JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 9, Number 1, 2002
Mary Ann Liebert, Inc.
Pp. 55–66

Improved Background Correction for Spotted DNA Microarrays

CHARLES KOOPERBERG,[1] THOMAS G. FAZZIO,[2] JEFFREY J. DELROW,[3] and TOSHIO TSUKIYAMA[2]

ABSTRACT

Most microarray scanning software for glass spotted arrays provides estimates for the intensity for the "foreground" and "background" of two channels for every spot. The common approach in further analyzing such data is to first subtract the background from the foreground for each channel and to use the ratio of these two results as the estimate of the expression level. The resulting ratios are, after possible averaging over replicates, the usual inputs for further data analysis, such as clustering. If, with this background correction procedure, the foreground intensity was smaller than the background intensity for a channel, that spot (on that array) yields no usable data. In this paper it is argued that this preprocessing leads to estimates of the expression that have a much larger variance than needed when the expression levels are low.

Key words: Bayesian statistics, low intensity spots.

1. INTRODUCTION

GENE EXPRESSION REGULATES THE PRODUCTION OF PROTEIN, which in turn governs many cellular processes in biological systems. The knowledge of gene expression has applications ranging from basic research and trying to better understand the mechanism of protein production to applications such as diagnosing, staging, and finding treatments of diseases. With cDNA microarrays, it is now possible to measure rapidly and efficiently the expression level of genes expressed in a biological sample.

In this paper, we focus on how to process data that arises from glass-spotted arrays. On these arrays, typically several thousand cDNA spots corresponding to different ORFs are applied. Two types of probes labeled from different isolates of messenger RNA are hybridized to this array. One of these types is labeled with Cy3 ("green") dye, the other with Cy5 ("red") dye. After the hybridization has been carried out, the intensity of the green dye and the intensity of the red dye on a particular spot indicate how much

[1]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, MP 1002, Seattle, WA 98109-1024.

[2]Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, MP 1002, Seattle, WA 98109-1024.

[3]DNA Array Facility, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, MP 1002, Seattle, WA 98109-1024.

expressed RNA of the particular type was hybridized to that spot. The relative intensities of the green and the red dyes provide an estimate of the expression ratio for the particular gene of the two tissues from which the RNAs were extracted.

Two problems associated with this procedure are that some of the probe will attach to the array, even when there is no cDNA available. This is known as "background intensity." To some extent, the data can be corrected for background intensity by combining the intensities of the dye at a particular spot with the intensities of the dye at a (nearby) area of the array where there was no cDNA spotted. Another problem is that, not only does a probe that corresponds to the "correct gene" hybridize with the cDNA, but also some probe corresponding to other genes may hybridize. This is known as cross-hybridization. It is much less clear how to correct for this problem.

In this paper, we primarily deal with the problem of background correction. As we will see, this problem is intimately related to the estimation of the expression ratio. The standard approach to background correction is to subtract an estimate of the background intensity from the intensity measured in the spot (the foreground intensity). This approach can cause problems when the foreground intensity is low, for example, of the same magnitude as the background intensity. This situation will cause estimates of the expression ratio to become very noisy, or even undefined, when the background intensity is higher than the foreground intensity. Still, these spots do contain valuable information. In Fig. 1 we show the foreground intensity of the red channel for two repeat arrays. Triangles (crosses) in this plot have a foreground intensity that is below the background for one (two) of the arrays. As can be seen, most of these spots are as reproducible as the other spots on the array, which suggests that the intensities for these spots are real. As such, inclusion of the data of such spots in further analysis, such as clustering (e.g., Eisen *et al.*, 1998; Hastie *et al.*, 2000; Tibshirani *et al.*, 1999) or analysis of variance (Kerr *et al.*, 2000), will improve the results of such analysis. In this paper, we introduce an alternative Bayesian method to correct for background noise. A feature of this method is that it only uses the summary statistics (mean, median, SD) for each spot that are provided by the typical scanning software and does not require analysis of the raw pixel data that is obtained from the scanner (GenePix, 1999).

We are not aware of a method to "correct" spotted glass array data for cross-hybridization. In Section 3.2, we give an example in which we compare expression ratios estimated using glass spotted arrays and

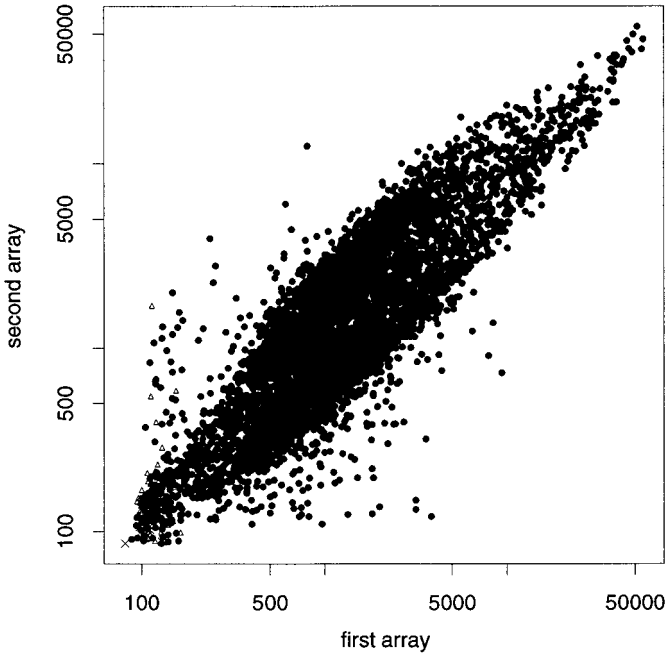

FIG. 1. Median foreground expression for the red channel of two repeat arrays. Spots for which the expression level is below background in one (two) channels are shown as triangles (crosses). There are 82 triangles and 16 crosses in this figure.

estimated using northern blot. Northern blot analysis does not suffer from cross hybridization to the extent that glass spotted arrays do. We here find out that this is the main cause of inconsistency between northern blot and glass array results.

One of the earliest papers discussing algorithms for the computing of expression ratios for glass spotted arrays is by Chen *et al.* (1997). There, the background area is identified as a group of pixels that have a significantly lower (using a Wilcoxon statistics) intensity than the remaining pixels. As in this paper, they assume that the median intensity level of the foreground pixels has a normal distribution. This assumption is being used to identify spots for which the expression ratio (not the log-expression ratio) is significantly different from zero. It is not clear who first proposed the method which we label as "traditional" to correct for background and compute expression ratios, though it may go back to Eisen *et al.* (1998). Newton *et al.* (2000) propose a Bayesian approach to computing the log-ratio, after background correction has already taken place. They assume that the true foreground intensities for the red and green channels follow a Gamma distribution. Theilhaber *et al.* (1999) also proposes a Bayesian algorithm to estimate the fold change. Their paper also assumes that background correction of the intensities has already taken place. They do, however, acknowledge the possibility that the observed background intensity may sometimes be larger than the foreground intensity by putting a prior on the foreground intensity. Theilhaber *et al.* (1999) carry out the computations for the expression ratio on a regular rather than a logarithmic scale (as did Chen *et al.* [1997]). We feel that this is somewhat unfortunate, since the resulting algorithm is not "symmetric" for exchanging the red and green channels.

In the next section, we discuss a Bayesian algorithm for background correction. All parameters in the proposed methodology can be estimated based on the limited number of summary statistics that are provided by most scanning/image analysis software packages. In Section 3, we apply our algorithm to a number of glass spotted arrays of yeast. We end with a brief discussion.

2. A BAYESIAN APPROACH TO BACKGROUND CORRECTION

Most scanning devices and software for the analysis of glass spotted arrays provide the user with estimates of a number of quantities for both channels for every spot on an array. In particular, we will have the following information: the mean of the foreground intensities X_f over all pixels within a region of interest, the standard deviation S_f of these intensities, the median of the intensities Y_f, and the number of pixels within the foreground region n_f. We will also have estimates of the same quantities over an appropriately defined background region. We refer to the background quantities as X_b, S_b, Y_b, and n_b. The actual background region will be different depending on the scanning software used; that is, the background region could be all the pixels near the spot that are not in the foreground region, they could be all the pixels near the spot that are at least a few pixels away from the foreground region, or they could be a selected other region. The calculations that are discussed in this section are carried out for every spot and both channels separately. We omit subscripts g and r for the Cy3 dye ("green") and Cy5 dye ("red"), respectively, and an index i for the spot, whenever this causes no confusion.

It is assumed that the amount of probe that hybridizes to a particular spot has an approximate linear relation with the intensity. Within the background region, the RNA can attach to the (unprocessed) array. Within the foreground region, the RNA can hybridize to the target cDNA or to the glass itself. These two effects are assumed to be additive. Formally, the intensity of every pixel in the foreground is assumed to be a random variable with mean $\mu_f = \mu_t + \mu_b$, $\mu_b, \mu_t \geq 0$ and the intensity of every pixel in the background is assumed to be a random variable with mean μ_b. The goal of this section is to obtain a posterior distribution for the mean of the target μ_t.

From the statistical theory on point processes (e.g., Ripley, 1981) it follows that if the number of RNA molecules that hybridize to a particular area is independent to those that hybridize at another area then the intensity of a single foreground (background) pixel i Z_{if} (Z_{ib}) would be proportional to a Poisson distribution. In particular, αZ_{if} and αZ_{ib} would each have a Poisson distribution with means μ_f/α and μ_b/α, respectively. The proportionality constant α could depend on the channel. We will get back to these assumptions shortly.

For the model described above, the maximum likelihood estimate for μ_t, $\hat{\mu}_t$, is $X_f - X_b$. In practice, either μ_t or the difference between the median levels of foreground and background $\mu_t = Y_f - Y_b$ is used

to estimate the target intensity. Using these estimates can cause problems if $\mu_f \approx \mu_b$ so that $\mu_t \approx 0$, as now it is possible that $\mu_t < 0$ or $\mu_t < 0$. This causes difficulties, as further analysis is often based on an expression ratio μ_{tg}/μ_{tr}, so that spots for which either $\mu_{tg} < 0$ or $\mu_{tr} < 0$ have to be ignored in further analysis. As we will see in Section 3, this approach can increase the variance substantially. To circumvent this, we take a Bayesian approach.

2.1. *Informal description of the Bayesian approach*

Informally the Bayesian approach can be summarized as follows. If the true background intensity is μ_b, we will observe a background intensity X_b that may be smaller or larger, roughly following a normal distribution with mean μ_b. Without any further information, our best estimate for μ_b would be X_b.

Similarly, if the true foreground intensity is $\mu_f = \mu_t + \mu_b$, we will observe a foreground intensity X_f that may be smaller or larger, roughly following a normal distribution with mean μ_f. If we would just look at the foreground, our best estimate for μ_f would be X_f.

However, we know that $\mu_t \geq 0$, and thus $\mu_f \geq \mu_b$. So, if $X_b > X_f$ it has to be true that the X_b that we observed was larger than the true background intensity μ_b or that the X_f that we observed was smaller than the true foreground intensity μ_f, since otherwise $\mu_t < 0$, which we know to be false. Reversing these arguments, we would now guess that $\mu_b < X_b$ or that $\mu_f > X_f$, and thus that $\mu_t = \mu_f - \mu_b > X_f - X_b$. If X_f is only slightly larger than X_b, this does not have to be the case, but it is still likely that $\mu_t > X_f - X_b$.

After we completely describe the relation of X_b to μ_b, the relation of X_f to μ_f and our prior beliefs concerning μ_f and μ_b, we can formalize this argument and use the posterior distribution to get a better estimate of μ_t.

2.2. *Posterior distribution of* μ_t

Since X_f and X_b are means over fairly large numbers of pixels, because of the central limit theorem $X_b \sim N(\mu_b, \sigma_b^2)$ and $X_f \sim N(\mu_t + \mu_b, \sigma_f^2)$ for some σ_b and σ_f. We assume that, conditional on μ_b, μ_t, σ_b, and σ_f, X_b and X_f are independent. Let p_{μ_t} and p_{μ_b} be the prior distributions of the target and background intensities, respectively. We assume that these prior distributions of μ_t and μ_b are independent of each other and do not depend on σ_b and σ_f. (As we will see below, the posterior distributions are not independent though.) Using Bayes' theorem we get

$$
\begin{aligned}
p(\mu_t, \mu_b | \sigma_f, \sigma_b, X_f, X_b) &= \frac{p(X_f, X_b | \mu_t, \mu_b, \sigma_f, \sigma_b) p(\mu_t, \mu_b | \sigma_f, \sigma_b)}{p(X_f, X_b | \sigma_f, \sigma_b)}, \\
&= \frac{p(X_f | \mu_t, \mu_b, \sigma_f) p(X_b | \mu_b, \sigma_b) p(\mu_t) p(\mu_b)}{\int\int p(X_f, X_b | u, v, \sigma_f, \sigma_b) p_{\mu_t,\mu_b}(u, v | \sigma_f, \sigma_b) du dv}, \\
&= \frac{\phi\left(\frac{X_f - \mu_b - \mu_t}{\sigma_f}\right) \phi\left(\frac{X_b - \mu_b}{\sigma_b}\right) p_{\mu_t}(\mu_t) p_{\mu_b}(\mu_b)}{\int\int \phi\left(\frac{X_f - u - v}{\sigma_f}\right) \phi\left(\frac{X_b - v}{\sigma_b}\right) p_{\mu_t}(u) p_{\mu_b}(v) du dv},
\end{aligned}
$$

so that

$$
p(\mu_t | \sigma_f, \sigma_b, X_f, X_b) = \frac{\int \phi\left(\frac{X_f - v - \mu_t}{\sigma_f}\right) \phi\left(\frac{X_b - v}{\sigma_b}\right) p_{\mu_t}(\mu_t) p_{\mu_b}(v) dv}{\int\int \phi\left(\frac{X_f - u - v}{\sigma_f}\right) \phi\left(\frac{X_b - v}{\sigma_b}\right) p_{\mu_t}(u) p_{\mu_b}(v) du dv}, \tag{1}
$$

where $\phi(\cdot)$ is the density of the standard normal distribution.

Let $\sigma_d = \sqrt{\sigma_b^2 + \sigma_f^2}$, the standard deviation of the difference $X_f - X_b$. If we take a noninformative improper uniform prior for μ_t and μ_b on $[0, \infty)$ equation (1) reduces to

$$p(\mu_t|\sigma_b, \sigma_f, X_b, X_f) = \frac{\int_0^\infty \phi\left(\frac{X_f - v - \mu_t}{\sigma_f}\right)\phi\left(\frac{X_b - v}{\sigma_b}\right)dv}{\int_0^\infty\int_0^\infty \phi\left(\frac{X_f - u - v}{\sigma_f}\right)\phi\left(\frac{X_b - v}{\sigma_b}\right)dudv}.$$

A noninformative prior is typically assumed if we do not want to make any strong prior assumptions about the parameters (Box and Tiao, 1973). In the current setup, the noninformative prior can also be seen as the limit of a gamma prior when we let its parameter go to infinity. A simple transformation of variables now yields

$$p(\mu_t|\sigma_b, \sigma_f, X_b, X_f) = \frac{\phi\left(\frac{X_f - \mu_t - X_b}{\sigma_d}\right)\Phi\left(\frac{(X_f - \mu_t)\sigma_b^2 + X_b\sigma_f^2}{\sigma_f\sigma_b\sigma_d}\right)}{\sigma_d\int_0^\infty \Phi\left(\frac{X_f - v}{\sigma_f}\right)\phi\left(\frac{X_b - v}{\sigma_b}\right)dv}, \tag{2}$$

if $\mu_t \geq 0$ and 0 otherwise, where $\Phi(x) = \int_{-\infty}^{x}\phi(x)dx$, the cumulative standard normal distribution.

We now return to the assumptions and their implications. It is fairly common that in practice not all pixels in a spot show hybridization at the same level. The reason for this is that the probe may not be uniformly distributed on the array. (For example, it is not uncommon that because of the mechanics of spotting arrays the center of the spot has less target deposited. This is sometimes known as the doughnut effect. See Fig. 3 of Chen *et al.* [1997].) To circumvent this problem, we propose to use the median intensities in the foreground and background region, Y_f and Y_b, instead of the mean intensities, X_f and X_b, respectively. Since the number of DNA molecules that hybridize at an individual pixel will be large, the Poisson distribution of αZ_{if} (and αZ_{ib}) is well approximated by a normal distribution. This implies that, like X_f and X_b, Y_f and Y_b have approximately normal distributions with means μ_f and μ_b, respectively. Therefore, Equation (2) remains approximately valid, even if we substitute the medians Y_f and Y_b for the means X_f and X_b.

Thus, for each of the two channels, Equation (2), with Y's substituted in for the X's, provides a way to compute the posterior distribution of μ_t, provided we have estimates for σ_f and σ_b. It is tempting to use $\sigma_f = S_f\sqrt{\pi}/\sqrt{2n_f}$ and $\sigma_b = S_b\sqrt{\pi}/\sqrt{2n_b}$. (Note that the standard deviation of the median of n independent identically distributed normal random variables is $\sigma\sqrt{\pi}/\sqrt{2n}$.) However, this would be a valid approach only if the intensity levels of the individual pixels within one foreground (background) region are independent and identically distributed. While, depending on the image segmentation algorithm, the assumption of identical distributions may be questionable, it is also clear that pixel values are not independent.[1] If we assume that the intensities of the pixels do have an identical distribution, and that the correlation between the intensity of two spots depends only on the distance between these two spots, it follows that

$$\sigma_f = a\frac{S_f}{\sqrt{n_f}} \tag{3}$$

and

$$\sigma_b = a\frac{S_b}{\sqrt{n_b}} \tag{4}$$

for some constant a, ignoring boundary effects (Ripley, 1981).

[1] This is, for the background pixels, the assumption that is being made by Theilhaber *et al.* (1999).

An implicit assumption of background correction is that the background intensity is locally constant. If we assume that μ_b is (approximately) constant from one spot to the next spot, the empirical standard deviation of Y_b for that spot and the (four) spots that are physical neighbors of that spot is an alternative estimate σ_b for that spot. We estimate a separately for each channel by regression $\frac{S_b}{\sqrt{n_b}}$ on σ_b for all spots on an array. This estimate $\hat{a}$ is then combined with both the background (4) and foreground (3) standard deviations provided by the software package to obtain estimates σ_f and σ_b.

The Bayesian approach to background correction implicitly assumes that, for pixels for which the background intensity Y_b is larger than the foreground intensity Y_f, this usually occurs when the target mean μ_t is very small, and that we observe a chance event. In particular, we assume that

$$Y_b - Y_f \sim N(\mu_t, \sigma_f^2 + \sigma_b^2).$$

As $\mu_t \geq 0$, the quantity

$$q = \Phi\left(\frac{Y_f - Y_b}{\sqrt{\hat{\sigma}_f^2 + \hat{\sigma}_b^2}}\right)$$

provides some sort of significance level for the assumption that $\mu_t \geq 0$ and thus that the model is in this aspect reasonable. Since we will usually compute q for both channels for all spots simultaneously, we apply a Bonferoni correction to circumvent that we will identify 5% of all spots if the model was reasonable, but only identify any spots in 5% of the experiments (e.g., Bickel and Doksum, 1977). For example, for the yeast arrays discussed in Section 3 there are 6,309 spots. This means that we label spots for which

$$\frac{Y_f - Y_b}{\sqrt{\hat{\sigma}_f^2 + \hat{\sigma}_b^2}} < \Phi^{-1}\left(\frac{.05}{2 \times 6309}\right) \approx -4.5 \tag{5}$$

as suspicious.

2.3. *Combination of both channels*

Using the methodology described in Section 2, $E(\mu_{tg}|\hat{\sigma}_{fg}, \hat{\sigma}_{bg}, Y_{fg}, Y_{bg})$ and $E(\mu_{tr}|\hat{\sigma}_{fr}, \hat{\sigma}_{br}, Y_{fr}, Y_{br})$ are obtained from (2) using numerical integration (which is straightforward assuming that $\phi(\cdot)$ and $\Phi(\cdot)$ are readily available) as estimates of the posterior distribution of the mean of the target μ_{tg} and μ_{tr} for the green and red channels, respectively. A natural estimate for the log-ratio $\log(\mu_{tr}/\mu_{tg})$ is

$$\text{LR} = E(\log(\mu_{tr}|\hat{\sigma}_{fr}, \hat{\sigma}_{br}, Y_{fr}, Y_{br}) - E(\log(\mu_{tg}|\hat{\sigma}_{fg}, \hat{\sigma}_{bg}, Y_{fg}, Y_{bg})).$$

Kerr *et al.* (2000) discuss analysis of variance models for combining (normalized and background corrected) estimates of the log expression for different channels in situations more complicated than the comparison of two channels on the same array.

While it is unlikely that the posterior distributions of $\log(\mu_{tr})$ and $\log(\mu_{tg})$ are independent, it is tempting to take $V = \text{var}(\log(\mu_{\text{tr}}|\sigma_{\text{fr}}, \sigma_{\text{br}}, \text{Y}_{\text{fr}}, \text{Y}_{\text{br}}) + \text{var}(\log(\mu_{\text{tg}}|\sigma_{\text{fg}}, \sigma_{\text{bg}}, \text{Y}_{\text{fg}}, \text{Y}_{\text{bg}})$ as a measure of the relative accuracy of the log-ratio for different spots on the same array and for spots of the same gene on different arrays. In our experience though, the correlation between the two channels is high (correlations between median background levels of the green and the red channels upwards of 0.95 are common), so that fairly small changes in the local correlation structure from from one spot to another spot would dominate the variation. As such, we would recommend more empirical methods to determine the variance of the log-expression ratios. For example, after appropriate preprocessing, we could use a local polynomial estimator of the variance as a function of the mean expression ratio, for example using data like that displayed in Fig. 2. An alternative approach is to estimate the variance of the log-expression ratio using repeat data, after which analysis could be made using standard t-statistics. As for the data used for the examples in the next section, finding genes with "significant" expression ratios was not a goal of the experiment; we do not pursue that approach further.

3. APPLICATION

3.1. Wildtype–wildtype

We applied the methodology described in the previous section to fifty glass spotted yeast arrays produced by the Fred Hutchinson Cancer Research Facility array facilities. Ten of these arrays are so called "wildtype-wildtype" experiments, where both channels show expression of genes in the same wildtype yeast cells, so that we know that the true expression ratio is 1. In Fig. 2, we show for one of these arrays both the traditional estimate of the log-expression ratio

$$\log\left(\frac{Y_{fr} - Y_{br}}{Y_{fg} - Y_{bg}}\right) - \text{(correction factor)} \tag{6}$$

and an estimate based on applying the method described in the previous section

$$\text{LR} - \text{(correction factor)'}. \tag{7}$$

The two correction factors in Equations (6) and (7) are chosen such that the median expression ratio for eight yeast control genes on the array is one for the particular method. One reason to correct the estimate is that the proportionality constant, α, may be different for both channels. For the traditional method (6), we exclude spots for which either $Y_{br} \geq Y_{fr}$ or $Y_{bg} \geq Y_{fr}$ and spots that were labeled as "bad" by the investigator. For the new method (7), we exclude spots that are suspicious according to (5) and spots that were labeled as "bad" by the investigator. The effect of this is that many spots that are excluded for the traditional approach are not excluded for the new approach. These spots are shown as triangles in Fig. 2. There were 14 spots excluded for the new proposed method; all of these spots were also excluded for the traditional approach. Figure 3 shows for the red channel both the estimate of the intensity for the traditional approach ($Y_f - Y_b$) and the new proposed estimate $\hat{E}(\mu_t | \sigma_f, \sigma_b, Y_f, Y_b)$.

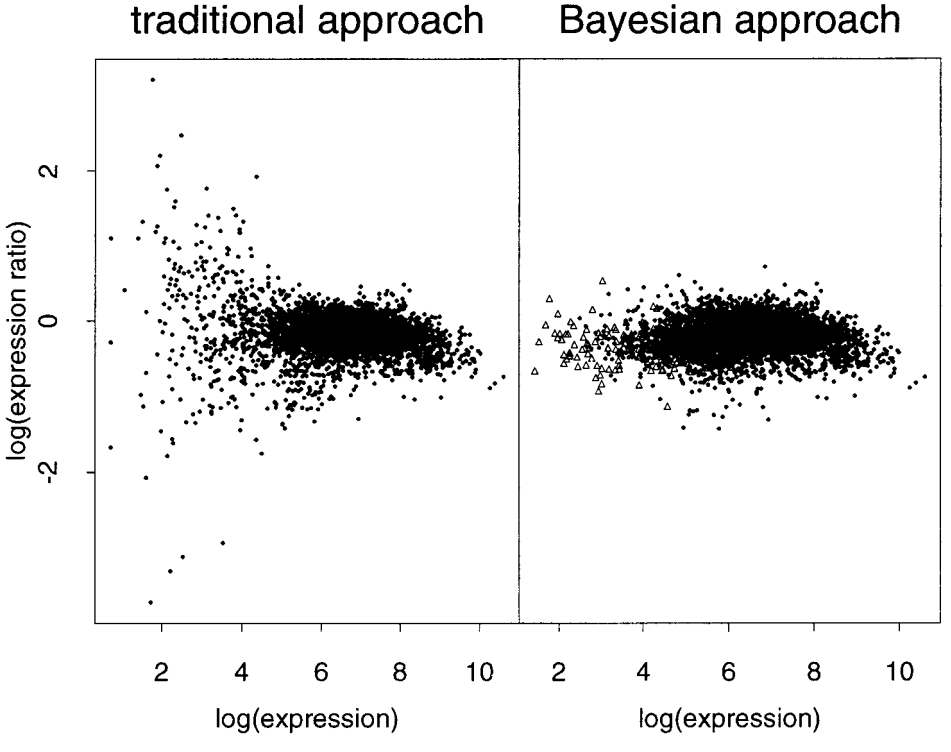

FIG. 2. Two estimates of the log-expression-ratio against the mean of the log-expression for both channels. The 255 genes for which the traditional method did not yield an estimate of the ratio are triangles for the Bayesian method.

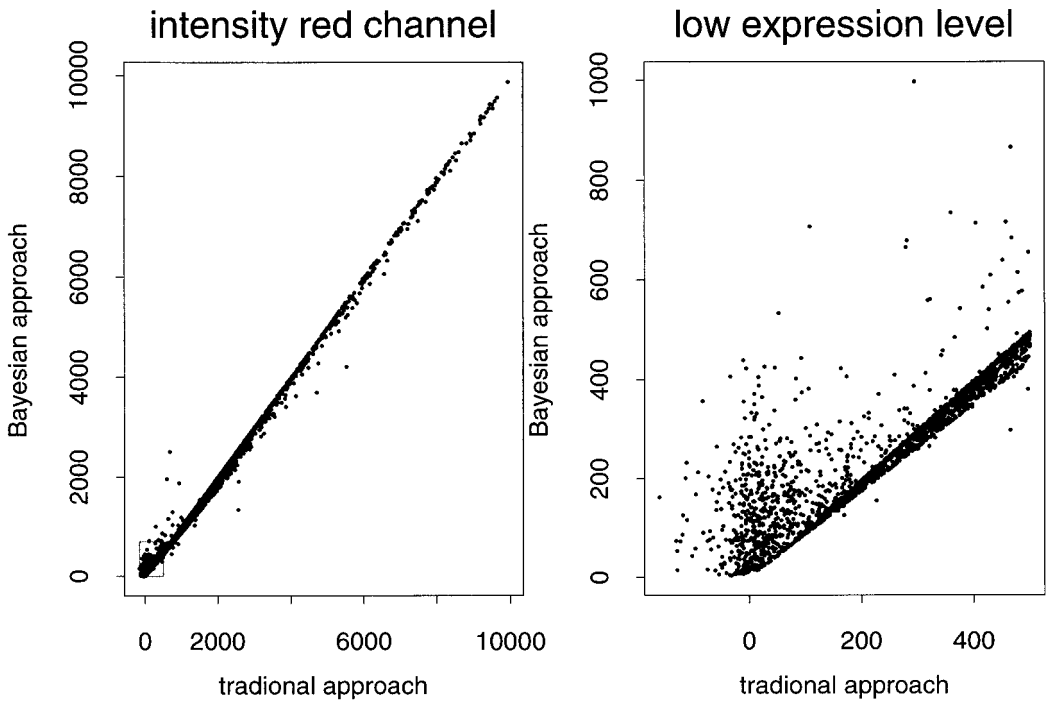

FIG. 3. Traditional and Bayesian estimate of the expression level for the red channel. The right side enlarges the area in the box that contains genes that have low expression levels.

The zoom-in on the right side shows that the proposed methodology essentially modifies the estimates for which $Y_f - Y_b$ is small, while if $Y_f - Y_b$ is large the proposed estimate essentially equals the traditional estimate. This is partly the effect of using a noninformative uniform prior: if $Y_f - Y_b$ is large, the prior does not modify the traditional estimate, but if $Y_f - Y_b$ is small, the prior, which requires $\mu_t > 0$, will yield a larger estimate than the traditional approach. The results shown in Figs. 2 and 3 are typical for all ten wildtype–wildtype arrays and both channels. In Table 1, we summarize for the ten wildtype–wildtype arrays the mean squared error between the estimated log-expression ratio and the true log-expression ratio of 1. The SD is the standard deviation of the mean squared error over the ten arrays. As can be seen from

TABLE 1. AVERAGE MEAN SQUARED ERROR PER ARRAY IN THE LOG-EXPRESSION RATIO[a]

	Traditional approach		*Bayesian approach*	
Spots averaged over	*Mean*	*SD*	*Mean*	*SD*
All spots not labeled suspicious by (7)	—	—	0.042	0.021
All spots for which the traditional method yields an expression ratio	0.051	0.029	0.038	0.020
Spots for which the traditional method does not yield an expression ratio	—	—	0.187	0.122
Spots for which the expression level is among the 10% lowest levels on the chip	0.233	0.149	0.096	0.061
Spots for which the expression level is not among the 10% lowest levels on the chip	0.031	0.017	0.032	0.017
Fraction of spots for which an estimate was obtained	0.975	0.008	0.997	0.001

[a]The SD shows the array-to-array variation.

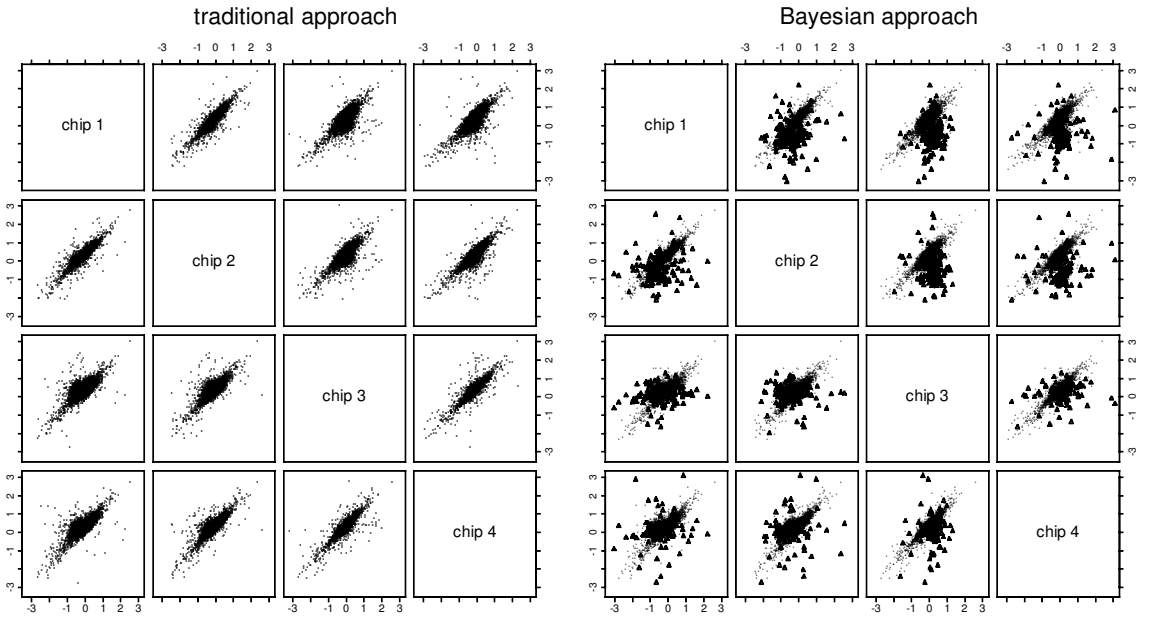

FIG. 4. Scatterplot matrix for the traditional and Bayesian estimates of the log-expression ratio for four replicate chips of the same mutant. Combinations for which the traditional method did not yield an estimate of the ratio are triangles for the Bayesian method; there are between 110 and 160 such points in each panel, most of which fall on top of the other points.

this table, the traditional and the Bayesian procedures perform virtually identically when the expression levels are high, but when the expression level is low the Bayesian procedure has a considerably reduced variance.

3.2. Mutants

We also analyzed data on forty arrays where one channel contained RNA from wildtype yeast cells and the other channel from yeast cells in which some genes were mutated.There were ten different type of mutants, each of which was used on four arrays, twice as the green channel and twice as the red channel ("reverse fluoring"). In contrast to the wildtype–wildtype experiments, for these arrays we do not know what the true expression ratio is.

In Fig. 4 we show a pairs plot of the traditional (left) and Bayesian estimate (right) of the expression ratio for the four replicates of one of the mutants. The average correlation coefficient of the six correlation coefficients is 0.59 for the traditional approach and 0.82 for the Bayesian approach (0.77 if we exclude the points for which the traditional approach does not yield an estimate). Combining all ten sets of replicates, the average correlation for the traditional approach is 0.58 and the average correlation for the Bayesian approach is 0.77 (0.72 excluding the points for which the traditional approach does not yield an estimate).

The expression ratios of the mutants for some of the genes were independently verified using Northern Blot. In particular, we have Northern Blot data for 126 gene/mutant combinations. In Fig. 5 we show a plot of the traditional (left) and the Bayesian (right) estimates of the expression ratio for each of the four replicates (i.e., there are $4 \times 94 = 376$ points in these plots.[2] In Fig. 5, we show the (unweighted) average of the expression ratios against the Northern Blot data. Visually, there seems to be little difference between these approaches, although the root mean square difference between the cDNA estimate of the expression ratio and the Northern Blot estimate of the expression ratio is somewhat lower for the Bayesian estimate (0.694) than the traditional estimate (0.769).

[2]Two expression ratio are missing for the traditional approach.

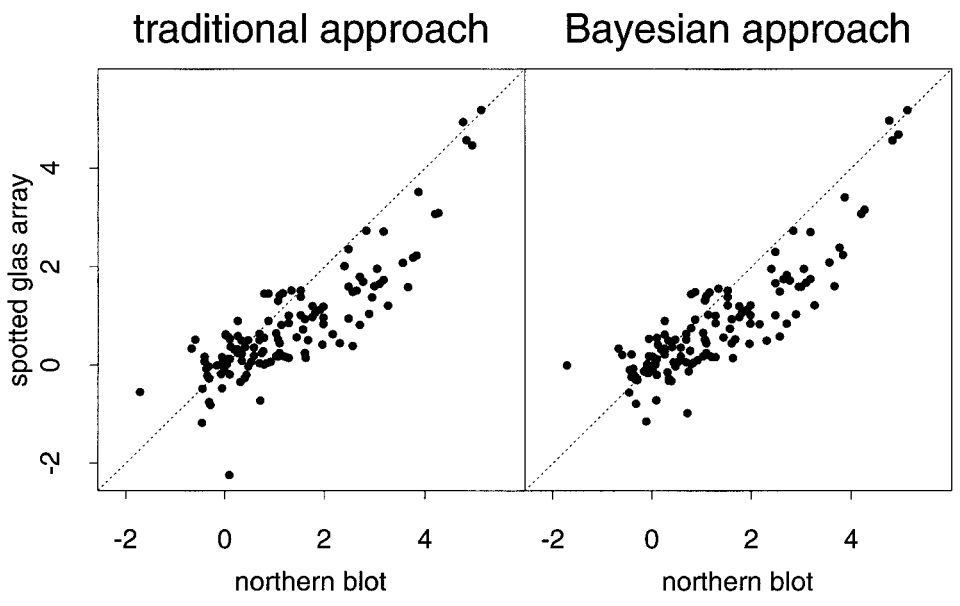

FIG. 5. Traditional and Bayesian estimate of the log-expression ratio based on the average of four arrays against a Northern Blot estimate of the expression ratio.

We do notice from Fig. 5 that there seems to be a systematic difference between the array measurements (independent of the computational approach) and the Northern Blot measurements. In particular, Northern Blot yields larger log-expression ratios for those gene/mutant combinations that already yield high expression ratios. We eventually determined that the reason for this apparent bias was that the standard Northern Blot analysis only measures the expression level for RNA of the correct length and excludes cross-hybridization, while for glass spotted arrays the expression level includes cross-hybridization. Effectively this "inflates" both the numerator and the denominator for the glass spotted array technique,

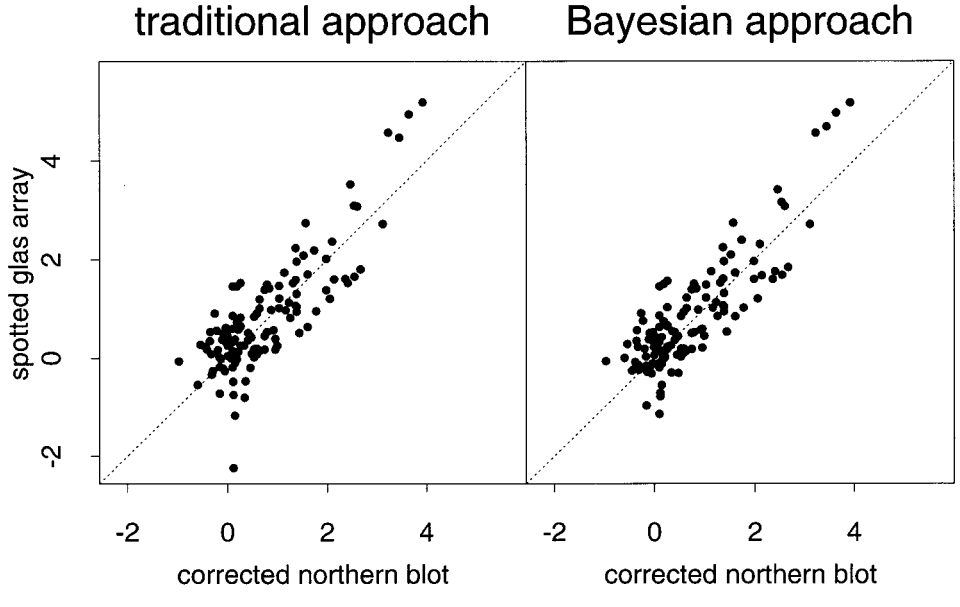

FIG. 6. Traditional and Bayesian estimate of the log-expression ratio based on the average of four arrays against a Northern Blot estimate of the expression ratio that has been adjusted to include cross-hybridization.

Table 2. Average Mean Squared Difference between the cDNA Estimate and the Northern Blot Estimate of the Log-Expression Ratio[a]

Genes	*Number of gene/ mutant combinations*	*Traditional approach*	*Bayesian approach*
20 selected for biological reasons	94	0.393	0.394
4 with low expression levels	30	0.463	0.183
All	124	0.411	0.343

[a]The two spots for which the traditional approach did not yield an estimate have been eliminated.

yielding somewhat smaller log-expression ratios. It is, unfortunately, not possible to exclude this cross-hybridization. On the other hand, it is possible to estimate the Northern Blot expression level including cross-hybridization (something we usually would not want to do). To this extent, we reread the Northern Blot films that were used in Fig. 5 to include cross-hybridization. Figure 6 contains these corrected results; as can be seen, now there appears to be little bias and little difference between the two approaches. Numerically, the root mean squared difference between the Northern Blot estimates and the traditional estimates is 0.411, while the squared difference between the Northern Blot estimates and the Bayesian estimates is 0.343.

In total there were 24 genes for which Northern Blot was done for some of the mutants. Of these 24 genes, 20 were selected because they were of interest for biological reasons (their analysis will be reported elsewhere), the remaining four were selected because they had low expression levels, and it was suspected that we may see differences between the traditional and the Bayesian analysis method. In Table 2, we summarize the differences between the two analysis methods separately for the two groups of genes. As can be seen, similarly to the wildtype experiments, the gain of the Bayesian approach comes from the experiments with low expression levels.

4. DISCUSSION

In this paper, we propose a Bayesian method for background correction and the computation of log-expression ratios. The proposed approach reduces the variation of the estimates of the expression ratio when the expression levels are low. At the same time, it keeps the estimates for expression ratios virtually unchanged when the expression levels are higher. After this preprocessing, the standard error of estimates of the log-expression ratio appears to be approximately independent of the expression level (see Fig. 2). This is desirable, since it allows users to informally compare estimates of expression ratios, without having to worry too much about the standard errors of these estimates.

Undoubtedly, spots with low expression levels have many problems, for example image segmentation algorithms are likely doing a poorer job on spots where the expression level is low, and such spots are more likely to be the result of faulty spotting on the array. Still, reliable estimates of the expression level will improve further analysis, such as clustering, as good clustering algorithms will not depend on the precise expression level of a single spot, but rather depend mainly on the magnitude of (groups) of log-expression ratios.

ACKNOWLEDGMENTS

Charles Kooperberg was supported in part by NIH grant CA 74841. Thomas G. Fazzio was supported in part by a predoctoral fellowship from H.H.M.I. Toshio Tsukiyama was supported in part by the Pew Charitable Trust Biomedical Scholars Fellowhip and NIH grant GM58465.

The authors wish to thank Michael LeBlanc and Steve Self for a number of helpful discussions and Cassandra Neal and Ryan Basom for help in carrying out the cDNA experiments.

REFERENCES

Bickel, P.J., and Doksum, K.A. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.

Box, G.E.P., and Tiao, G.C. 1973. *Bayesian Inference in Statistical Analysis*, Wiley, New York.

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

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.

GenePix User's Guide. 1999. Axon Instruments, Foster City, CA.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., and Botstein, D. 2000. *Gene Shaving: A New Class of Clustering Methods for Expression Arrays.* Technical report, Department of Statistics, Stanford University.

Kerr, M.K., Martin M., and Churchill, G.A. 2000. Analysis of variance for microarray data. *J. Comp. Biol.* To appear.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. 2000. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol*. 8, 37–52.

Ripley, B.D. 1981. *Spatial Statistics*, Wiley, New York.

Theilhaber, J., Bushnell, S., and Fuchs, R. 1999. Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *Nature Genet.* 23, 78.

Tibshirani, R., Hastie T., Eisen, M., Ross, D., Botstein, D., and Brown, P. 1999. *Clustering Methods for the Analysis of DNA Microarray Data.* Technical report, Department of Statistics, Stanford University.

Address correspondence to:
Charles Kooperberg
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue N, MP 1002
Seattle, WA 98109-1024

E-mail: clk@fhcr.org