Statistical methods of background correction for Illumina BeadArray data

Yang Xie, Xinlei Wang, and Michael Story

Abstract

Motivation: Advances in technology have made different microarray platforms available. Among the many, Illumina BeadArrays are relatively new and have captured significant market share. With BeadArray technology, high data quality is generated from low sample input at reduced cost. However, the analysis methods for Illumina BeadArrays are far behind those for Affymetrix oligonucleotide arrays, and so need to be improved.

Results: In this article, we consider the problem of background correction for BeadArray data. One distinct feature of BeadArrays is that for each array, the noise is controlled by over 1000 bead types conjugated with non-specific oligonucleotide sequences. We extend the robust multi-array analysis (RMA) background correction model to incorporate the information from negative control beads, and consider three commonly used approaches for parameter estimation, namely, non-parametric, maximum likelihood estimation (MLE) and Bayesian estimation. The proposed approaches, as well as the existing background correction methods, are compared through simulation studies and a data example. We find that the maximum likelihood and Bayes methods seem to be the most promising.

Contact: yang.xie@utsouthwestern.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Illumina BeadArrays are long oligonucleotide arrays. Compared with Affymetrix oligonucleotide arrays, Illumina expression BeadArrays often generate data of similar quality, require less sample input and reduce the cost of array experiments (Shi et al., 2006). These features have allowed the BeadArray platform to capture a significant market share. BeadArray technology has been applied to a variety of genomic studies including expression profiling, single-nucleotide polymorphism genotyping and copy number variation analysis. In this article, we will focus on genome-wide expression data.

The essential element of BeadArray technology is the 3-micron silica bead coated with hundreds of thousands of copies of a specific oligonucleotide sequence, which will be referred to as one bead type. For genome-wide expression arrays, the majority of the genes are represented by one bead type, which is a specific 50 mer oligonucleotide sequence, derived from the national center for biotechnology information reference sequences. Each bead type is replicated about 30 times on each array to increase the reproducibility and the quality of the data. Besides gene sequences, Illumina also allocates more than 1000 control bead types to each array, which do not correspond to any expressed sequences in the genome. So the associated beads are not expected to hybridize to any genes in the RNA samples. They serve as negative controls for the non-specific binding or background noise in an experiment.

Microarray data preprocessing plays a crucial role in obtaining valid biological results from array experiments (Bolstad et al., 2003; Durbin et al., 2002; Irizarry et al., 2003; Li and Wong, 2001; Lin et al., 2008). Data preprocessing methods for cDNA spotted arrays and Affymetrix oligonucleotide arrays have undergone rigorous testing and validation, and these methods have been shown to greatly improve the quality of data. In contrast, few statistical methods have been developed for processing BeadArray data. Dunning et al. (2008) discussed some important statistical issues in processing and analyzing Illumina BeadArray data; they focused on comparing the existing tools and statistical methods. Lin et al. (2008) proposed a variance-stabilizing transformation (VST) method that takes advantage of the technical replicates available on an Illumina microarray to preprocess BeadArray data. In this article, we will focus on the background correction step for Illumina BeadArrays, the purpose of which is to remove the non-specific signal from total signal. Background correction is the main step separating various microarray data preprocessing methodologies (Irizarry et al., 2006). Furthermore, it is platform specific because of the different sources of noise for each platform (Irizarry et al., 2003; Ritchie et al., 2007). Although the methodologies in background correction for cDNA and Affymetrix arrays are well studied (Irizarry et al., 2006; Ritchie et al., 2007), few have been proposed for BeadArrays. Several inherent platform differences make statistical methods for processing Illumina BeadArrays different from other platforms. First, more than 1000 negative control bead types are allocated on each Illumina array to control background noises. Second, the negative control beads on
Illumina arrays are conjugated to non-specific sequences that are not associated with gene-specific probes; this is different from Affymetrix mismatch probes, each of which has one single nucleotide different from the corresponding perfect match (PM) probe. Third, the number of replicated beads for each bead type is random, 30 on average, and the replicated beads have the same sequence for one bead type.

As the most commonly used software for preprocessing BeadArray data, BeadStudio (Illumina, CA, USA) provides only two options for background correction: no background correction and background subtraction using the average value of the negative control beads. However, a large number of negative data values can be generated by using the background subtraction approach. For example, in our motivating leukemia study (Ding et al., 2008), half of the probe values on one array became negative after the background subtraction step. Without using a more advanced data transformation method, such as VST by Lin et al. (2008), the probes with negative values need to be excluded in the purpose of log transformation and it will result in the loss of large amounts of information on the arrays. On the other hand, without background correction, the commonly used measure for downstream analysis, fold change (i.e. the expression ratios between two experimental groups), can be compressed. For illustration, let \( E \) and \( C \) represent the true expression value for the treatment group and the control group, respectively; and let \( B \) be the noise value for both groups. Suppose we are interested in the ratio \( R = E/C \). Without background correction, we would use the observed total intensities, \( E+B \) and \( C+B \), to calculate the ratio \( R' = (E+B)/(C+B) \). It is easy to show that \( R' \) is always biased to 1 compared with \( R \), which results in fewer genes identified as differentially expressed than expected. Dunning et al. (2008) also concluded that the background normalization recommended by Illumina can introduce substantial variability into the data, increase the numbers of false positives and a large number of low expression values become negative and cannot be log2 transformed. Our research was motivated by the need for alternatives to BeadStudio background correction methodologies.

Besides BeadStudio, two R packages deal with data preprocessing for Illumina BeadArrays: beadarray (Dunning et al., 2007) and lumi (Du et al., 2008). In those packages, there are three choices for background correction: no background correction, background subtraction and Normexp background correction. Normexp uses the same normal-exponential convolution model as the robust multi-array analysis (RMA) background correction, and we refer to this as the RMA model in our article. RMA is a popular R package used to preprocess Affymetrix oligonucleotide array data. The background correction step in RMA uses a normal-exponential convolution model to fit the PM probes on the array. Empirical evidence shows that RMA background correction works well in practice (Bolstad et al., 2003; Irizarry et al., 2003; Wu et al., 2004). Recently, Silver et al. (2008) proposed using saddlepoint approximation to estimate parameters for the normal-exponential convolution model, and they showed that the approach achieved good performance in background correction for two-color microarray data. This normal-exponential convolution model with saddlepoint approximation method is now available in the beadarray package. Ding et al. (2008) proposed a model-based background correction method (MBCB) for Illumina BeadArrays, which extends the RMA model to incorporate information from negative control data generated with Illumina BeadArrays. Ding et al. (2008) showed that MBCB can lead to more precise determination of gene expression and better biological interpretation of Illumina BeadArray data. In this article, we will further improve parameter estimation of the MBCB model. We will consider and compare three methods for parameter estimation including non-parametric, maximum likelihood and Bayesian approaches. A complete mathematical development and detailed parameter estimation are given in this article. The R package and data files used for this article will be available through the Bioconductor web site. We will use both simulation studies and a real data example to illustrate and compare these methods, and provide guidance for their practical use.

2 THE MODEL

The statistical model we use is motivated by a leukemia study, which will be described in Section 5. Figure 1A shows the smoothed histograms of the gene intensities observed from all BeadArrays in the study. It is not surprising that the distributions appear to be similar to the Affymetrix signal distributions (Bolstad et al., 2003). Figure 1B shows one example of the smoothed histograms of intensities for both the genes and negative controls on a single array, from which we can see that the distribution of the negative controls might be approximated by a normal distribution. In addition, the modes of the genes and the negative controls are close to each other, with the mode of the genes being larger than that of the controls. These observations motivated us to extend the convolution model in the RMA background correction method for Affymetrix arrays to the BeadArray signal intensities. That is, observed intensity=true signal+background noise, where the true signal intensities (if not zero) are modeled by an exponential distribution with mean \( \alpha \) and the background noise is modeled by a normal distribution with finite mean \( \mu \) and variance \( \sigma^2 \). We further assume that \( \mu \) and \( \sigma^2 \) take appropriate values (e.g. \( \mu > 2\sigma \)) so that negative background noise occurs rarely and has a negligible impact. Note that an additive model is used here, which was considered for modeling the background noise component by several researchers (Cui et al., 2003; Durbin et al., 2002; Huber et al., 2002; Irizarry et al., 2003; Wu et al., 2004).
2004), and was regarded as a better model than a multiplicative model in the context of background correction.

Throughout the article, we use $i$ to index the genes, and $j$ to index the negative controls. For $i=1, \ldots, I$, we have

\begin{equation}
X_i = \frac{S_i + B_i}{\sigma}, \quad S_i \sim \text{Exponential}(\alpha),
\end{equation}

where $X_i$ is the observed intensity, $S_i$ is the signal intensity and $B_i$ is the noise intensity for gene $i$; and $I$ is the number of the genes. For $j=1, \ldots, J$,

\begin{equation}
X_{0j} = \frac{B_{0j}}{\sigma}, \quad B_{0j} \sim N(\mu, \sigma^2),
\end{equation}

where $X_{0j}$ is the observed intensity, and $B_{0j}$ is the noise intensity for negative control $j$; and $J$ is the number of the negative controls. Note that all $X_i$'s and $X_{0j}$'s are assumed to be independent; and for each negative control bead, the signal intensity $S_{0j}$ is naturally assumed to be zero.

Under (1), for the $i$-th gene, the marginal density of $X_i$ is given by

\begin{equation}
p(x_i) = \int_0^{+\infty} p(x_i | s_i) p(s_i) ds_i = \int_0^{+\infty} \frac{1}{\sigma} \phi \left( \frac{x_i - s_i - \mu}{\sigma} \right) \frac{1}{\alpha} \exp \left( -\frac{s_i}{\alpha} \right) ds_i = \frac{1}{\alpha} \exp \left( \frac{\sigma^2}{2\alpha^2} - \frac{x_i - \mu}{\alpha} \right) \cdot \Phi \left( \frac{x_i - \mu - \frac{\sigma^2}{\alpha}}{\sigma} \right),
\end{equation}

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative density function (cdf) of a standard normal distribution. The conditional distribution of the true signal $S_i$ given the total intensity $X_i=x_i$ can be given by

\begin{equation}
p(x_i) = \int_0^{+\infty} p(x_i | s_i) p(s_i) ds_i = \int_0^{+\infty} \frac{1}{\sigma} \phi \left( \frac{x_i - s_i - \mu}{\sigma} \right) \frac{1}{\alpha} \exp \left( -\frac{s_i}{\alpha} \right) ds_i = \frac{1}{\alpha} \exp \left( \frac{\sigma^2}{2\alpha^2} - \frac{x_i - \mu}{\alpha} \right) \cdot \Phi \left( \frac{x_i - \mu - \frac{\sigma^2}{\alpha}}{\sigma} \right),
\end{equation}

where $a_{i} = x_i - \mu - \frac{\sigma^2}{\alpha}$ and $b = \sigma$. Then the expected true signal given the observed intensity can be given by

\begin{equation}
E(S_i | X_i = x_i) = \int_0^{+\infty} s_i p(s_i | x_i) ds_i = a_i + b \frac{\phi (a_i)}{\Phi (a_i)}.
\end{equation}

As in the RMA method, $E(S_i | X_i = x_i)$ will be used as the background corrected intensity for gene $i$.

It is worth mentioning that (5) is different from that used in the original RMA background correction method (Bolstad, 2004), namely
This is because in the original RMA method, $B_i$'s were assumed to be non-negative; however, for algebraic simplicity, the normalizing constant of the truncated normal distribution was ignored when calculating $p(x_i)$ and the subsequent $E(S_i|X_i=x_i)$ in (6). In contrast, the $B_i$'s in our model can take negative values but with a very small probability so that we can ignore the effect of negative values. Both approaches lead to closed-form formulas. However, we prefer using (3) and (5) because the mathematical correctness of $p(x_i)$ plays an important role in obtaining the correct maximum likelihood estimator.

3 PARAMETER ESTIMATION

In the original RMA convolution model, there is no closed form to estimate the parameters $\alpha$, $\mu$ and $\sigma^2$. The numerical maximum likelihood algorithms cannot converge well, as pointed out by the authors themselves and others (Bolstad et al., 2003; McGee and Chen, 2006). Therefore, an ad hoc approach is used to estimate the parameters. In (Bolstad et al., 2003), it is described as follows. First, a non-parametric density function is fitted from the observed intensities, and the mode of this density is used to estimate $\mu$. Then the lower tail of the density (to the left of the mode) is used to estimate $\sigma$ and the right tail is used to estimate $\alpha$. Based on the program in Bioconductor, the parameters are estimated in a slightly different way: one first obtains the overall mode $m_0$ of the whole density; then obtains the local mode $m_1$ from the left tail of the density (to the left of $m_0$) and use $m_1$ to estimate $\mu$; and finally, one uses the left tail of the density (to the left of $m_1$) to estimate $\sigma$ and the right tail (to the right of $m_1$) to estimate $\alpha$.

The model described in Section 2 incorporates extra information from negative controls, which can easily solve the problem encountered in the RMA parameter estimation. Let $\theta^T$ denote the vector of parameters $\alpha$, $\mu$ and $\sigma^2$ in our model. We will discuss three methods for estimating $\theta$ below.

3.1 A non-parametric estimator

A simple method to estimate $\theta$ is to estimate $\alpha$ using all the observations, but estimate $\mu$ and $\sigma^2$ using observations from negative control beads only. Noting that $EX_0=\alpha+\mu$ and $EX_0^2=\mu$, an unbiased estimator of $\alpha$ can be given by $\tilde{\alpha}$, where $\tilde{\alpha}=\frac{\tilde{X}}{n^2}$. If we further estimate $\mu$ and $\sigma^2$ using sample mean and variance of $(X_0)_{i=1}^J$, then an unbiased estimator of $(\alpha, \mu)$ is given by

$$
\tilde{\alpha} = \frac{\tilde{X} - \tilde{X}_0}{J}, \\
\tilde{\mu} = \tilde{X}_0, \\
\tilde{\sigma}^2 = \frac{\sum_{i=1}^J (X_0 - \tilde{X}_0)^2}{J-1},
$$

with variances

$$
\text{Var}(\tilde{\alpha}) = \frac{\alpha^2 + \sigma^2}{J}, \\
\text{Var}(\tilde{\mu}) = \frac{\sigma^2}{J}, \\
\text{Var}(\tilde{\sigma}^2) = \frac{2\sigma^4}{J(J-1)}.
$$

An approximate interval estimator of $\alpha$ can be given by $\tilde{\alpha}$, with $\tilde{\alpha} \pm 1.96 \frac{\sigma}{\sqrt{n}}$. Also, an exact confidence interval of $\alpha$ can be computed using the fact that $\tilde{\alpha}$ is equivalent to the sum of independent Gamma($I, \beta/I$) and Normal($0, \sigma^2(1/\beta + 1/J$)).

Though simple in nature, it might be attractive for practitioners. First of all, it is a non-parametric estimator of $\theta$ that requires no distributional assumptions to both true signals and background noise. Second, it is indeed the least square estimator of $(\alpha, \mu)$, which minimizes the sum of squared distances

$$
D = \sum_{i=1}^{I} (X_i - \mu - \alpha)^2 + \sum_{j=1}^{J} (X_0j - \mu)^2.
$$

More importantly, it is very easy to compute; and the associated variances can be estimated readily by plugging $\tilde{\alpha}$ in (8). Note that we use the
3.2 The maximum likelihood estimation (MLE)

When estimating \( \mu \) and \( \sigma^2 \), the non-parametric estimator ignores the information expressed through the genes and this leads to great simplicity in calculation. However, since \( I \), one may improve estimation efficiency of \( \mu \) and \( \sigma^2 \) by incorporating such information. In order to do so, it is natural to consider the maximum likelihood estimator, due to its well-established theoretical properties.

The likelihood function is given by

\[
L(\theta) = \prod_{i=1}^{l} p(x_i | \alpha, \mu, \sigma^2) \cdot \prod_{j=1}^{J} p(x_{0j} | \mu, \sigma^2)
\]

\[
= \prod_{i=1}^{l} \frac{1}{\alpha} \exp \left( \frac{\sigma^2}{2\alpha^2} - \frac{x_i - \mu}{\alpha} \right) \Phi \left( \frac{x_i - \mu - \sigma^2}{\sigma} \right).
\]

\[
\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi \sigma}} \exp \left[ -\frac{(x_{0j} - \mu)^2}{2\sigma^2} \right].
\]

The maximization of the log likelihood function \( l(\theta) \) over \( \theta \) has no closed form solution. The MLE can be computed numerically through the Newton–Raphson algorithm, which iteratively updates the parameter estimate using

\[
\theta^{k} = \theta^{k-1} - \left[ \partial^2 l(\theta^{k-1}) \right]^{-1} \partial l(\theta^{k-1}), \quad k = 1, 2, \ldots,
\]

where \( \partial \) denotes the vector of the first-order derivatives of \( \theta \), denotes the Hessian matrix of the second-order derivatives of \( l(\theta) \), and \( \theta^0 \) is the initial value. Here, we use our non-parametric estimate for \( \theta^0 \). We note that the good initial point, plus the fact that \( \partial \) and \( \partial^2 \) are both available in closed forms, leads to quick convergence of the algorithm in our problem setting. Also, the covariance matrix of \( \theta \) is estimated routinely through \( \partial^2 \), so that we can construct an interval estimator of \( \theta \) readily. We also show that for large samples, \( \partial \) can improve estimation efficiency of \( \theta \). The details can be seen in the Supplementary Material.

3.3 A Bayesian approach

Bayesian approaches are frequently used in microarray data analysis due to the flexibility of incorporation of complicated data structure and prior information (Reilly et al., 2003; Xiao et al., 2006). Here, we describe a Bayesian method for parameter estimation in the model.

The joint posterior distribution for the parameters \( \alpha, \mu \) and \( \sigma^2 \) is given by

\[
p(\theta | X, X_0) \propto \prod_{i=1}^{l} p(x_i | \alpha, \mu, \sigma^2) \cdot \prod_{j=1}^{J} p(x_{0j} | \mu, \sigma^2) \cdot \pi(\alpha, \mu, \sigma^2)
\]

\[
= \prod_{i=1}^{l} \frac{1}{\alpha} \exp \left( \frac{\sigma^2}{2\alpha^2} - \frac{x_i - \mu}{\alpha} \right) \Phi \left( \frac{x_i - \mu - \sigma^2}{\sigma} \right).
\]

\[
\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi \sigma}} \exp \left[ -\frac{(x_{0j} - \mu)^2}{2\sigma^2} \right] \cdot \pi(\alpha, \mu, \sigma^2),
\]

where \( \pi(\alpha, \mu, \sigma^2) \) is the prior density function, which can easily incorporate useful information available from direct knowledge or previous experiments. Here, we consider weak, but proper independent prior distributions for the three parameters, to represent the case that no real prior information is available, for the purpose of fair comparison in our numerical experiments. For \( \mu \), we choose a normal prior with a sufficiently large variance to make the mean irrelevant. For both \( \alpha \) and \( \sigma^2 \), we use the conjugate prior with vague information, that is, the
The posterior distribution \( p(\alpha, \mu, \sigma^2 | X_i) \) is not analytically tractable. So a Metropolis–Hastings (MH) procedure is used to simulate samples from the posterior distribution, which is outlined in the Supplementary Material. We now proceed to discuss the background correction method under the Bayesian approach. As indicated in Equation (4), the conditional distribution of \( S_i | X_i \) is a truncated normal distribution, where \( 0 < S_i < X_i \). In the Markov chain Monte Carlo (MCMC) procedure, we can simulate a sample of \( S_i \) from this distribution at each step, obtain the conditional distribution of signals for each gene and then use the average of \( S_i | X_i \) samples as the background corrected signal for gene \( i \). An alternative way is to calculate the posterior means of the parameters, and then plug them into Equation (5) to get the background corrected signals. These approaches lead to almost identical results, and we use the second approach in our numerical experiments for comparison with the other methods.

4 SIMULATION

4.1 Comparison

We have discussed three methods for estimating \( \theta \) based on the model in (1) and (2). In this section, we first compare the performance of the three methods in parameter estimation and background correction under various parameter settings.

In our first experiment, we set \( \alpha \) to be 20, 50 or 100; \( \mu \) to be 100, 150 or 200; and \( \sigma \) to be 25, 35 or 45. For each possible \( \theta \) value, we simulated 100 datasets and for each dataset, we generated 45 000 observations for genes from (1), and 1000 observations for negative controls from (2). Supplementary Table 1 reports the mean squared errors (MSE) of parameter estimation for each method. Table 1 reports the MSEs of background corrected intensities. Under each setting, for example, the MSE of the MLE is defined as \( \text{MLE} \); and the MSE of background corrected intensities is defined as

\[
\frac{1}{100} \sum_{t=1}^{100} \frac{1}{45000} \sum_{i=1}^{45000} \left( E(S_{ti} | X_i = x_{ti}, \hat{\theta}_i) - E(S_{ti} | X_i = x_{ti}, \theta) \right)^2.
\]

(11)

The MSEs for \( RMA \), non-parametric and Bayes estimators are defined similarly. We also compare the MSEs of background corrected intensities for raw data without any background correction and the normexp model with saddlepoint estimation. In the tables, \( NP \) stands for the non-parametric estimator, \( B \) stands for the Bayes estimator, \( NES \) stands for the normexp model with saddlepoint estimation. \( RMA \), \( MLE \) and \( RAW \) are obvious.

| Table 1. | MSE of background corrected intensities |

We first discuss the performance in parameter estimation. From Supplementary Table 1, we can see that the \( RMA \) estimator is substantially worse than the \( NP \), \( MLE \) and \( B \) estimators. The overall performance of \( MLE \) and \( B \) are very similar. \( MLE \) and \( B \) work very well for estimating \( \alpha \) and \( \mu \), and the estimates are nearly as good as the truth, as indicated by the close-to-zero MSE values. There is actually not much difference in estimating \( \alpha \) and \( \mu \) among \( MLE \), \( NP \) and \( B \), though \( MLE \) and \( B \) appear to always be slightly better. However, \( MLE \) and \( B \) are much better than \( NP \) for estimating \( \sigma^2 \). As to background correction, we can conclude easily from Table 1 that the performance has the order

\( MLE > B > NP > NES > RMA > RAW \).

For \( RMA \), \( NP \), \( MLE \) and \( B \) methods, we also calculated the biases of parameter estimation (e.g. the bias of \( MLE \) is defined as \( \text{MLE} \)), and the average computing time over all the settings. \( RMA \) is seriously biased in all the settings, while all the other methods appear to be unbiased. In terms of computing speed, \( NP \) is the fastest, followed by \( RMA \) and \( MLE \), and \( B \) requires much more time (i.e., the average time in seconds is 0.00, 0.23, 1.46 and 26.18 for \( NP \), \( RMA \), \( MLE \) and \( B \), respectively).

4.2 Robustness checking
In practice, the normality assumption for background noise may not always hold. In our second experiment, we test the robustness of the estimators when the normality assumption is violated.

We generated true signal intensities from exponential(100), and background noise from Gamma($a$, $b$). Here, ($a$, $b$) takes values (64,3.125), (32.65,6.125) and (19.75,10.125) so that the mean $\mu$ of the background noise is fixed at 200, and the SD $\sigma$ takes values 25, 35 and 45, respectively. Supplementary Figure 1 shows clearly that the gamma distribution with the largest variance deviates the most from the normality assumption.

For each ($a$, $b$), we generated 100 datasets and proceeded with the parameter estimation and background correction as if the underlying true model had been (1) and (2). The top panel of Table 2 reports the MSEs of background corrected intensities defined in (11). Note that in (11), the first expectation was calculated under the assumed model, while the second expectation was calculated under the true data generating model.

From Table 2, we can see that again, all our methods perform much better than RMA. They work reasonably well, especially when the deviation from normality is not big. As the deviation gets larger, the performance worsens. Clearly, NP is the best amongst all methods. This is not surprising, since NP requires no distributional assumption when estimating parameters.

Another model assumption is that the distribution of $B_i$'s (the background noise for the beads carrying gene sequences), is the same as that of $B_{0j}$'s (the background noise for the beads carrying negative control sequences). Ideally, the noise sources should be the same for $B_i$'s and $B_{0j}$'s because the beads on the same array are hybridized and scanned simultaneously. In this sense, the assumption is reasonable. However, it is possible that the negative control beads also contain some weak signals because of the sequence selection, which may lead to the higher intensity levels of the negative controls compared with the true background noise. In this case, the mean intensity of the negative controls is greater than the mean intensity of the random noise. In our third experiment, we check the performance of the methods under this scenario.

We generated true signal intensities from exponential(100), the true background noise from normal distributions with mean $\mu$ and variance $\sigma^2$, and the negative control intensities from normal distributions with mean $\mu_0$ and variance $\sigma^2$. We set $\mu$ to be 100 or 150, $\delta=\mu_0-\mu$ to be 10 or 50 and $\sigma$ to be 25 or 35. The bottom panel of Table 2 reports the MSEs of background corrected intensities defined in (11).

Table 2 shows that NP, MLE and B perform better than RMA even when the noise distributions are different for negative controls and genes. When $\delta$ increases, the MSEs increase for NP, MLE and B. The performance of RMA does not depend on $\delta$ because its estimation uses information from genes only. On the other hand, NP is the most sensitive to $\delta$, because its estimator is calculated from negative controls only. MLE performs the best in all the cases, which is slightly better than B. Although both MLE and B use the information from negative controls, they rely more on the information from genes since the number of genes is much larger than the number of negative controls.

5 AN EXAMPLE

We use a leukemia study to examine the different approaches of parameter estimation and the subsequent background correction. In this study, mouse models of radiation-induced leukemia were applied to study the leukemogenic process, which have been proved to be potential tools for studying the pathogenesis of leukemia in humans. Illumina Mouse-6 V1 BeadChip mouse whole-genome expression arrays were used to obtain the gene expression profiles of acute myeloid leukemia (AML) samples from irradiated CBA mice who subsequently developed AML and the control samples. In this type of array, 46 120 genes and 1655 negative controls are randomly allocated on each array. The goal of the study was to identify the genes that express differently between leukemia and control samples. Ding et al. (2008) described the experiment in detail and demonstrated that using a background correction strategy enhanced the biological findings of the study.

We applied our methods to all the arrays in the study. Supplementary Table 1 gives an example of the point estimates and standard errors of the parameters for one array. Note that the RMA method does not provide standard errors for parameter estimation. The variances of the
To compare the performance of different methods, reverse transcriptase-polymerase chain reaction (RT–PCR) experiments were conducted on randomly selected genes in the study. RT-PCR is a highly sensitive technique for the detection and quantification of mRNA (messenger RNA) level, and so is regarded as the gold standard for gene expression levels. However, RT-PCR experiments are relatively time and labor intensive since they measure the expression level for one gene at one time. In our study, RT-PCR experiments were limited to 14 genes because of cost. We applied the raw data without background correction (RAW), the background subtraction method as mentioned in Section 1, RMA, NP, MLE and B for background correction, and then used the quantile–quantile normalization to remove the systematic variation amongst arrays. Note that after background subtraction, 12 out of the 14 genes have negative expression values for either AML or control samples. So the subtraction method is not considered in further comparisons. The log ratios of gene expression levels between leukemia samples and control samples were calculated for each method. The method that can generate the most consistent results with the RT-PCR results will be thought as the best method.

Supplementary Figure 2 indicates that for the randomly selected 14 genes, the background corrected expression levels from BeadArrays are highly correlated with the RT-PCR results. Background correction using MLE and B lead to the most consistent results with RT-PCR, and RAW performs the worst. We notice that without background correction, the log ratios are closer to 0 compared with the RT-PCR results, which is consistent with the data compression problem discussed in the Section 1.

Table 3 summarizes the association between the background corrected microarray results and the RT-PCR results based on linear regression models for the various methods. In the regression models, the response variable is \( \log_{10} \) ratio of gene expression between leukemia and normal tissues generated by microarrays with the different methods of background correction, and the independent variable is \( \log_{10} \) ratio of gene expression between leukemia and normal tissues generated by RT-PCR. If a background correction method generates consistent results with the RT-PCR method, then the slope of the corresponding linear regression model is expected to be close to 1. The slopes for RT-PCR versus MLE and B are both 0.87, and the slope for RT-PCR versus NP is 0.95, RT-PCR versus RMA is 1.20 and RT-PCR versus RAW is 0.52. In addition, using the RT-PCR results as the gold standard, the MSEs for MLE and B are smallest, followed by NP, and last RMA and RAW. All of the above indicate that applying background correction methods can increase the BeadArray data quality and amongst the methods, it appears that MLE and B are the best, with NP next, and RMA the worst. Also note that if we use the background subtraction method provided by Beadstudio software, most information for those 14 genes is lost. The normexp method with saddlepoint approximation method (NES) implemented in beadarray library was applied to the bead-level data for background correction. We applied this method to the leukemia study, and it does not generate more consistent results with RT-PCR results (MSE=0.19 and \( \beta=0.54 \)) compared with the methods we described in this article.

6 DISCUSSION

The Illumina BeadArray platform has become increasingly important because it is carefully designed to control noise and variation. However, the statistical methodology development for this platform is far behind that for Affymetrix and cDNA arrays, and there is ample space for improvement. In this article, we have described model-based background correction methods for Illumina BeadArrays. Built on the RMA convolution model, our model incorporates the information from over 1000 negative control beads, and improves the efficiency of background correction significantly, compared with the existing methods. We have considered three methods, namely, non-parametric, MLE and Bayes methods, for parameter estimation. All these methods have their own merits and are better than the RMA estimation method. The non-parametric method is very simple, and fast in calculation, which can provide a good starting point for the other two methods. The MLE is attractive in theory and has the best estimation efficiency overall. The Bayes method has nearly identical performance as the MLE. It is computationally slow, but it offers the flexibility to incorporate more complicated data structure and prior information into the model, which
is very useful when extra sources of data are available. Note that the methods of background correction compared in this article focus on removing background noise from auto fluorescence of the non-specific oligonucleotide on a spot. Because these methods do not intend to address local background noise, they can be applied to both bead-level data and summarized bead-type data. Our real data example does not show much benefit of applying the methods to bead-level data.

The model we used relies on two assumptions: the background noise is assumed to be normally distributed, and has the same distribution for both gene and negative control beads. The assumptions were made based on the empirical distributions from real data examples that we have been exposed to, as well as mathematical convenience. We caution that the assumptions might be violated for BeadArrays in some experiments. To lessen this concern, we have examined the robustness of our methods through simulations where the assumptions did not hold and through the leukemia study where the truth is not known. We find that our methods, especially \textit{MLE} and \textit{B}, perform reasonably well. Of course, it will be of great interest to develop a more flexible model to address these issues for potentially better results.

\textbf{Funding:} National Institute of Health (UL1RR024982, and NASA (NSCORS NNJ05HD36G, NAG9-1569).

\textbf{Conflict of Interest:} none declared.

\section*{Supplementary Material}

[Supplementary Data]
Click here to view.

\section*{Article information}

Published online Feb 4, 2009. doi: 10.1093/bioinformatics/btp040

PMCID: PMC2654805

Yang Xie,\textsuperscript{1,2} Xinlei Wang,\textsuperscript{3} and Michael Story\textsuperscript{2,4}

\textsuperscript{1}Division of Biostatistics, Department of Clinical Sciences, \textsuperscript{2}Simmons Cancer Center, University of Texas Southwestern Medical Center, \textsuperscript{3}Department of Statistical Science, Southern Methodist University and \textsuperscript{4}Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, USA

*To whom correspondence should be addressed.
Associate Editor: David Rocke

Received September 28, 2008; Revised December 18, 2008; Accepted January 17, 2009.

\textbf{Copyright} © The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

This article has been cited by other articles in PMC.

Articles from Bioinformatics are provided here courtesy of \textit{Oxford University Press}

\section*{REFERENCES}


