Background Adjustment for DNA Microarrays Using a Database of Microarray Experiments

YUNXIA SUI,¹ XIAOYUE ZHAO,³ TERENCE P. SPEED,⁴ and ZHIJIN WU^{1,2}

ABSTRACT

DNA microarrays have become an indispensable technique in biomedical research. The raw measurements from microarrays undergo a number of preprocessing steps before the data are converted to the genomic level for further analysis. Background adjustment is an important step in preprocessing. Estimating background noise has been challenging because background levels vary a lot from probe to probe, yet there are limited observations on each probe. Most current methods have used the empirical Bayes approach to borrow information across probes on the same array. These approaches shrink the background estimate for either the entire sample or probes sharing similar sequence structures. In this article, we present a solution that is truly probe specific by using a database of large number of microarray experiments. Information is borrowed across samples and background noise is estimated for each probe individually. The ability to obtain probe specific background distributions allows us to extend the dynamic range of gene expression levels. We illustrate the improvement in detecting gene expression variation on two datasets: a Latin Square spike-in experiment from Affymetrix and an Estrogen Receptor experiment with biological replicates. An R package *dbRMA* implementing our method can be obtained from the authors.

Key words: background adjustment, gene expression, high-density oligonucleotide microarrays, preprocessing.

1. INTRODUCTION

HIGH-DENSITY OLIGONUCLEOTIDE MICROARRAYS have been used in biomedical research for over a decade now. The application of these microarrays has expanded from monitoring gene expression levels to quantifying genomic variations such as DNA copy number changes, comparative genomic hybridization (CGH) (Barrett et al., 2004; Huang et al., 2004), and DNA methylation (Schumacher et al., 2006; Shi et al., 2002). Regardless of the biological application, the microarray technology quantifies the abundance of nucleic acid molecules by employing the hybridization property between nucleic acid molecules with complementary sequences. The complex biological sample, composed of a large variety of DNA or RNA fragments, is labeled with fluorescent dyes and sorted by the probes immobilized on a solid surface. Ideally,

¹Department of Community Health and ²Center for Statistical Sciences, Brown University, Providence, Rhode Island.

³Bionovo Inc., Emeryville, California.

⁴Department of Statistics, University of California, Berkeley, California.

only complementary sequences bind together (hybridize), allowing the mixture of DNA/RNA fragments to be separated by binding to their specific probes. The measurement on each probe is a fluorescent intensity representing the amount of nucleic acid attached to that probe. The hybridization between complementary sequences is referred to as *specific binding*. However, in all microarray applications, there is a fair amount of *nonspecific binding* going on at the same time as specific binding happens. Thus, the observed intensity on

FIG. 1. Log raw probe intensities of a yeast experiment versus different human tissues. The four different human tissue experiments are found from the GEO website (GSM180626, GSM224762, GSM263915, and GSM263919).

each probe is often a combination of specific and nonspecific binding, as well as optical noise. Intensity on probes that is not due to specific binding is often referred to as background noise. Failure to adjust for background noise leads to bias in estimating relative variations in the amount of DNA or RNA in the biological samples.

Various approaches have been taken to estimate and adjust for background noise in oligonucleotide arrays. The original design of Affymetrix arrays, one of the most popular oligo arrays, included mismatch probes that differ from the perfect match arrays by only one nucleotide in the middle of the sequence. Intensities on the mismatched probes were treated as direct measurements of the background noise for the perfect match probes. This approach was used in Affymetrix algorithm for gene expression measure MAS 5.0 (Hubbell et al., 2002). This background adjustment turned out to introduce a lot of variability in the log transformed gene expression measures (Irizarry et al., 2003; Wu et al., 2004). Irizarry et al. (2003) estimate a global background for the entire array by borrowing strength across all probes in the same hybridization. This background estimate is used in expression measure Robust Multiarray Analysis (RMA). Huber et al. (2003) also estimate a global background for all probes on an array in the Variance Stabilization Normalization (VSN) algorithm. Noticing that nonspecific binding can vary a great deal across different probes, Wu et al. (2004) use a simple model to predict the affinity of probes towards nonspecific binding based on the sequence composition of the probes. Instead of considering all probes within one array exchangeable in background noise, this model allows information borrowed from probes sharing similar nonspecific binding affinities.

All of the above approaches to estimate background noise are based on borrowing information across probes within the same arrays. This has been a natural choice since the microarray is a famous example of the “large p small n ” problem. There are tens of thousands of probes on each microarray but usually only a small number of arrays used in each experiment. Thus, the number of observations on each probe is very limited in an experiment. This situation has changed after commercial platforms of microarrays have been used for over a decade now. Although the number of arrays used in most experiments are often still limited, the biomedical research community has been sharing the raw data through public data repositories. Now for a given short oligonucleotide, we have accumulated hundreds even thousands of observations across various biological conditions. This presents us the opportunity of studying each probe’s characteristic over a large number of arrays.

In Wu and Irizarry (2004), an experiment was introduced in which yeast genomic DNA was hybridized to a human gene expression array. A BLAST search confirmed no complementary match between the probes and the yeast genomic sequence; thus, the intensities observed on this array is all due to nonspecific binding. Figure 1 shows that intensities on this array appear extremely similar to those observed in experiments involving entirely different human tissues. This suggests that nonspecific binding accounts for a large portion of the observed intensity and that nonspecific binding appears to be probe specific and rather consistent across arrays, even when the target sample varies. In this article, we take advantage of the consistency of nonspecific binding across different samples to estimate the distribution of background noise on each probe from a large collection of microarrays.

2. DESCRIPTION OF DATA

2.1. Database

An accumulation of microarray data from public data depository consists of 491 oligonucleotide arrays from various experiments. Each experiment was conducted on the same type of microarray chip, which is the Human Genome U-133 A platform from Affymetrix. The Hgu133a chip includes 22300 probe sets and each probe set comprises 11–20 probe pairs (PM/MM), and thus there are 506944 probes on each array. Since we only use the PM probes, the data in use is a 248152×491 matrix, each cell of the matrix is the observed intensity of the hybridization for that PM probe. This huge database provides 491 repeated measurements for each probe, and it motivates us to estimate the probe specific background.

2.2 Latin Square hgu133a spike-in experiment

This data set consists of 3 technical replicates of 14 separate hybridizations of 42 spiked transcripts in a complex human background at concentrations ranging from 0 to 512 pM (picoMolar). This data set

and a detailed description of the Latin Square design are available at www.affymetrix.com/support/technical/sample_data/datasets.affx. Each spike-in gene is composed of 11 or 20 probes, and in total there are 498 probes with known concentrations. As the concentrations are known for them, this data set is utilized for bias assessment of our proposed method.

2.3. Estrogen Receptor (ER) experiment

Tetracycline-inducible U2OS cells stably transfected with *ER α* or *ER β* were treated for 6 hours with 10 nM E2, 125 μ g/ml MF101, or 1 μ M raloxifene (RAL), tamoxifen (TAM), genistein (GENT), ERB-041, niasol (NYA), liquiritigenin (LIQ), glycosylated form LIQ (LGL), diarylpropionitrile (DPN), or propyl pyrazole triol (PPT). The final labeled cRNA samples were hybridized overnight against Hgu133a2 GeneChips. All treatments were done in triplicate. The ability to detect differentially expressed genes and consistency among control genes and replicates is evaluated in this experiment.

3. PROBE SPECIFICITY AND EFFICIENCY

In microarray experiments, the concentration of a target molecule (DNA or RNA) is measured indirectly by the fluorescent intensity read from its complementary probes after hybridization take place. The intensity as a result of specific binding between complementary sequences is believed to be proportional to the target concentration, $E[S] = aC$, where S is the specific binding intensity, and a is a probe specific parameter representing the efficiency of a probe in response to the change of its target concentration C . Empirical evidence suggests constant coefficient of variation in specific binding, and most researchers have modeled the measurement error as multiplicative $S = aCe^\epsilon = \exp\{\log(a) + \log(C) + \epsilon\}$ (Rocke and Durbin, 2001; Durbin et al., 2002; Huber et al., 2003; Wu and Irizarry, 2007). Since the efficiency parameter a is unknown, the quantification of target concentration is only relative: the (log) ratio of specific binding intensity represents the (log) ratio of concentration: $\log(S_1) - \log(S_2) = \log(C_1) - \log(C_2) + \epsilon^*$. The simple cancellation of the probe efficiency a is only possible when the probe is absolutely specific and thus only binds to its target sequence. The observed intensity for most probes, however, is a combination of the background and specific binding: $Y = B + S$. As a result, $E[\log(Y_1) - \log(Y_2)] \neq E[\log(C_1) - \log(C_2)]$. The apparent log concentration ratio calculated from observed intensities, $\log(Y_1) - \log(Y_2)$, is biased towards 0 if any of these conditions apply:

1. if the probe has low specificity (thus there is considerable nonspecific binding, resulting in large B),
2. if the probe has low efficiency a ,
3. if the target concentration C is low.

The larger the background B is, relative to specific binding intensity S , the stronger the bias. For highly expressed genes, the target concentrations C can be high enough such that the impact of additive background is negligible. In Figure 2, we show the log ratio of probe intensities from two of the Latin Square spike-in arrays. All but the spike-in genes have equal concentration, and the spike-in genes have a concentration ratio of 2 over the two arrays. The probes for the spike-in genes are highlighted in red (if the lower concentration of the two target arrays is less than 4 pM) or green (if the lower concentration of the target is at least 4 pM). For the high concentration spike-in targets (green points), most of the probes have positive log ratios that are clearly separated from the probes for the unchanging genes even when no background adjustment is done. But the log ratios of the probes for the other spike-in genes (red points) are biased towards 0 and buried in the noise. However, the histogram of the average log intensities indicates that vast majority of the probes do not reach the high-intensity range where background can be ignored. In the Latin-square experiments, the concentration of target genes range from 0 to 512 pM and are evenly distributed in the log scale. In other words, the spike-in genes are selectively biased towards high intensities ranges. As a result, in a bias-variance tradeoff, one can sacrifice the bias to avoid false positive in the low-intensity range and still be able to detect considerable amount of differential expression in the high-intensity range. However, in practical situations, we are less lucky to have as many genes with such high expression levels. Therefore, adjusting for probe specific background is an essential step in preprocessing.

FIG. 2. Log ratios of raw intensities of two Latin Square spike-in arrays and the histogram of the average log intensities. Green points are spiked-in genes whose lower concentration of the two target arrays is at least 4 picoMolar. Red points are spiked-in genes whose lower concentration is less than 4 picoMolar.

4. PREVIOUS WORK IN ESTIMATING BACKGROUND

Several approaches have been taken to estimate background intensity on the probes. (Notice that by background we refer to the intensity observed on probes due to nonspecific binding and optical noise, in contrast to the image background on cDNA arrays, which is often read from regions surrounding the probe area.) Mismatch probes, which differ from the Perfect Match probes by one base in the center of the probe, were used to directly measure probe specific background intensity. These were shown to be biased and lead to large variance in log transformed expression levels (Irizarry et al., 2003).

Other methods have taken advantage of the large number of probes on an array and borrowed strength over all probes to estimate an array-specific background common to all probes. RMA (Irizarry et al., 2003) assumes a normal distribution for the background and uses the lower tail of the empirical distribution of probe intensities to estimate the mean and variance of the background. The VSN algorithm in Huber et al. (2002) also estimates a common mean background (baseline) value for all probes on an array using an iterative procedure: a robust variant of the maximum likelihood method under the assumption that the

transformed probe intensity follows a normal distribution. One drawback of using global background parameters is the fact that probes with different sequences have different chemical properties and thus probe specific background levels. Lacking an accurate model to directly predict background intensity from probe sequence alone, GCRMA (Wu et al., 2004) makes use of probe sequences to calculate an affinity measure for each probe's tendency of nonspecific binding. Instead of assuming one global background distribution for all probes, GCRMA borrows information over probes sharing similar affinity levels and obtains background parameters that are probe specific. By borrowing information over probes with similar sequence compositions, physical models that use the stacking energy of DNA/RNA hybridizations have also been employed to predict the hybridization stability of specific and nonspecific bindings. Zhang et al. (2003) proposed the Position Dependent Nearest Neighbor method (PDNN) that use a stacking energy model and compute both gene-specific and non-specific binding energies of a probe as weighted sums of nearest neighbor stacking energies given the sequence of that probe. However, low dimensional models based on probe sequence composition of a probe cannot explain all the variation of non-specific binding.

5. ESTIMATING PROBE-SPECIFIC BACKGROUND FROM A DATABASE OF MICROARRAYS

One major reason that previous approaches of estimating probe specific background borrow information across different probes is that the number of observations on the same probe is very limited, since most microarray experiments have only a few replicates. With data from nearly 500 arrays on the same platform, we are now able to study the distribution of intensities on each probe. Since the arrays in the database are collected over a variety of experiments from different labs, removing the array effect with normalization is the first key step since systematic variations such as hybridization condition and scanner settings are known to affect readings on the entire array. Specifically, we use quantile normalization to normalize all microarrays in the database to the distribution of the Latin-square array.

Imagine one probe whose target gene is not present in any of the samples hybridized to the arrays in the database. For this probe, we would have repeated measures of the background intensity. In reality, the target genes have various concentrations in different samples, and the distribution of probe intensity across samples is a mixture

$$f(x) = \sum_{i=0}^p \pi_i f_i(x) = \pi_0 f_0(x) + \sum_{i=1}^p \pi_i f_i(x), \quad (1)$$

where the target gene's expression is assumed to take p possible levels and at each expression level the intensity distribution is $f_i(x)$. Here f_0 represents the background component, when the target gene is absent in the sample. Since most genes are not expressed in all tissues at all times, when we sample across a large variety of experiments, $f_0(x)$ is expected to be one of the major components of the mixture distribution $f(x)$. The empirical distribution of probe intensities is consistent with the mixture model. In Figure 3, we present empirical density and quantile-quantile plot of several randomly chosen probes. The majority of the probes, like the ones shown here, appear to have a major component with the smallest mean in the mixture. The quantile-quantile plots with normal distribution as the theoretical distribution start with a portion of straight line, suggesting a normal-like distribution for the first component of the mixture.

Estimating the distribution of background intensity now simplifies to estimating the first component of the mixture distribution. We could use the EM algorithm to estimate the number of components as well as the parameters for each component by assuming a normal mixture model. However, we want to avoid running an EM for more than 200,000 probes. Fortunately, we are only interested in one component, namely f_0 , for the purpose of estimating background. The portion of the straight line in the quantile-quantile plot seems a natural start: if the distribution is indeed one component normal distribution, the intercept and slope of the line represent the mean and standard deviation. However, in the mixture distribution, the straight line portion is shifted upwards by the other stochastically larger components. In order to estimate the parameters of f_0 , we consider the observed intensities on a probe as a distribution of background

FIG. 3. Empirical distribution of probe intensities from the database. Density of log intensity and quantile-quantile plot of intensity for a random selection of 8 probes. A normal distribution is used as the theoretical distribution on the x-axes of the quantile-quantile plots.

FIG. 4. Observed versus estimated intensity at concentration zero. The straight line is the 0-1 identical line.

intensities contaminated by large outliers. The left tail of the empirical distribution is least affected by the outliers. We consider the first M order statistics of the N probe intensities, whose likelihood is

$$L = [1 - F_x(y_M)]^{N-M} f_x(y_1) \dots f_x(y_M).$$

We now approximate this likelihood by ignoring the contribution from other components for the smallest M order statistics:

$$L \approx [1 - \pi_0 \Phi(y_M; \mu, \sigma)]^{N-M} \phi(y_1; \mu, \sigma) \dots \phi(y_M; \mu, \sigma) \pi_0^M,$$

where $\Phi(y_M; \mu, \sigma)$ is the cumulative density function of $N(\mu, \sigma)$, and $\phi(y_M; \mu, \sigma)$ is the density function.

The mean and variance of each probe are estimated by maximizing this approximated likelihood. In the Latin Square dataset, a number of genes are known to be absent (concentration zero) in some samples. The observed intensities on the probes for these genes are due to background noise alone. In Figure 4, the observed intensity versus estimated intensity for the zero concentration probes are plotted. The background estimates from the database are much closer to the observed values than predicted values based on sequence alone. In Figure 5, we also show that the estimates of both mean and variance obtained from the approximate likelihood are very close to that obtained from EM for these probes.

6. ESTIMATING THE SPECIFIC BINDING INTENSITY AND EXPRESSION LEVEL

With the background parameters for each probe available, we are ready to estimate specific binding intensities. Each new array is normalized at the probe level to the same reference distribution to that of the arrays used to estimate background are normalized. Now with the additive background model $Y = B + S$, where background B is modeled as $N(\mu, \sigma^2)$, we can estimate the specific binding intensity S given the total observed intensity Y . One option is to assume a prior distribution for S and calculate a posterior expectation. GCRMA uses a flat prior on $\log(S)$. However, there is no analytical solution of the posterior mean of $\log(S)$, and it has to be computed via numerical integration. To lower the computation burden, a simple but fast approximation is provided in GCRMA that uses a weighted linear average of a small constant k and $(Y - E[B])^+$ with weights determined by minimizing the mean squared error of the weighted average. This

FIG. 5. (A,B) Background parameters from approximate likelihood are similar to those obtained from EM algorithm. Estimated log background mean by our approximate likelihood method (order.stat.) against EM algorithm.

essentially shrinks intensities lower than estimated background to a constant k and adjusts very high intensities by subtracting the estimated mean background. The constant k is a tuning parameter. Larger values of k minimizes variance in the lower tail of expression, but loses the ability to detect subtle changes. Here, we use the same approximation to estimate $\log(S)$. Normalization and summarization remain the same as RMA and GCRMA, and we name the new method *dbRMA* to reflect the use of database in background estimation and the other preprocessing steps as RMA.

7. ESTIMATING THE STANDARD ERROR OF EXPRESSION MEASURES

The log transformation taken in expression measures stabilizes variance for many genes. This is apparent in the specific binding model since $S = aCe^\epsilon$ implies $\log(S) = \log(a) + \log(C) + \epsilon$. If there were no background, the log transformation would remove the dependence of mean and variance. However, since specific binding S is not directly observed, a log transformation of background adjusted intensity does not have the desired constant variance. As specific binding increases, the effect of background adjustment becomes smaller, the variance of $\log(\hat{S})$ decreases and converges to the variance of ϵ . When specific binding is very weak, the observed intensity is mostly due to background, and the variance of $\log(\hat{S})$ becomes large if unbiasedness is required, as seen in MAS 5.0 expression measures. Therefore, it is often beneficial to shrink the estimate of $\log(S)$ on probes with very low intensities. As a result, the variance of $\log(\hat{S})$ decreases as the observed intensity decreases to background level. Such a relationship between variance and mean of log expression is observed on most expression measures in the log scale (Irizarry et al., 2006), as illustrated in Figure 6.

Knowing that the variances of gene expression values are not constant and depend on the gene expression level, we find it useful to provide an estimate of the standard error of expression measure that reflects this relationship. In the summarization step, we have the following model for each gene g ,

$$s_{gij} = c_{gi} + a_{gj} + \epsilon_{gij}$$

where s_{gij} is the log specific binding intensity for probe j in sample i , c_{gi} is the gene expression level in sample i , and a_j represents the probe efficiency ($\sum_j a_{gi} = 0$). If we knew the variance for ϵ_{gij} , the standard

FIG. 6. The variance of gene expression measures depends on the expression level.

FIG. 7. The bias assessment at probe and gene level. The straight line has slope 1. The tuning parameter k is set at 6 for dbrma6 and at 10 for dbrma10. **(A)** Estimated specific binding $\log_2(S)$ by three methods. **(B)** Estimated gene expression aligned at 8 for all three methods when the log nominal concentration is 4.

error for least squares estimate of c_{gi} would be easy to calculate. To achieve robustness, we fit the model using median polish instead and obtain the residuals $r_{gij} = \hat{e}_{gij}$. We sort the residuals r_{gij} by gene expression level c_{gi} and use a running Median Absolution Deviation to estimate a σ_{gi} that depends on expression measure. We further smooth the σ_{gi} as a function of c_{gi} by loess and use the fitted values as $\hat{\sigma}_{gi}$. Now for the gene expression in each sample, we calculate the standard errors for least square estimates and use these to approximate the standard errors $\tilde{\tau}$ for our estimates c_{gi} .

8. RESULTS

8.1. Assessment using the Latin-square spike-in data

To evaluate how the estimated background can improve the estimation of the specific binding S , we test it on the Latin Square hgu133a experiment data.

Since the observed probe intensities include both specific and nonspecific binding, $\log_2(B + S)$ is approximately $\log_2(S)$ when concentration is high and converges to $\log_2(B)$ when S decreases to 0. This leads to bias in $\Delta \log_2(S)$ if B is not accounted for. An ideal background adjustment would remove the bias and give $\log_2(\hat{S})$ a linear relationship with $\log_2(C)$ with slope 1 regardless of the concentration. Figure 7A compares the estimated specific binding from RMA, GCRMA and dbRMA versus the nominal concentrations. The estimated log specific binding from all three methods demonstrate a linear relationship with log nominal concentration for high concentrations. However, for lower concentrations, the slope from RMA decreases for log concentrations lower than 2. GCRMA does a more aggressive background correction but does not differentiate the log concentrations lower than -1 . This is probably because affinity to nonspecific binding is estimated by borrowing information among probes with similar sequences in GCRMA. So for the probes whose mean nonspecific binding is lower than what is predicted by the affinity model, background is overestimated and small variations in specific binding are not detected.

At the probe level, we see that dbRMA can extend the lower bound of concentration for detectable variation, thus increasing the dynamic range of detectable expression variation. We also evaluate the dynamic range expansion at the gene expression level. Figure 7B shows similar improvement by dbRMA at the low concentrations.

Reducing bias alone is not necessarily hard to achieve. The key question is whether we can better detect the variation of target expression at low concentrations above noise. Since the challenge is mostly in detecting moderately expressed genes, which are the majority as shown in Figure 2, we focus on the ability of detecting differential expression for sensitivity and specificity of genes expressed lower than 2 pMolar. The receiver-operating-characteristic (ROC) curve for such genes is shown in Figure 8. For each gene at concentrations c_1 and c_2 , the differential expression is defined as $E_1 - E_2$, where E is the estimated log expression. Since we also provide standard error for dbRMA, a statistic for ranking the differential expression is also calculated as $\frac{(E_1 - E_2)}{\sqrt{(\tilde{\tau}_1^2 + \tilde{\tau}_2^2)}}$ (we ignore the covariance between the estimates for simplicity).

Figure 8 shows that using the dbRMA method increases the ability of detecting differential expression at low concentrations. Accounting for the relationship between a gene expression estimate and its variance improves the result even more.

8.2. Assessment using biological replicates

The Latin-square experiment provides a unique dataset for which we know the true positives and negatives, making the assessment relatively easy. However, it does not include biological replicates, and the number of differentially expressed genes is limited. In order to evaluate how the new method works in practice, we assess the ability of detecting differentially expressed genes on the Estrogen Receptor (ER) experiment in which $ER\alpha$ or $ER\beta$ transfected U2OS cells were treated with 11 different ER modulators, and each treatment has 3 biological replicates.

After the preprocessing steps by RMA, GCRMA, or dbRMA, the Limma (Smyth, 2004) procedure is conducted to find the differentially expressed genes. The criterion of being differentially expressed is that genes have fold change greater than 2, and the BH (Benjamini and Hochberg, 1995) adjusted p -values are less than 0.05. Different preprocessing methods result in detecting different number of differentially expressed genes. Table 1 lists the number of detected genes for each ligand in each cell line.

FIG. 8. The ROC curves for genes expressed lower than or equal to 2pM.

Table 2 compares the sensitivity of detecting positive genes for treatments E2, RAL, and TAM. The positive genes are compiled from the intersect of two sets: one set from a previous study (Tee et al., 2004) and the other set from the intersection of all three preprocessing methods (RMA, GCRMA, and dbRMA) with BH-adjusted p -values < 0.05 . In cell line α , there are 120, 22, and 53 positive genes for treatment E2, RAL, and TAM, respectively. In cell line β , the number of positive genes are 58, 35, and 52, respectively. The sensitivity is listed at four cutoffs of the BH-adjusted p -values. Table 2 shows that the new method is sensitive in detecting positive genes, usually more sensitive than RMA and slightly less sensitive than GCRMA.

Another important property for a good expression measure is to have small variance for genes expected to be non-differentially expressed. We consider two comparisons here. First, we compute the median absolute deviation of expression measures for 100 control genes provided by Affymetrix (to download, see www.affymetrix.com/support/technical/byproduct.affx?product=hgu133-20) from the three preprocess-

TABLE 1. NUMBER OF DIFFERENTIALLY IDENTIFIED GENES IN ER α AND ER β TRANSFECTED U2OS CELLS, TREATED WITH 11 DIFFERENT ER MODULATORS

Treatment	Cell line α			Cell line β		
	RMA	GCRMA	dbRMA	RMA	GCRMA	dbRMA
E2	607	1165	725	299	712	436
RAL	42	155	68	246	623	332
TAM	206	504	265	185	516	260
MF101	11	78	16	298	681	385
NYA	117	280	152	313	711	429
LIQ	2	23	3	332	746	430
LGL	0	34	2	292	701	413
DPN	174	501	238	326	705	432
ERB41	0	33	2	259	653	379
GENT	596	1132	691	276	685	405
PPT	791	1421	893	—	—	—

Threshold: fold change > 2 and BH adjusted p -value < 0.05 .

TABLE 2. SENSITIVITY OF DETECTING DIFFERENTIAL EXPRESSION AT DIFFERENT CUTOFF OF THE BH ADJUSTED p -VALUES

<i>Treatment</i>		<i>p-Value cutoff</i>							
		<i>Cell line α</i>				<i>Cell line β</i>			
		<i>0.05</i>	<i>0.001</i>	<i>1.00E-04</i>	<i>1.00E-05</i>	<i>0.05</i>	<i>0.001</i>	<i>1.00E-04</i>	<i>1.00E-05</i>
E2	RMA	0.692	0.692	0.692	0.683	0.569	0.569	0.569	0.569
	GCRMA	0.833	0.825	0.783	0.775	0.776	0.776	0.776	0.741
	dbRMA	0.783	0.783	0.775	0.767	0.672	0.672	0.672	0.655
RAL	RMA	0.318	0.318	0.318	0.318	0.686	0.686	0.686	0.686
	GCRMA	0.636	0.545	0.455	0.364	0.857	0.857	0.857	0.771
	dbRMA	0.409	0.409	0.409	0.364	0.743	0.743	0.743	0.714
TAM	RMA	0.396	0.396	0.377	0.377	0.423	0.423	0.404	0.404
	GCRMA	0.623	0.623	0.509	0.491	0.615	0.577	0.558	0.538
	dbRMA	0.472	0.453	0.453	0.434	0.519	0.519	0.519	0.500

TABLE 3. VARIABILITY IN HOUSEKEEPING GENES AND BIOLOGICAL REPLICATES

	<i>RMA</i>	<i>GCRMA</i>	<i>dbRMA</i>
MAD among 100 control genes	0.152	0.182	0.164
Average pairwise L2-distance among biological replicates	0.046	0.069	0.036

sing methods. A small MAD indicates good precision of gene expression measure. Second, we compare the similarity of biological replicates. Average pair-wise L2 distance is computed for each method. Table 3 shows that dbRMA has MAD between RMA and GCRMA for the 100 housekeeping genes, and gives better similarity among biological replicates.

9. DISCUSSION

This article presents a new background correction method for DNA microarrays. Unlike the shrinkage methods that either assume exchangeability of all probes or use sequence models, the background parameters in dbRMA are estimated from observations on each probe separately and are truly probe specific. The probe specific background is estimated by using a database of microarrays collected from the public data depository. We evaluate this method using both technical replicate arrays with spiked-in controls as well as data with biological replicates. The comparison is done with two widely used preprocessing procedures, RMA and GCRMA. The assessment using the ER experiment data suggests that GCRMA tends to be more sensitive (Table 2) and identifies many more positive genes (Table 1). However, this is achieved by a price of increased noise, as shown in Table 3 for the analysis on control genes that are expected to be non-differentially expressed. This is probably because, by shrinking background estimates according to probe sequence, background adjustment is done aggressively for many probes. By contrast, RMA estimates a small global background, and has less noise, but lower sensitivity. dbRMA estimates background parameters that are truly probe specific, since it does not borrow information across probes. Its sensitivity is lower than GCRMA, but also more specific: data from biological replicates appear more similar in dbRMA than in both RMA and GCRMA. We prefer dbRMA for the balance demonstrated in real biological experiments. This is also supported by the ROC curve in Figure 8.

The database of microarrays used in this article includes about 500 arrays over a large number of different experiments. We assume that the smallest component in the mixture distribution represents the background. This may not be accurate for the entire transcriptome. For example, the housekeeping genes are considered to be expressed in most, if not all, tissues and at all times. For these genes, we may not have observed their background distribution component. The estimated background parameters may actually represent a “minimum expression” state, rather than no expression. Thus, dbRMA may remove some specific binding signal for these genes. Since microarrays are only used to measure relative, not absolute,

expression levels, we are mostly concerned about the impact on relative expression measures. This possible overestimation of background can lead to a positive bias in differential expression magnitude, as insufficient background adjustment leads to underestimating differential expression. Fortunately, the number of housekeeping genes is small, and we expect that, for the vast majority of the genes, the background component is observed. Table 3 also suggests that dbRMA does not inflate the variance much for the 100 housekeeping genes, compared to RMA. As the collection of data in public repositories continues to grow, each probe will have more and more repeated measurements over a greater variety of tissues and conditions. This will further decrease the probability that a gene is expressed in all arrays included in the database.

The dbRMA method is demonstrated on Affymetrix gene expression arrays, but the idea is platform independent and can be applied to other types of chips. We have observed a mixture distribution similar to that seen in Figure 3 in Affymetrix Exon arrays from a database of 615 arrays downloaded from Gene Expression Omnibus (GEO) (result not shown).

In principle, a large collection of repeated measurement on each probe across many conditions also makes it possible to estimate the other components in model (1), which can be used to estimate specific binding efficiency parameter a . However, for genes that are not expressed in a considerable number of arrays, there is not much information from the components other than f_0 . As microarray data continue to accumulate, we may be able to study probe properties beyond nonspecific binding in the near future.

ACKNOWLEDGMENTS

We want to thank Dr. Dale Leitman from University of California, San Francisco, for providing the Estrogen Receptor experiment data.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Barrett, M.T., Scheffer, A., Ben-Dor, A., et al. 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci. USA* 101, 17765–17770.
- Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300.
- Durbin, B.P., Hardin, J.S., Hawkins, D.M., et al. 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18, S105–S110.
- Huang, J., Wei, W., Zhang, J., et al. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Human Genomics* 1, 287–299.
- Hubbell, E., Liu, W.-M., and Mei, R. 2002. Robust estimators for expression analysis. *Bioinformatics* 18, 1585–1592.
- Huber, W., von Heydebreck, A., Sultmann, H., et al. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 1, 1–9.
- Huber, W., von Heydebreck, A., Sultmann, H., et al. 2003. Parameter estimation for the calibration and variance stabilization of microarray data. *Statist. Appl. Genet. Mol. Biol.* 2, 3.
- Irizarry, R.A.B., Hobbs, F.C., Beazer-Barclay, Y., et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.
- Irizarry, R.A., Bolstad, B.M., Collin, F., et al. 2003. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.* 31, e15.
- Irizarry, R.A., Wu, Z., and Jaffee, H.A. 2006. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 22, 789–794.
- Rocke, D.M., and Durbin, B. 2001. A model for measurement error for gene expression arrays. *J. Comput. Biol.* 8, 557–569.
- Schumacher, A., Kapranov, P., Kaminsky, Z., et al. 2006. Microarray-based dna methylation profiling: technology and applications. *Nucleic Acids Res.* 34, 528–542.

- Shi, H., Maier, S., Nimmrich, I., et al. 2002. Oligonucleotide-based microarray for dna methylation analysis: principles and applications. *J. Cell. Biochem.* 88, 138–143.
- Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.* 3, 3.
- Tee, M.K., Rogatsky, I., Tzagarakis-Foster, C., et al. 2004. Estradiol and selective estrogen receptor modulators differentially regulate target genes with estrogen receptors α and β . *Mol. Biol. Cell* 15, 1262–1272.
- Wu, Z., and Irizarry, R. 2004. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *Proc. RECOMB 2004*.
- Wu, Z., Irizarry, R., Gentleman, R., et al. 2004. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Statist. Assoc.* 99, 909–917.
- Wu, Z., and Irizarry, R.A. 2007. A statistical framework for the analysis of microarray probe-level data. *Ann. Appl. Statist.* 1, 333–357.
- Zhang, L., Miles, M.F., and Aldape, K.D. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* 21, 818–821.

Address correspondence to:

Dr. Zhijin Wu
Center for Statistical Sciences and Department of Community Health
Brown University
Brown University Box G-S121, 7th floor
121 South Main Street
Providence, RI 02912

E-mail: zwu@stat.brown.edu