to ratios may greatly influence the entire data set, and introduce heterostochasticity and non-linearity (Poynton 1993). This is especially if the global additive errors are substantially different in one channel or the other. In the Figure 2 the nonlinear example ($y = x + x^2$ fits r-squared = 0.996) is a product of additive error. After removing the bias based on the log-normality of the net channel intensities (Table 1) the data was plotted in figure 3. Note the now linear form.

Table 1.

Channel 1 (Cy3)	Range correction	skew	kurtosis	observations
minimum	+73	0.08	2.64	11903
optimum (10th lowest)	-146	-0.20	2.97	11894
1st percentile	-327	-0.61	3.97	11783

Channel 2 (Cy5)	Range correction	skew	kurtosis	observations
minimum	+386	0.07	2.39	11901
optimum (10th lowest)	+167	-0.41	2.91	11892
1st percentile	+45	-0.78	3.89	11782

Note 3.0 is the kurtosis for a perfectly normal distribution.

In our procedure, the range correction is determined in an iterative process. First the minimum net intensity is found for a given channel. If this value is greater then zero then the minimum is subtracted from all spots. If the minimum value is negative then this value is added to all spots. The corrected values are log-transformed and their kurtosis is tested. Kurtosis for a perfectly normal distribution is three. If kurtosis is larger than 3.0 (+ or - 0.05) then a larger number must be subtracted or a smaller number added. This may cause some data points to become negative values. As a result these data points are lost because they can not be log-transformed. We recommend that no more than 1% of the data that should be removed in this way. In fact one advantage in using a range correction method is that generally fewer spots are

discarded after correction than would be from uncorrected spots.

Kurtosis = $[\sum(\text{observation-mean})^4]/N(\text{variance})^2] - 3$
where N is the number of observations.
Variance = $[\sum(\text{observation-mean})^2]/(N-1)$

If Kurtosis is below 3.0 then either a smaller number must be subtracted or a larger number added. Our practice is to rank the spots by intensity and set the additive correction so that each subsequent gene is set to zero starting with the minimum spot. If this correction fails then the values for the second lowest net intensity spot is used.

Interestingly, this simple additive method works well for SAGE data but not for GeneChip® data (Affymetrix, Santa Clara), membrane data or spotted microarray data that was not background corrected. In each case, global correction may improve the normality of the log-transformed data, but it cannot achieve perfect log-normality in our experience. It is presumed that feature or spot-specific error is large enough to prevent full correction. When full correction is possible each successively larger addition increases the kurtosis (relative peak height) of the distribution and each successively larger subtraction reduces the kurtosis. For two-color microarray, the kurtosis is typically high ((Finkelstein, Ewing et al. 2001), in press) before correction so that a small number must be subtracted.

In both the GeneChip® data, membrane data (Andrews, Bouffard et al. 2000) background correction is not local but by sector or subarray. However, the GeneChip® data is more complex as average differences reflect the difference in signal between several pairs of specific (perfect match) and non-specific (single mismatch) oligos (Lockhart, Dong et al. 1996). In our hands, the population of independent probe pairs cannot be fully log-normalized by additive error. This indicates that the difference between perfect match oligo and mismatch oligo signals is the cause of sufficient oligo-specific error to frustrate any attempt to find a fully normalizing global correction.

Systematic Bias Correction

Two-color arrays are robotically printed from a series of 384 well plates with a set of slotted pins. Each pin produces each own sector and each pin is used to spot clones from each plate. So, clones in each plate are evenly dispersed throughout the sectors. Therefore these two biases are expected to be independent. However, cases have been found for which correcting spatial bias by sector increased the apparent level of plate bias (J. Gollub, unpublished). Whether this is due to an interaction or simply due to improved detection of plate bias in the absence of sector bias is unclear. In any case, the need to remove both biases is apparent.

The underlying assumption in the removal of both plate and sector bias is that the clones in each plate or sector are a random representative sample of the whole population of clones on an array. This assumption is certainly not true for some arrays that are deliberately printed in functional groups (e.g., Atlas ™ Arrays, Clontech Laboratories, Inc., Palo Alto, CA) or for arrays where each plate is a functional class of genes. In these cases any plate or sector bias cannot be distinguished from authentic biology and these correction should not be made.

Given that clones within plates and sectors were randomly selected, then it follows that the distribution of signals within a plate should be representative of the whole. In practical terms this means that the mean and standard deviation of log-normal signals should be very similar. Deviations in these measures indicate bias and are used to remove them. Note that our detection of plate and sector bias is currently achieved through ANOVA tests, which are based on means. If we were to correct plate and sector bias by means only, we would be unable to detect it. However, as mentioned above, means methods may not handle extreme values well. A more complete correction method must consider the distribution of log-signals within each sector or plate. It is also worth noting that if the spatial effects are on a finer scale than sector, so this method may not detect or fully correct this bias.

If each plate or sector is truly representative, then a log-signal in the first percentile of the global population should also be close to the first percentile in each sector or plate population. Since signals are log-normalized by additive correction, the percentile of each gene for each channel can be found using Fisher's Z transformation (Ott 1988). The difference in Z for the global population and the local plate or sector difference can be found for each clone. The difference in means is then calculated and corrected serially, after Z difference correction. Both corrections achieve a fully corrected spatial and plate biases. The importance of a Z correction is that outliers are handled proportionally. That is, outliers in the global population are preserved. All corrections are performed on log-transformed data as only additively corrected log-data is normally distributed.

First the mean and standard deviations of the signals for each plate and each sector must be found. This must be performed on each channel separately so that for each spot the corrections for the Cy5 channel are independent of the corrections for Cy3 channel. Next the global standard deviation and mean must be found. From these values it is possible to calculate the global Z of each signal and the local Z of that signal.

One set of Z differences is calculated for plate and another for sector. The difference is found and multiplied by the local (plate or sector) standard deviation to remove the variability introduced by this number and to put the correction on the proper scale. This z difference is then added to each spot.

Z_{local} = [log(signal)-$mean_{local}$]/standard $deviation_{local}$
Z_{global} = [log(signal)-$mean_{global}$]/standard $deviation_{global}$
$Z_{difference}$ = Standard $deviation_{local}$[Z_{global}- Z_{local}]
log-signal $_{Zcorrected}$ = log(signal) + $Z_{difference\text{-}plate}$ + $Z_{difference\text{-}sector}$

Next the means for the Z-corrected signals are found for each plate and sector and the global mean of the Z-corrected log-signals is also found for each channel. Then the difference between means for each plate and sector is found and added to the signal.

$mean_{difference}$ = $mean_{global}$ of log-signal $_{Zcorrected}$ - $mean_{local}$ of log-signal $_{Zcorrected}$
log-signal $_{fully\text{-}corrected}$ = $\log(signal)_{Zcorrected}$ + $mean_{difference\text{-}\mathbf{plate}}$ + $mean_{difference\text{-}\mathbf{sector}}$

For two-color arrays each spot will have two fully corrected log-signals. Prior to creating a ratio it is still necessary to find a regression function to standardize the channels. Now that the data has been normalized by additive correction the regression of log-signal cy3 vs log-signal cy5 is valid. Once the slope and intercept are found the log-signal of cy5 is altered so that the regression function is equivalent to y = x. Log ratios are created from Cartesian signals. Log_2-ratios are adjusted from natural log ratios by division of log(2).

Regression model:

log-signal of Cy3 $_{fully\text{-}corrected}$ =
$B_{intercept}$ + B_{slope}(log-signal $Cy5_{fully\text{-}corrected}$)
Normalization:

log-signal $Cy5_{standardized}$=
$B_{intercept}$ + B_{slope}(log-signal $Cy5_{fully\text{-}corrected}$)

log-ratios $(Cy5/Cy3)_{standardized}$=
log[signal of $Cy5_{fully\text{-}corrected}$ /signal of Cy3 $_{fully\text{-}corrected}$]

log_2-ratios $(Cy5/Cy3)_{standardized}$ =
log-ratios $(Cy5/Cy3)_{standardized}$/log[2]

An Example

The experiment in this nonlinear example is publicly available on the Stanford Microarray database (SMD)(Sherlock, Hernandez-Boussard et al. 2001) Figure 1. The experiment is an *Arabidopsis thaliana*

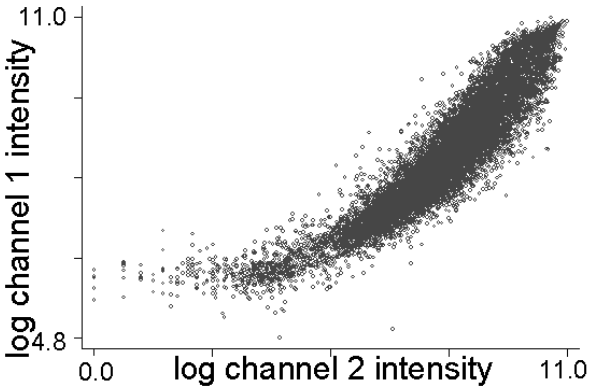

Figure 1. Background Corrected Data
This nonlinear data was locally background corrected then log transformed.

microarray ID: 7561. The experimental treatment was a 16-day elevated $C0_2$ gas exposure at twice ambient levels (800 ppm). The leaves of the plants were harvested, RNA extracted, labeled, hybridized to the arrays, washed, and scanned (complete details are posted at http://afgc.stanford.edu/afgc_html/site2.htm). Data was extracted with Genepix Pro 3.0 (Axon, Palo Alto).

Table 2.
ANOVA Model

y = x	df	F	Prob > F	R-squared
range corrected log(Cy3) = plate				
	30	42.42	0.0000	0.0969
range corrected log(Cy3) = sector				
31	26.94	0.0000	0.0658	
range corrected log(Cy5) = plate				
30	47.43	0.0000	0.1071	
range corrected log(Cy5) = sector				
31	18.54	0.0000	0.0462	
Z-corrected log(Cy3) = plate				
30	31.10	0.0000	0.0729	
Z-corrected log(Cy3) = sector				
31	18.11	0.0000	0.0452	
Z-corrected log(Cy5) = plate				
30	**32.36**	**0.0000**	**0.0757**	
Z-corrected log(Cy5) = sector				
31	13.09	0.0000	0.0331	
All corrections log(Cy3) = plate				
30	0.00	1.0000	0.0000	
All corrections log(Cy3) = sector				
31	0.00	1.0000	0.0000	
All corrections log(Cy5) = plate				
30	0.00	1.0000	0.0000	
All corrections log(Cy5) = plate				
31	0.0	1.0000	0.0000	

Each line represents the results of a separate ANOVA test. Plate and sector are treated as categorical variables, rather than continuos variables.After the experimental data was additively corrected (Figure 2), the means and standard

deviations of each sector and each plate was found for each channel. Then Z corrections were calculated and applied to the data Figure 3. Note that the changes from Figure 2 to Figure 3 appear subtle, however the scale has been substantially increased. Plate and sector bias was detected using ANOVAs where the model was y = log-signal and x = plate or sector. Table 2 shows the changes in plate and sector bias as detected with ANOVAs as correction progresses. The original biases are both below r-squared of 0.1, but above 0.01. Z correction reduces but does not eliminate both biases, only final means correction brings the r-squared and F statistic to 0.0. Figure 4 shows the data after complete means correction. Figure 5 shows that data without Z correction.

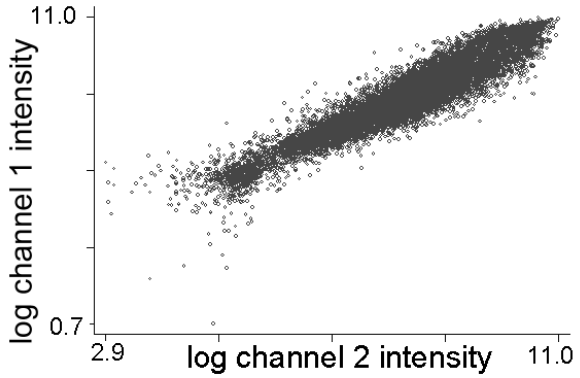

Figure 2. Data After Range Correction
A global number was subtracted or added all data points in each channel the data was log transformed. This corrects the range bias and linearizes the data.

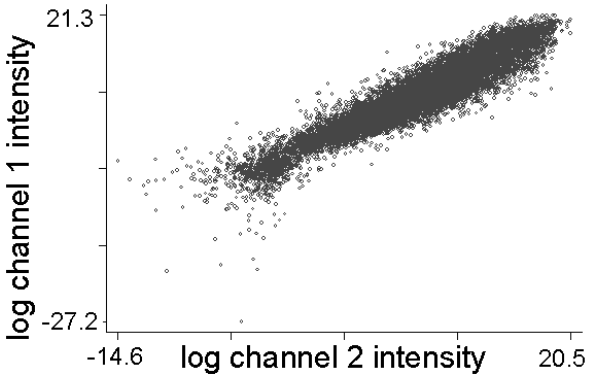

Figure 3. After Z correction and Range Correction
After range correction the Z correction was performed. Cautionary note: this correction greatly increases the scale. The value of this correction is unproven.

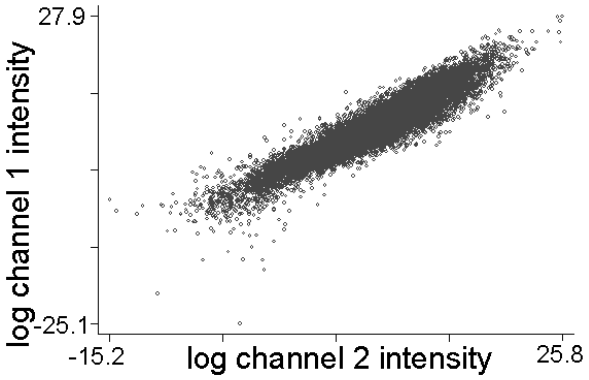

Figure 4. After Z , Range and Mean Corrections
After Z correction the mean corrections remove the influence of sector and plate biases.

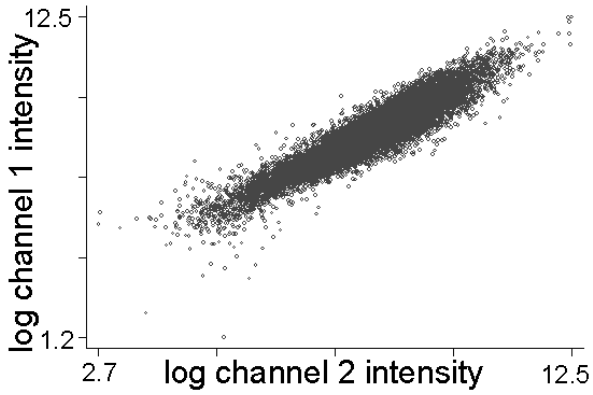

Figure 5. Range and Mean Corrected Data
After range correction the mean corrections remove the influence of sector and plate biases. In this case no Z correction was applied and the scale was not increased. Note this combination of corrections is the method we recommend.

RESULTS

In order to compare the value of each step in the process after each correction the log channel intensities were normalized using simple linear regression. That is, the channel 2 values were multiplied by the least squares slope and then the intercept was added so that the final relationship of log channel 1 to log channel 2 was equivalent to y = x. Then these log values were exponentiated (returned to the original Cartesian scale). Finally, the normalized channel 2 value was divided by the channel 1 values and log base 2 transformed.

Each correction step changed the range, distribution, standard deviation and mean of the resultant $\log_2$ratios. (Table 3.).

Table 3. Changes in log_2ratios due to corrections

Log_2ratios range	mean	Std dev.	Skew	Kurtosis
Background Corrected:				
-8.9 to 4.0	-0.3	1.03	-1.93	11.65
Background and Range Corrected:				
-3.6 to 6.65	4.9e-07	0.63	0.11	4.96
Background, Range and Z corrected:				
-23.0 to 19.9	5.5	6.64	-0.44	2.95
Background, Z, Range and Mean Corrected:				
-14.3 to 32.4	3.4e-07	2.83	0.28	5.8
Background, Range and Mean Corrected:				
-3.1 to 6.5	5.4e-07	0.61	0.18	5.16

It is important to note that, although the Z correction created a nearly normal distribution, it greatly altered the mean and the range of values. This correction is not a valid stopping point because, although it corrects for the shape and deviation distribution of the individual plates and sectors it is still not corrected for the means (Table 2). This is made clear by Table 4 where after Z correction (**BRZ**) the correlate very poorly with the other methods although this is corrected once the mean correction is also applied (**BRZM).**

The other clear result from both table 3 and 4 that range correction is the most effective single correction in eliminating skew, kurtosis and centering the mean around zero. The range and mean corrections together eliminate the plate and sector biases, but do not create the change in scale seen in the Z correction. Yet this set of correction is highly correlated r = 0.99 to method that includes Z correction. Beyond the change in scale it appears that the Z correction has little comparative effect on the population of log_2ratios.

Table 4. Correlation of log_2ratios

	B	BR	BRM	BRZM	BRZ
B	1.0				
BR	0.55	1.00			
BRM	0.51	0.96	1.00		
BRZM	0.5	0.95	0.99	1.00	
BRZ	0.69	0.04	0.04	0.06	1.00

B denotes background correction. **R** denotes range correction. **M** is for mean correction and **Z** is for Z correction. **BR** denotes both background correction and range correction were applied.

Table 5. gives specific examples of genes and how each correction log_2ratio. The first, most notable result from the comparison of the highest and lowest-expressed genes is that only after range correction were the genes related to photosynthesis the most changed, as expected (Cheng, Moore et al. 1998). Furthermore far more genes associated with carbon assimilation were present in the highest and lowest-expressed genes after range correction. In fact, only one carbon metabolism gene that was listed as highly down-regulated by elevated carbon dioxide in the background only corrected data (**B**) was not in the range corrected data. This gene was isocitrate lyase. Isocitrate lyase is a well-known juvenility gene that assists oil breakdown in developing seeds (Eastmond and Graham 2001; Rylott, Hooks et al. 2001). Sixteen day old plants are expected to have only low levels of this gene. Changes in a rare juvenile specific gene in mature plants may indicate that isocitrate lyase is a false positive. Others have observed unexpected behavior of this gene on similar arrays for unrelated experiments (Katrina Ramonell, pers. com). In the fully corrected expression ratios this gene was altered, but to a much lower degree.

Another possible false positive is the BT gene, which is a bacterial gene that is not present in *Arabidopsis* and has no homolog. Again this gene was high in background corrected data that was directly normalized and low in any method that includes the range correction. Note that with the mean correction had only modest effects in general while the Z correction greatly altered the range of the data.

Also of interest is the fact that cytochrome P450 proteins are reported as down-regulated with elevated carbon dioxide by the full correction method and hemolysin, a protein that may degrade these P450 proteins (Armstrong and Renton 1994), is now reported as highly up-regulated. Furthermore carbonic anhydrase was up with a log_2ratio of 2.22 by the regression-only method is now up with a log_2ratio of 10.08. This gene is known to increase in response to elevated carbon dioxide in Arabidopsis (Raines, Horsnell et al. 1992).

Table 5. Changes in Specific Genes

Stanford Unique ID	Gene Description	BRZ M	log_2ratio B	BRM	BR
143603	GH3 like protein	1.6	-8.7	0.2	-0.01
139641	isocitrate lyase	-1.6	-8.0	-0.3	-0.5
227329	BT	-2.4	-8.0	-0.6	-0.5
143102	carbonic anhydrase	9.6	2.2	2.1	2.1
143048	hemolysin	11.7	3.0	2.5	2.6
133336	HMG-coA synthase	12.1	3.0	2.6	2.8
133996	photosystemII 44kd	-12.2	-3.1	-2.7	-2.9
135832	cytochrome P450	-11.5	-3.4	-2.5	-2.7
141998	pathogenesis related protein*	4.0	-8.4	0.7	0.5
142648	putative amino-transferase*	8.1	-8.0	1.6	1.2

* These genes are in the lowest percentile for Cy3 intensity.

BRZM are the ratios calculated from data corrected by background, range, Z correction and mean correction. **B** are the log_2ratios calculated from background corrected data only. **BRM** are the log_2ratios calculated by background, range and mean

corrections. **BR** are the $\log_2$ratios calculated for background and range only.

DISCUSSION

Corrections performed on this data set were based on a linear model of gene expression. Each subsequent correction is based on the removal of errors regularly detected on real microarrays. The additive range correction is the most intuitive correction and the most valuable. It is clear that negative values are artifacts and that high values for negative controls are likewise due to technical causes. The resulting improved log-normality of the data is a fortuitous result. This method should attenuate the high variance in measurement observed at low intensities.

In principle, any method that accounts for nonlinear the data with respect to intensity will have some of the same beneficial effects that range correction provides. Conceptually, however an additive correction is superior to a nonlinear fitting method. This correction is based on the fundamental biology and can be derived empirically with the use of negative controls. It is essentially a global background correction based on biological principles. This range should also allow data to be compared across arrays, as long as the same negative controls are employed.

In the absence of negative controls, choosing the additive correction that results in the most log-normal distribution requires assuming that log-normality is correct. In biological terms, this implies that the majority of genes are expressed at moderately low range with a relatively small number of highly expressed genes and rare genes.

Unlike the range correction the mean and Z corrections were intended to address the plate and sector bias of this example. As these biases were in the modest but typical range (r squared less than 0.1), we expect that changes would also be modest. This was the case for mean corrected data. However, the effect of the Z method does not appear to be in proportion to the bias it is intended to adjust.

The use of a simple means correction alone appears to be superior to the means correction coupled with the Z correction. The Z correction alters the scale of the data so that it agrees with expected ranges of expression from Northerns (Schena, Shalon et al. 1995). However, this effect may not be consistent and reliable. In order to validate this scale change numerous Northerns would have to be tried. Furthermore the Z correction is computationally heavy compared with the mean correction alone. In addition the Z correction does not fully correct for the plate and spatial biases and must be coupled with the mean correction (Table 3). Lastly since these results correlate (r = 0.99) to the results from background, mean and range corrected data, we see no reason to recommend the Z correction at this time. Only if the increase in range were proven stable and verified by empirical means would it be acceptable. It is included here as an example of a method, which can be derived from reasonable principles, yet is not practical. It also highlights the limitations of mathematical analysis and the temptation to over correction.

Another difficulty in mathematical correction of biases is the accurate measurement of the bias. Systematic biases such as sector and plate bias may not be ideal factors for predicting spatial error or DNA-related error. Spatial bias may be subtler than the pin level sector and plate bias may be related to DNA concentration or another clone-dependent continuous variable (Finkelstein, unpublished). Despite this coarseness, the choice of printing plate and sector is based on the physical methods used in microarray production. A single printing pin produces each sector and the clones from each printing plate are evenly distributed among the sectors. For any spatial scale there is some probability that an authentic functional group may be printed together purely by chance. The finer the scale the greater the probability that this may occur. It is also important to note that the exact probability can not be known without knowing *a priori* how many genes will respond to a given stress.

Despite these problems we can envision an alternative to the sector method by finding the mean corrections for each row and each column. Each spot would be a member of unique pair of large presumably representative classes. This method should increase the scale while avoiding the risk of functional groups occurring by row. Furthermore a row column method should find edge effects well compared to sector methods. This method is currently under investigation.

However there are concerns about the row column approach. First all plates are not represented in all columns or rows. The plate bias may influence the row and column biases. Also methods that treat the sector, rows and columns as separate entities ignore trends within a slide. Methods for handling trends over a two-dimensional surface are also under investigation (J. Gollub, unpublished).

No matter how effective any normalization method may be there is some point where the error in the data is too complex to model or correct. This method will fail to correct data sets whenever the assumptions are violated. Whenever sectors and plates are not truly representative or a bias not contemplated by this model is significant then the method will not be fully corrective. We are currently developing diagnostic tools for determining when biases are too extreme to correct.

Examination of the log-ratios reveal that some false positives are no longer significantly different

after range correction. Furthermore, the expected up and down regulated genes are now at the tails of the distribution, where they are expected. The most changed log-ratios are generally from low intensity spots that are most affected by additive error and it is still advisable to view these ratios with caution. A comparison of the effects of each step in the process indicate that the Z correction in unnecessary and potentially could yield inconsistent changes in the range of log-ratios. By contrast however the conceptually intuitive range correction is easily added to other methods and provides the most beneficial results. the plate and sector mean corrections are simple and effective yet are likely to be improved whether replaced by medians or by more sophisticated methods. In summary, this method removed non-linearity, increased the range of ratios, removed plate and spatial bias, and handled outliers proportionally.

ACKNOWLEDGEMENTS

We would like to thank Shauna Somerville, David Ehrhardt, Todd Richmond, Orly Alter, Mike Fero, Fredrik Sterky, Sandrine Dudoit, Gavin Sherlock and Rob Tibshirani for their insights.
Special thanks to the National Science Foundation for their financial support through the postdoctoral fellowship in bioinformatics (David Finkelstein) and through the Plant Genome Research that supports the Arabidopsis Functional Genomics Consortium (grant number 9872638).

REFERENCES

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proc Natl Acad Sci U S A **97**(18): 10101-6.

Andrews, J., G. G. Bouffard, et al. (2000). "Gene discovery using computational and microarray analysis of transcription in the Drosophila melanogaster testis." Genome Res **10**(12): 2030-43.

Armstrong, S. G. and K. W. Renton (1994). "Factors involved in the down-regulation of cytochrome P450 during Listeria monocytogenes infection." Int J Immunopharmacol **16**(9): 747-54.

Bassett, D. E., Jr., M. B. Eisen, et al. (1999). "Gene expression informatics--it's all in your mine." Nat Genet **21**(1 Suppl): 51-5.

Cheng, S. H., B. Moore, et al. (1998). "Effects of short- and long-term elevated CO2 on the expression of ribulose-1,5-bisphosphate carboxylase/oxygenase genes and carbohydrate accumulation in leaves of Arabidopsis thaliana (L.) Heynh." Plant Physiol **116**(2): 715-23.

Eastmond, P. J. and I. A. Graham (2001). "Re-examining the role of the glyoxylate cycle in oilseeds." Trends Plant Sci **6**(2): 72-8.

Finkelstein, D. B., R. Ewing, et al. (2001). "Microarray data quality analysis: lessons from the AFGC." Plant Molecular Biology **in press**.

Finkelstein, D. B., J. Gollub, et al. (2000). Iterative linear regression by sector: re-normalization of cDNA microarray data and cluster analysis weighted by cross homology. Stanford, Carnegie Institution**:** 7.

Kruskal, W. H. and W. A. Wallis (1952). "The use of ranks in one-criterion variance analysis." Journal of the American Statistical Association **47**: 583-621.

Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.

Newton, M. A., C. M. Kendziorski, et al. (2001). "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data." J Comput Biol **8**(1): 37-52.

Ott, L. (1988). An Introduction to Statistical Methods and Data Analysis. Boston, PWS-Kent Publishing Co.

Poynton, C. A. (1993). ""Gamma" and its Disguises: The Nonlinear Mappings of Intensity in Perception, CRTs, Film and Video." The Society of Motion Picture and Television Engineers Journal **102**(12): 1099-1108.

Raines, C. A., P. R. Horsnell, et al. (1992). "Arabidopsis thaliana carbonic anhydrase: cDNA sequence and effect of CO2 on mRNA levels." Plant Mol Biol **20**(6): 1143-8.

Richmond, T. and S. Somerville (2000). "Chasing the dream: plant EST microarrays." Curr Opin Plant Biol **3**(2): 108-16.

Rylott, E. L., M. A. Hooks, et al. (2001). "Co-ordinate regulation of genes involved in storage lipid mobilization in Arabidopsis thaliana." Biochem Soc Trans **29**(2): 283-7.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Sherlock, G., T. Hernandez-Boussard, et al. (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-5.

Yang, Y. H., S. Dudoit, et al. (2000). Normalization for cDNA Microarray Data. UC Berkeley Tech Report.