A Nonparametric Background Correction Method for Oligonucleotide Arrays

Monnie McGee, ZhongxueChen, Richard H. Scheuermann*, and Feng Luo†

Department of Statistical Science, Southern Methodist University
Dallas, TX 75275
mmcgee@smu.edu & zhongxue@smu.edu

**Department of Pathology, University of Texas Southwestern Medical Center*
scheurm@utsw.swmed.edu
6000 Harry Hines Blvd., Dallas, TX 75390

†*Department of Computer Science,Clemson University*
Clemson, SC, 29634-0974
luofeng@cs.clemson.edu

Abstract

Affymetrix GeneChip high-density oligonucleotide arrays are widely used in biological and medical research because of production reproducibility, which facilitates the comparison of results between experiment runs. In order to obtain high-level classification and clustering analysis that can be trusted, it is important to perform various pre-processing steps on the probe-level data to control for variability in sample processing and array hybridization. The Robust Multichip Average (RMA) model is a popular method of obtaining background corrected and normalized gene expression values from microarray data. It is based on the assumptions that the background noise is normally distributed, and the true signal is exponentially distributed. The quality of the final results depends on the validity of these underlying assumptions.

We propose a nonparametric (distribution-free) method to circumvent observed deficiencies in the distributional assumptions in models for background correction used by the standard signal processing algorithms. Our analysis of the data indicates that the intensities of mismatched (MM) probes that correspond to the smallest perfect match (PM) intensities can be used to estimate the background noise. Specifically, we obtain the smallest q_2 percent of the MM intensities that are associated with the lowest q_1 percent PM intensities, and use these intensities to estimate background. We compare several variations of this method with MAS 5.0 and RMA using various Affymetrix spike-in datasets. Comparisons with other methods on three spike-in data sets show that our nonparametric methods are a superior alternative for background correction of Affymetrix data.

Acknowledgement

The authors wish to thank William Schucany for a careful reading of this manuscript during its preparation.

1 Introduction

Affymetrix GeneChip arrays are widely used in biological and medical research. To interrogate a gene, 11-20 probe pairs are used. Each probe pair consists of a perfect match (PM), which is the perfect complement of a subsequence of the target transcript of interest (gene), and a mismatch (MM), which has the same composition as the PM sequence except that the middle base (13th) is changed to its Watson-Crick complement. Both the PM and MM probes are twenty-five nucleotides long. The MM probes were originally designed to be different at one base pair so that their intensities could be subtracted from those of the PM as a measure of non-specific hybridization.

In order to estimate gene expression values and perform high level analyses, such as classification and clustering, probe-level pre-processing of the data is necessary. Typically, there are three steps of preprocessing: background correction, normalization and summarization, although not necessarily in this order. It has been argued that background correction is the most crucial step for probe level processing (Bolstad, 2004, Choe, *et. al* ., 2005). Many methods have been proposed to preprocess microarray data. Among these methods, Microarray Analysis Suite (MAS5.0) (Affymetrix, 2001, 2002) and robust multichip average (RMA) (Irizarry, 2003a,b) are the most popular.

MAS 5.0 corrects background by dividing an array into sixteen (by default) regions (Affymetrix, 2001, 2002). The mean of the lowest 2 percent intensities in each region is used as an estimate of background noise. This approach to background correction assumes that the background noise is spatially biased and normally distributed, with the bottom 2 percent cutoff chosen somewhat arbitrarily.

For RMA, the background is estimated based on a convolution model, given by

$$X = S + Y, \tag{1}$$

where X is the observed PM intensity for a probe on the array, $S \sim \exp(\frac{1}{\alpha})$ is the true signal, and $Y \sim \mathcal{N}(\mu, \sigma^2)$ is the background noise. The normal noise distribution is truncated at zero so that the model does not return negative intensity values. The true signal can be estimated by

$$E(S|X = x) = a + b \left(\frac{\phi(\frac{a}{b}) - \phi(\frac{x-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{x-a}{b}) - 1} \right), \tag{2}$$

where $a = x - \mu - \sigma^2\alpha$, $b = \sigma$, $\Phi(\cdot)$ is the cumulative distribution function of the normal distribution, and $\phi(\cdot)$ is the density function of the normal distribution. The convolution model used by RMA is based on the assumptions that the background noise across the array is normally distributed and that the true signal is exponentially distributed.

Comparisons based on spike-in data have shown that RMA tends to be more accurate in estimating expression differences than other techniques (Irizarry 2003 a,b). However, there are some limitations for the RMA model. First, it is difficult to estimate the parameters for the model. We have observed that the parameter estimation method used in the affy package implementation of RMA (http://www.bioconductor.org) produces very poor estimates for μ, σ, and α for a simulated convolution of a truncated normal and an exponential distribution (unpublished).

Second, RMA does not use the information given in MM probes. The original intent of the MM probes was to provide a measure of non-specific hybridization, which could be subtracted from the PM intensities, leaving the true signal. MAS 5.0 was developed under the assumption that the PM and MM intensities within a probe pair have the same capability to catch non-specific hybridization. However, many researchers have reported that, for about one third of probe pairs, the MM probe has a larger intensity than the corresponding PM probe, suggesting that cross-hybridization is a problem. In addition, the correlation between PM and MM probe intensities ranges from 0.7 to 0.75, suggesting that the MM probes are hybridizing to the true target. Both of these observations seem to be true for all Affymetrix platforms. These facts combined have led researchers to believe that MM probes should be excluded from analyses (Bolstad, 2004, Irizarry *et.al.*, 2003a), but, in doing so, one is deleting fifty percent of the data. Choe, *et. al* . (2005) have shown that MM-corrected methods perform better on a DrosGenome1 spike-in data set than do PM-only methods. GCRMA (Wu, *et. al* .,2004) is a modification of RMA that uses additional information from the GC content of the probes and the MM intensities to perform background correction. GCRMA typically performs better than other methods, indicating that there is some use for the information in the MM intensities. However, the gains from using GCRMA are modest, and RMA continues to be the standard method for analysis among statisticians.

In this paper, we concentrate on the distributional assumptions underlying the background correction method in RMA. We show that these assumptions are unreasonable for most data sets. This observation, plus the difficulty of checking assumptions in the first place, motivate a nonparametric method. Three variations of the nonparametric method are explored, all of which use MM probe intensities for background correction. Two of the variations of the nonparametric method perform better than RMA and MAS 5.0 on three different spike-in data sets.

2 Materials and Methods

Equation 1 is built on the reasonable assumption that flourescence intensities from a microarray experiment are composed of both signal and noise, and that the noise is ubiquitous throughout the signal distribution. A convolution model of a signal distribution and a noise distribution is a good choice in such a situation. Figure 1 shows density estimates of one replicate from each of the fourteen experiments in the HGU133 spike-in data set. From this picture, it seems that a combination of a normal and an exponential might fit the data. However, Figures 2 and 3 (see Results) show that the convolution of a normal and an exponential distribution is not generally a good fit for microarray data. Considering this limitation, we propose a nonparametric method to estimate the background noise.

Our new background estimation method uses the information given in the MM intensities but in a different way than previous methods. While our method also assumes that the observed intensity is a sum of two parts, one is the background and the other is the true signal, it does not set any distributional constraints on the background and signal. Our methods are based on the assumption that MM probe signals corresponding to low intensity PM probes contain mainly background intensity. We can use the intensities of a proportion of the MMs that are associated with those PMs that have the lowest q_1 percent intensities. However, in some cases, high values of MM are associated with low PM value due to cross-hybridization. Thus, we should eliminate abnormally large MM intensities and use only the smallest q_2 percent of MM intensities to estimate background.

To choose the parameter q_1, we use an algorithm which calculates q_1 such that the proportion of MM intensities greater than the PM intensities for the smallest q_1% of the

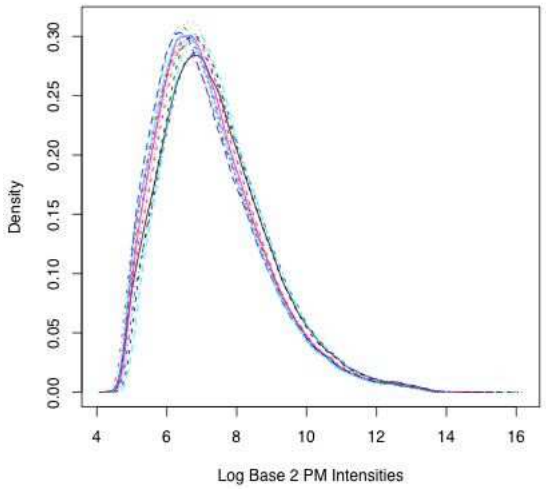

Figure 1: Density estimates of the log base 2 PM intensities from the HGU133 spike-in data set. Each line represents one replication from a different experiment.

data $\approx 50\%$. We believe that one of the reasons that MM intensities are sometimes greater than their corresponding PMs is non-specific hybridization. Therefore, in estimating q_1, we wish to get some measure of non-specific hybridization.

For finding q_2, we make the assumption that a PM intensity is bigger than a given number (say, median of PM) when the corresponding proportion of MM > PM is close to some small percentage (say 10%). One way to think of this percentage is as an estimate of the chance that a MM probe is cross-hybridizing to another probe (or that most of its signal is from non-specific hybridization). Therefore we take $q_2 = 1 - 10\% = 90\%$ as an estimate. In practice, q_2 does not affect the results very much. Figure 6 (Discussion) and supplemental boxplots of the distribution of MM intensities corresponding to PMs for various values of q_1 and q_2, for several platforms and experiments, confirm this result.

The Nonpar algorithm for background correction proceeds as follows:

1. Obtain the lowest q_1 percent PM intensities. q_1 is typically a small number (less than 30%).

2. Obtain lowest q_2 (typically 90% or 95%) of MM intensities associated with the PMs obtained in step 1. These MM intensities are a measure of background noise, and will be termed "noise" in the sequel.

3. Use a nonparametric density estimate to find the mode of the noise distribution. The mode is the estimated mean of the noise ($\hat{\mu}$).

4. Estimate the standard deviation of the background noise by calculating the sample standard deviation of the noise for values which are smaller than $\hat{\mu}$. Then $\hat{\sigma}$ is the sample standard deviation multiplied by $\sqrt{2}$.

5. Correct the background noise using one of the following methods:

A. For a given number k (*e. g.* 2), if the PM intensities are greater than $\hat{\mu} + k$, then the background corrected values equal [PM $- \hat{\mu}$; otherwise, PM $= k$. We term this method "Nonpar A".

B. For a given number k, if PM $> \hat{\mu} + k$, then PM $=$ PM $- \hat{\mu}$; otherwise, use a linear interpolation such that min(PM) $= \ell$ (a number < 1) and values of PM such that PM $= \hat{\mu} + k \rightarrow k$. This method is called "Nonpar B".

C. This method considers the contribution of $\hat{\sigma}$. For a given number k, if PM $> \hat{\mu} + k\hat{\sigma}$, then PM $=$ PM $- \hat{\mu}$. Otherwise, use a linear interpolation such that the min(PM) $= \ell$ and PM : PM $= \hat{\mu} + k\hat{\sigma} \rightarrow k\hat{\sigma}$. We call this method "Nonpar C".

Note that Nonpar B and Nonpar C are the same method with different parameters.

3 Results

In order to test the validity of the RMA model assumptions, we compared background noise distribution estimated by RMA with the standard normal distribution in both quantile-quantile (QQ) plots and density plots using the HGU95 and HGU133 spike-in data sets provided by Affymetrix http://www.affymetrix.com/support/technical/sample_data/datasets.affx, and the DrosGenome1 spike-in data set of Choe et al. (2005), also called the "Goldenspike" experiment. All calculations were done using the Bioconductor suite in the R software package for statistical analysis (Gentleman, *et. al* ., 2004). R code is provided in the supplemental material.

Quantile-quantile (QQ) plots are designed to compare the distributions of two data sets usually a "gold standard" and a test data set. Sometimes, the gold standard consists of simulated values from a distribution of interest (e.g. the normal distribution), and sometimes it is simply data observed from another experiment. If the gold standard is simulated from a known distribution, the purpose of the plot is to see if the observed data have that particular distribution. In either case, the sorted values for one data set (quantiles) are plotted on the horizontal axis, and the sorted values of the other data set on the vertical axis. If the plot results in a straight line, then the two data sets are assumed to have the same distribution.

Figure 2(a) shows a QQ plot of the log base 2 estimated background noise for all three replicates of the first experiment from the HGU95 spike-in data. The background was estimated using the RMA background correction method as coded in the affy package of Bioconductor (Gentleman, *et. al*, 2004). According to the RMA model assumption, the background noise should be normally distributed. Therefore, a plot of the background noise estimated from the RMA model versus values simulated from a normal distribution should produce a straight line, and we see that this seems to be the case. Thus, for this data set, assumption of normality for the background noise seems to be reasonable.

Figure 2(b) is a QQ plot of the log base 2 background corrected PM intensities (using the Bioconductor implementation of RMA) versus observations simulated from an exponential distribution. The data set is the HGU95 spike-in data. The rate parameter used for the simulated exponential distribution is equal to the estimated rate parameter of the signal using RMA. The QQ plot for the background corrected (signal) intensities does not show a straight line; in fact, it shows that the distribution of the signal is much heavier tailed than one would expect if the data were exponentially distributed. This suggests that either the exponential model is not a good one for the signal from the PM intensities, or the background correction algorithm used for RMA is flawed.

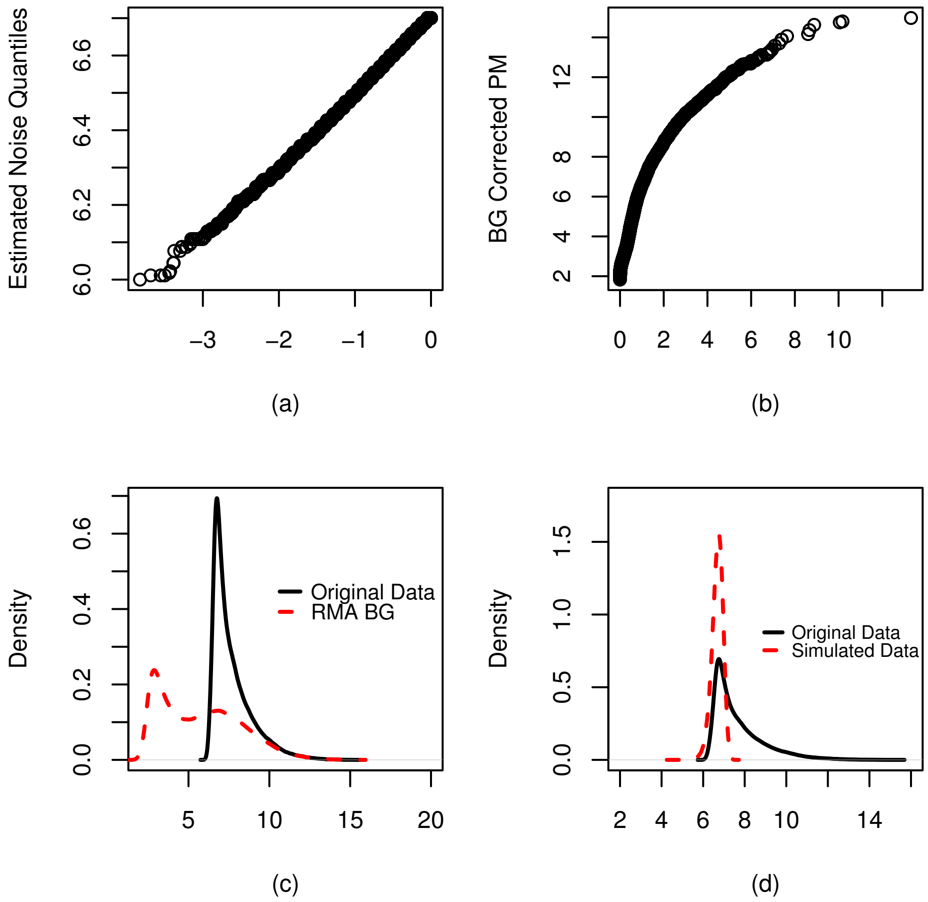

Figure 2: Quantile plots and density plots of background corrected HGU95 Spike-In data. (a) Quantile-quantile plot with quantiles of the standard normal distribution on the horizontal axis and quantiles of the noise distribution as estimated by RMA. (b) Quantiles of an exponential distribution versus the background corrected probe-level intensities from the RMA model. (c) Density estimates of the log base 2 PM intensities for the original (uncorrected) data (solid line) and the estimated background using RMA (dashed line). (d) Density estimates of the log base 2 PM intensities from the original data versus a simulated convolution of a normal distribution and an exponential distribution. The parameters for the normal and exponential distributions were obtained using estimates given by RMA.

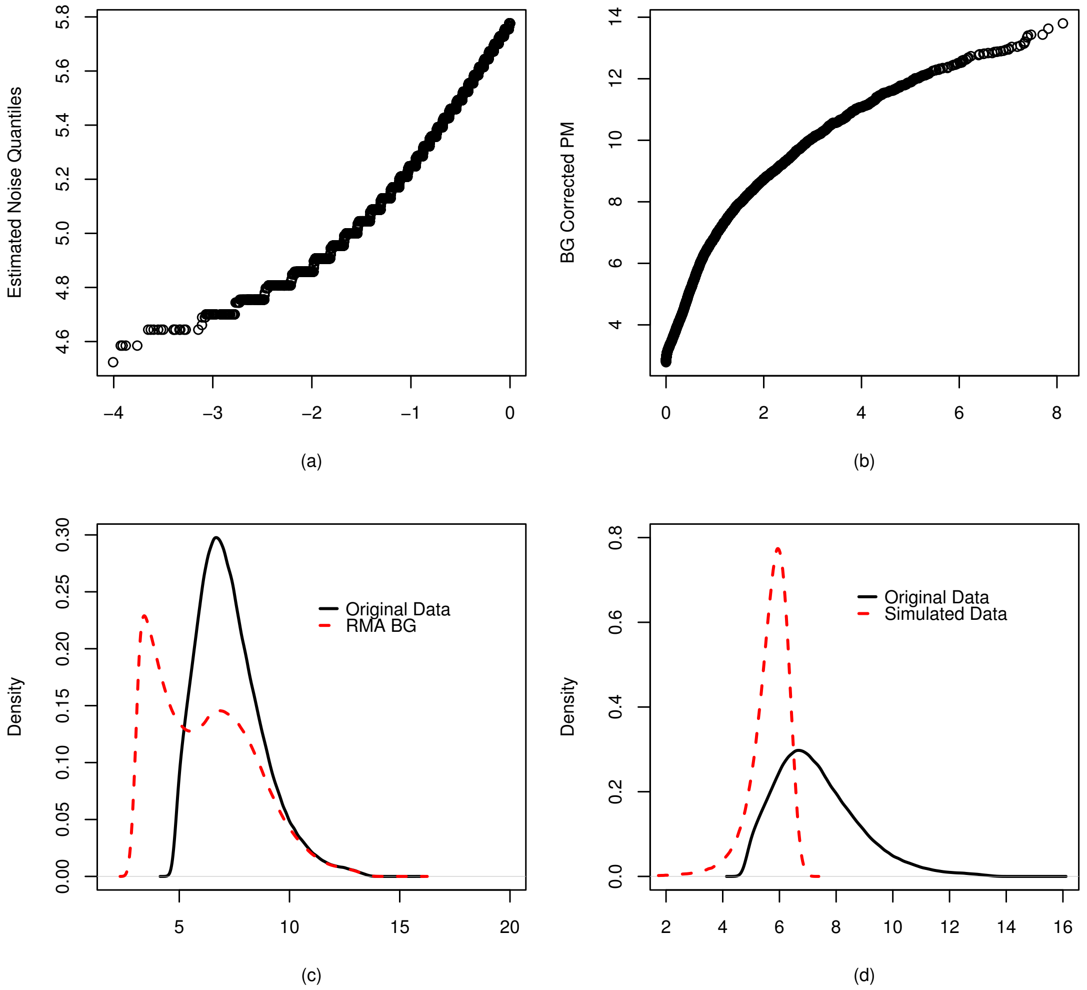

Figure 3: Quantile plots and density plots of background corrected HGU133 Spike-In data. (a) Quantile-quantile plot with quantiles of the standard normal distribution on the horizontal axis and quantiles of the noise distribution as estimated by RMA. (b) Quantiles of an exponential distribution versus the background corrected probe-level intensities from the RMA model. (c) Density estimates of the log base 2 PM intensities for the original (uncorrected) data (solid line) and the estimated background using RMA (dashed line). (d) Density estimates of the log base 2 PM intensities from the original data versus a simulated convolution of a normal distribution and an exponential distribution. The parameters for the normal and exponential distributions were obtained using estimates given by RMA.

Figure 2(c) shows density estimates of the original observed PM intensities (solid line) and the same intensities after background correction using RMA (dashed line). Under the assumption of the RMA model, the background corrected intensities should exhibit an exponential distribution. However, the signal from these data have two modes, suggesting that the estimated signal is composed of a mixture of two normal distributions rather than an exponential distribution, at least for this data set. Incidentally, this density estimate supports the idea that there are two groups of genes in this data set - genes that are expressed at low levels, and fewer genes expressed at higher levels.

Figure 2(d) shows the same density estimate of the original PM intensities that was seen in plot (c), but now this density is plotted against a density consisting of a simulated convolution of a truncated normal and an exponential, using parameters estimated by the RMA algorithm given in Bioconductor. It is clear that the convolution model does not fit the original data.

The results shown for the HGU95 spike-in data apply to the HGU133 spike-in data (Figure 3), with one notable exception. Figure 3(a) shows that the background as estimated by RMA does not have a normal distribution, since the QQ plot does not display a straight line. We have also analyzed colon cancer cell line data generated using the HGU133plus2.0 platform (data not shown), as well as data from the GoldenSpike experiment (Choe, *et. al* ,2005) (see supplemental material). The results from the colon cancer cell line data are much the same as from the HGU133 spike-in experiment: neither the distribution of the background or signal follow the respective distributions assumed by the RMA model. QQ plots from the GoldenSpike experiment seem to support a normally distributed background, but not an exponentially distributed signal. These observations lend credibility to the notion that different platforms may require different preprocessing methods (Allison, *et. al* ,2006), or that the preprocessing approach should not rely on these distributional assumptions.

We further apply Nonpar and RMA to the two Affymetrix Latin-Square spike-in data sets (HGU113 and HGU95). Each of these data sets contains several spiked-in transcripts in known locations on a set of chips. Affymetrix has already reported that certain probe pairs for transcripts 407_*at* and 36889_*at* had been found to perform poorly in the HGU95 spike-in data. In addition, other researchers have found that the number of spike-in probe sets should be 16 instead of 14. Wolfinger and Chu (2003) and Cope *et. al* . (2004) report that probe set 546_*at* should be considered with the same concentration as 36202_*at* since both of them were designed against the target Unigene ID Hs. 75209. Further, probe set 33818_*at* should be included as a spiked transcript in the 12th column of the Latin square design. Our definition of spike-ins for the HGU95 data does not include 407_*at* and 36889_*at* , but includes 546_*at* and 33818_*at*. We make no changes to the original spiked-in transcripts for the HGU133 data.

Figure 4 shows the ROC curves generated from results of analysis to identify differentially expressed genes using RMA, MAS 5.0, and the nonparametric methods on the HGU95 spike-in data. In this case, $k = 2$, $q_1 = 30\%$ and $q_2 = 90\%$ for all three nonparametric methods. For these spike-in data sets, true positive and false positive results can be determined based on the nature of the Latin square design. In this case, the Nonpar A approach gives very similar false positive rates to MAS 5.0. Both of these methods perform worse than RMA, which, in turn, performs worse than Nonpar B and Nonpar C. Recall that, according to Figure 2, the normal distribution is a reasonable fit to the background as estimated by RMA. In addition, there are few spiked-in transcripts for these data. These facts combined may explain the similar performance of RMA and the nonparametric methods.

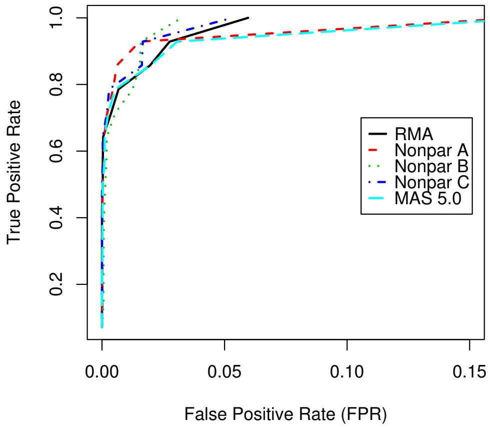

Figure 4: ROC curves generated from the Affymetrix HGU95 spike-in data for RMA (black solid line), MAS 5.0 (light blue solid line) and the three nonparametric methods. For all three nonparametric methods, $q_1 = 30\%$ and $q_2 = 90\%$. Details of Nonpar A (red dashed line), Nonpar B (green dotted line), and Nonpar C (blue dot-dash line) are described in Materials and Methods.

Figure 5 shows the ROC curves generated from the HGU133 data. For these data, Nonpar B and C outperform RMA, as does MAS 5.0. From Figure 3, we see that the normal distribution is not a good fit to the background noise as estimated by RMA. In this case, a nonparametric approach works better because there is no distributional assumption on the background. Other contributing factors could be larger number of the spike-in transcripts (42 for the HGU133 data versus 14 for the HGU95 data), and the different chip platform. Taken together, these findings are consistent with the concern of Allison, *et. al* . (2006) that current methods like RMA may be optimized for specific spike-in data sets and may not works as well for real data. Since RMA was developed before the HGU133 spike-in data was available, it may be the case that RMA was optimized to perform well on the HGU95 spike-in data.

We also use a third spike-in experiment to examine the relative performance of the nonparametric algorithms versus other algorithms. Choe. *et. al.* (2005) introduced a series of spike-in experiments on the DrosGenome1 chip, which they call the GoldenSpike experiment. In addition to being on a different platform from the Affymetrix spike-in data, the GoldenSpike experiment contains 1331 spiked-in transcripts whose levels are varied and 2,551 RNA species whose levels are held constant between the control and test array sets. The large number of spiked-in transcripts allows for more accurate estimates of the false positive and false negative rates and provides an RNA mix that more closely resembles total cellular RNA. Furthermore, no transcript targets were included for approximately two-thirds of the probe sets, allowing for an accurate definition of background data. In contrast, Affymetrix uses an uncharacterized RNA background for their spike-in data sets. Lastly, the fold differences between the test and control array sets for some of the spike-in transcripts are very low (1.2 fold), which allows an estimate of the reliability and sensitivity of detection of small fold differences.

No differences could be seen among the methods in a ROC plot using all spiked-in genes

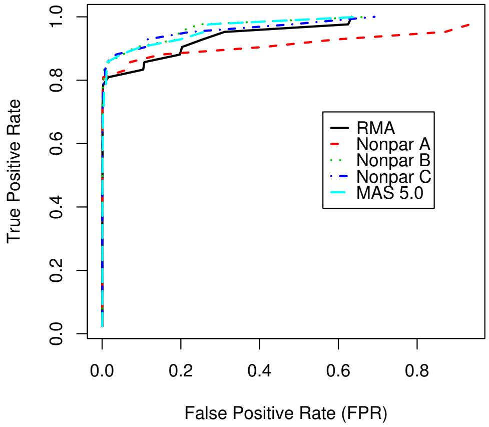

Figure 5: ROC curves generated from the Affymetrix HGU133 spike-in data for RMA (black solid line), MAS 5.0 (light blue solid line) and the three nonparametric methods. For all three nonparametric methods, $q_1 = 30\%$ and $q_2 = 90\%$. Details of Nonpar A (red dashed line), Nonpar B (green dotted line), and Nonpar C (blue dot-dash line) are described in the text.

on the GoldenSpike data (plot not shown). Therefore, we divided the data into low, medium, and high fold-difference values. Note that the actual concentrations are not known, only the fold-difference between spiked-in test samples and control samples. Low spike-ins were defined as those probe sets with fold-difference values between 1.2 and 1.5, medium had fold-differences between 1.7 and 2.5, and high spike-ins had fold-differences greater than 3.0. ROC curves for the comparisons of RMA, the three nonparametric methods, and MAS 5.0 were plotted for these three groups of spike-ins (see supplementary material). For very small false positive rates, MAS 5.0 and Nonpar C almost always give larger true positive rates than the other methods. In general, all methods are much better at detecting the large spike-ins than they are at detecting the low and medium spike-ins.

4 Discussion

The RMA convolution model for background correction of microarray data from Affymetrix platforms is very popular. This model assumes that the observed value of flourescence intensities is composed of an exponentially distributed signal with underlying normally distributed noise. This idea of a combination of signal and noise is quite reasonable, but the analysis presented here indicates that the distributional assumptions are not always correct. In order to examine the assumption of normally distributed background noise, we performed background correction using the convolution model and plotted the estimated background intensities versus a normal distribution using a quantile-quantile plot for three spike-in data sets. While the assumption of normally distributed noise is feasible for the GoldenSpike data and the HGU95 spike-in data, this assumption is not upheld for the HGU113 spike-in data. We also examined the background corrected intensities, which are purported to represent the true signal, against the exponential distribution. QQ plots show that the background corrected signal is clearly not exponentially distributed for any of the data studied here.

We devised a method which uses q_2th percentile of the MM signal corresponding to the smallest q_1 percentage of PM intensities to estimate background noise. The algorithm for choosing the value of q_1 is very stable, almost always choosing the same value of q_1 for a given platform. For example, two experiments completed on the HGU95 platform will have very similar values of q_1 (approximately 0.25). In other words, the values of q_1 are more platform dependent than they are experiment-dependent. This fact supports the notion that different normalization procedures are required for different platforms (Allison, *et. al.*, 2006).

Figure 6 shows boxplots of the MM intensities from one randomly selected chip from the HGU95 spike-in data. Values of q_1 are on the horizontal axis, and the distribution of intensities on the original scale can be seen on the vertical axis. Note that near $q_1 = 0.25$, the median of the boxplots are similar. One can think of this region as the point where we capture some background signal resulting from non-specific hybridization, as well as a global background fluorescence.

Once q_1 and q_2 are chosen, three different ways for computing the background corrected intensities were evaluated. One approach is to use a simple mean subtraction (Nonpar A), and the other two approaches are based on linear interpolations, where one method uses an estimate of the standard deviation of the noise to compute the corrected intensities (Nonpar B and Nonpar C). In general, the method (Nonpar C) that takes into account the standard deviation of the noise performs better than the other two nonparametric methods, and Nonpar A typically performs worse than RMA. Because the variability of the intensities can be quiet large, even for small intensities, accounting for this variability means that Nonpar C gives more accurate estimates of the background noise.

Testing the methods on the GoldenSpike data reveals that MAS 5.0 and Nonpar C methods are typically better than the other methods (see supplemental material). This finding would seem to be a contradiction as to what has been found in previous studies about the performance of MAS 5.0 vs. RMA. However, in searching the references listed below, we found that none of the articles, with the exception of Bolstad (2004) and Choe, *et.al.* (2005), examined ROC curves with both RMA and MAS 5.0. The former article

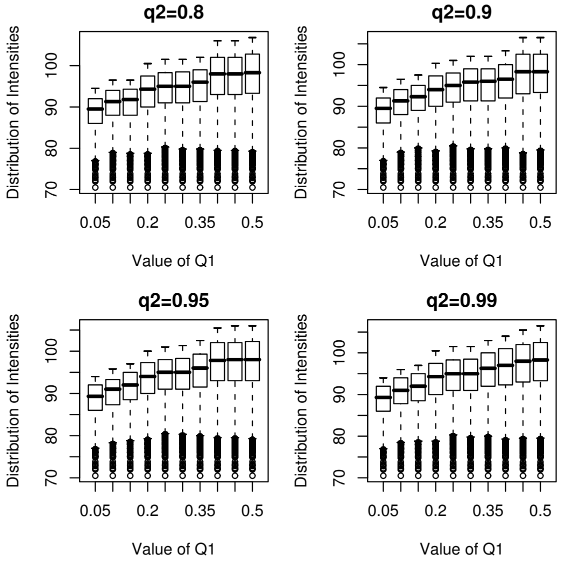

Figure 6: Boxplots of MM intensities for a Control Chip from the HGU95 spike-in data.

shows an ROC curve calculated using the HGU95 spike-in data, where MAS 5.0 performs almost as well as RMA. However, Ideal Mismatch and MAS 5.0/Ideal Mismatch perform very poorly. Choe, *et.al.* (2005) show MAS 5.0 outperforming RMA for the DrosGenome1 platform. Since the GoldenSpike data contain a large number of spiked-in transcripts, we expect that the estimates of false positive and false negative rates are more reliable than estimates from the Affymetrix spike-in data sets. Thus, the Nonpar C method appears to be consistently the best approach regardless of the data sets and array platforms used.

5 Conclusion

We have shown that microarray data from different Affymetrix platforms do not meet the assumptions of the convolution model used by RMA for background correction. In some cases, the estimated background does not follow a normal distribution (e.g. for the HGU133 platform). In every case examined, the resulting estimated real signal did not follow a simple exponential distribution. To circumvent these problems, we devised a distribution-free method to subtract background noise (Nonpar). This method, especially the Nonpar C variation, tended to perform better than RMA and MAS 5.0 across a variety of experiments and array platforms.

This finding has three important implications. First, any method that does not make an attempt to account for non-specific hybridization will not perform very well in practice. We attempted to account for this by using MM intensities to obtain an estimate of background noise. The convolution model subtracts a global estimate of noise that is not probe-dependent. Second, any background correction method based on assumptions that the background noise is normally distributed and that the real signal is exponentially distributed may not be valid for any given array platform. Third, it is clear that more research is needed into background correction methods for microarray data. In particular, we need to develop an understanding of the reasons certain methods perform better on certain platforms, and devise better ways to estimate the background noise.

References

Affymetrix, Inc (2001). ”Statistical Algorithms Reference”. Data Analysis Fundamentals Technical Manual, Chapter 5. www.affymetrix.com.

Affymetrix, Inc (2002). Statistical Algorithms Description Document. www.affymetrix.com.

Allison DB, Cui X, Page GP, and Sabripour M (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7, 55-65.

Bolstad, BM (2004). *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* Dissertation. University of California, Berkeley.

Bolstad BM, Irizarry RA, Astrand M, and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on the variance and bias. *Bioinformatics*, 19, 185-193.

Choe SE, Boutros M, Michelson AM, Church GM, and Halfon MS (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6:R16 http://www.geneomebiology.com/2005/6/2/R16.

Cope LM, Irizarry RA, Jaffee HA, Wu Z, and Speed TP (2004). A benchmark for Affymetrix GeneChip Expression measures. *Bioinformatics*, 20, 323-331.

Gentleman RC, Carey VJ, Bates DM, Bolstad BM, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry RA, Cheng Li FL, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JWH, and Zhang J (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249-264.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31, 4, e15.

Wolfinger, R and Chu, T (2003). Who are those strangers in the Latin Square? *Methods of Microarray Data Analysis III*, Johnson, KF. and Lin, SM (Eds.). New York: Springer.

Wu Z, Irizarry RA, Gentleman R, Martinez Murillo F, Spencer F (2004) A Model Based Background Adjustement for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99, 909-917.

Appendix A: Effect of q_1 and q_2 on the Noise Distribution

The following boxplots show various values of q_1 and q_2 for the GoldenSpike data set and the HGU133 spike-in data set. The value of q_2 is not as important as the value of q_1. Typically, q_1 is chosen as the value where the median for the data stabilizes. This value is 0.3 for the HGU133 spike-in data, and 0.2 for the GoldenSpike data (Choe, *et.al.*, 2005).

Appendix B: ROC Curves for the GoldenSpike Experiment

ROC Curves for the GoldenSpike Experiment were computed in order to compare the performance of RMA, MAS 5.0, and the nonparametric algorithms. No differences could be seen among the methods in a ROC plot using all spiked-in genes on the GoldenSpike data (plot not shown). Therefore, the data were divided into low, medium, and high fold-difference values. Only the fold-differences between spiked-in test samples and control samples are known for this data set. Low spike-ins were defined as those probe sets with fold-difference values between 1.2 and 1.5, medium had fold-differences between 1.7 and 2.5, and high spike-ins had fold-differences greater than 3.0. For very small false positive rates, MAS 5.0 and Nonpar C almost always give larger true positive rates than the other methods. In general, all methods are much better at detecting the large spike-ins than they are at detecting the low and medium spike-ins.

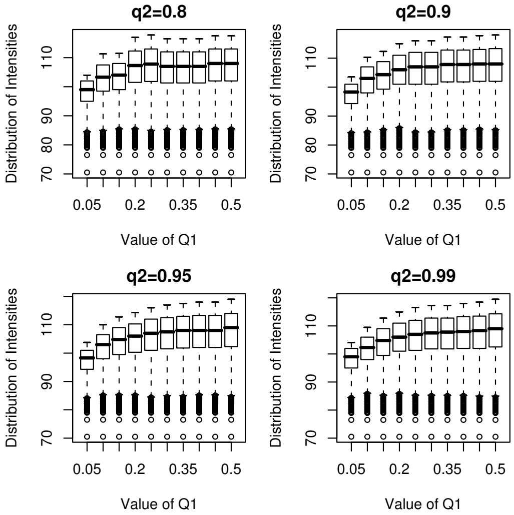

Figure 7: Boxplots of MM intensities for various values of q_1 and $q_2 = 0.8, 0.9, 0.95,$ and 0.99. The data are from the first replicate of the control chips (those with transcripts spike-in at no change from baseline) from the GoldenSpike experiment.

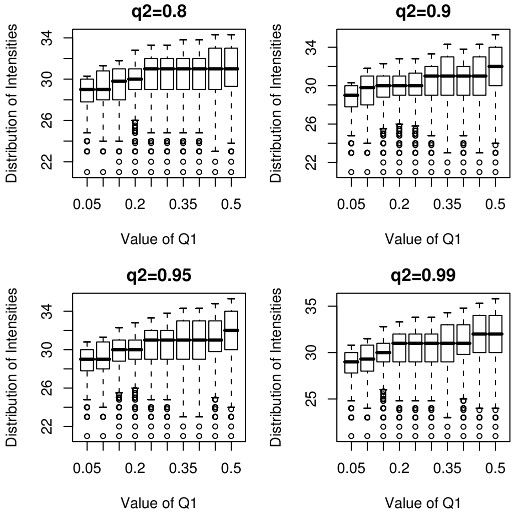

Figure 8: Boxplots of MM intensities for various values of q_1 and $q_2 = 0.8, 0.9, 0.95,$ and 0.99. The data are from the HGU133 spike-in data set.

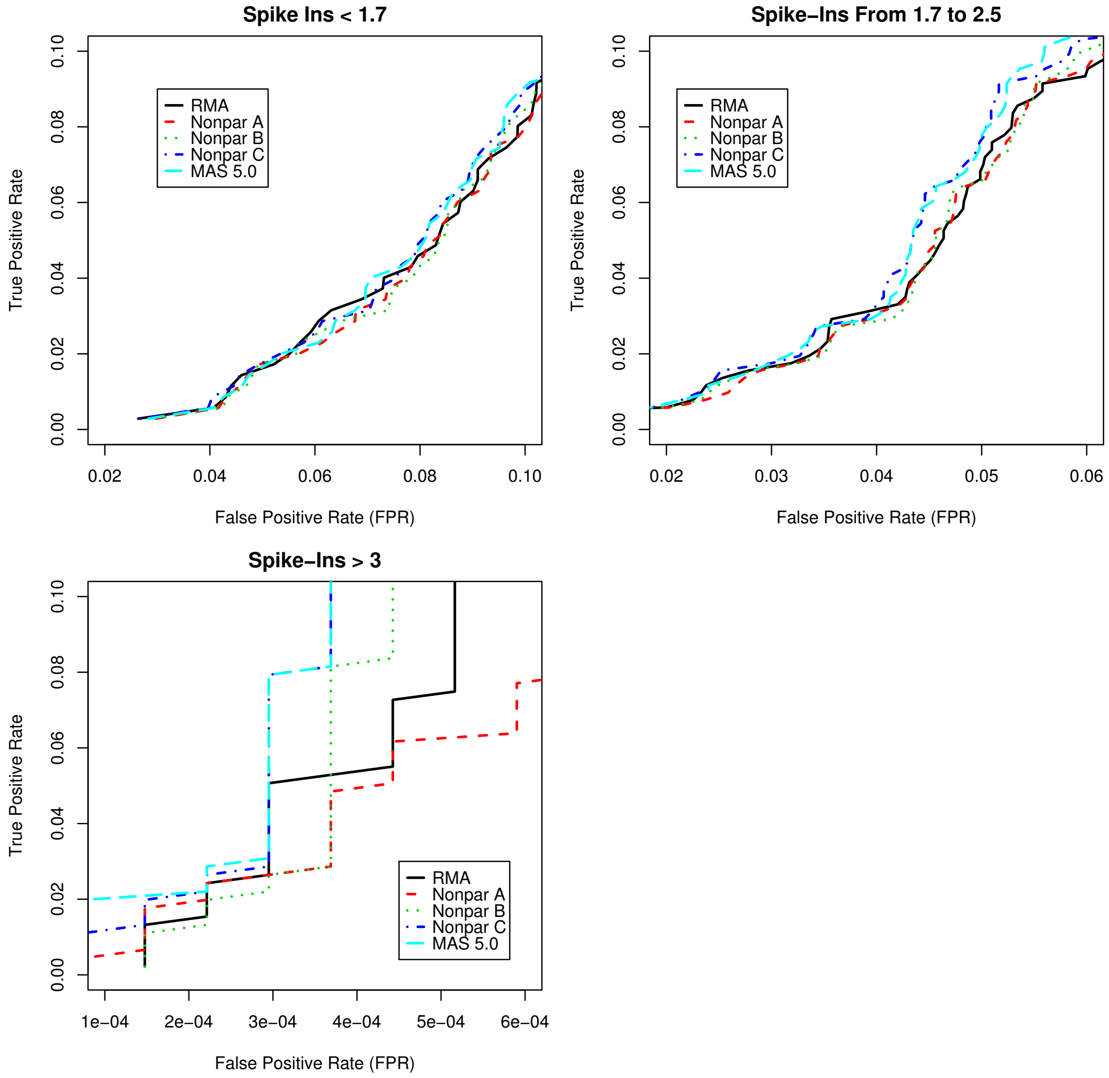

Figure 9: ROC Curves for the GoldenSpike experiment the scale for each graph was adjusted in order to enhance small differences among the methods.