

Parameter Estimation for the Convolution Model for Background Correction of Affymetrix GeneChip Data

Monnie McGee ^{a,*}, Zhongxue Chen ^{a,b}

^a*Southern Methodist University, Department of Statistical Science, 3225 Daniel
Avenue Room 144 Heroy, Dallas, Texas, 75275-0332, USA*

^b*Department of Pathology, U.T. Southwestern Medical Center, 5323 Harry Hines
Blvd., Dallas, TX 75390-9072, USA*

Abstract

There are many methods of correcting microarray data for non-biological sources of error. Authors routinely supply software or code so that interested analysts can implement their methods. Even with a thorough reading of associated references, it is not always clear how requisite parts of the method are calculated in the software packages. However, it is important to understand such details, as this understanding is necessary for proper use of the output, or for implementing extensions to the model.

In this paper, the calculation of parameter estimates used in Robust Multichip Average (RMA), a popular algorithm for background correction and normalization of microarray data, is elucidated. RMA models observed microarray data with a convolution of the true signal, assumed to be exponentially distributed, and a background noise component, assumed to have a normal distribution. A conditional expectation is calculated to estimate signal. Estimates of the mean and variance

of the normal distribution and the rate parameter of the exponential distribution are needed to calculate this expectation. Simulation studies show that the current estimates are flawed; therefore, new ones are suggested. When the new parameter estimates are used, it is shown that RMA is more sensitive and specific than previously thought.

Key words: microarray, RMA, Bioconductor, simulation, ROC curve

1 Introduction

Gene expression microarrays allow a researcher to measure the simultaneous response of thousands of genes to a stimulus. The availability of such technology has had a profound impact on molecular biology and related disciplines. For example, a search on PubMed for the term “microarray” appearing in the title or abstract of an article between January 1, 2000 and December 31, 2005, produced 12,479 articles. Correspondingly, interest in the analysis of microarray data has surged among the statistical community. The analysis of microarray data involves background correction and normalization of the data, with the purpose of identifying genes that are differentially expressed between two or more samples, tissues, or conditions. There are dozens of methods for analyzing microarrays, many with software packages to promote the use of the methods. Before explaining the analysis methods, it is important to explain how microarrays are produced.

* Corresponding Author.

Email addresses: mmcgee@smu.edu (Monnie McGee), zhongxue@smu.edu

(Zhongxue Chen).

URL: <http://faculty.smu.edu/mmcgee> (Monnie McGee).

There are four nucleotides which make up DNA, and they are denoted A, C, G, and T. A and T nucleotides always bind to each other, and G and C nucleotides always pair with each other to form double-stranded DNA. During transcription, DNA is split into single-stranded RNA. Microarray technology exploits the propensity of single-stranded messenger RNA (mRNA) or complementary DNA (cDNA) to bind to a sequence of complementary nucleotides to form DNA. In general, a single-stranded collection of probes, anywhere from 25 to 90 base pairs in length, is affixed to some sort of medium, usually a glass slide or a silicon chip. Then, a solution containing fluorescently labeled single-stranded target cDNA or mRNA probes is washed over the fixed probes, and it is assumed that the targets bind to their perfectly complementary sequences. Once the excess target is rinsed from the chip, the fluorescent label is excited using a laser scanner. The amount of fluorescence as measured by the scanner is proportional to the amount of binding (or gene expression) between the target cDNA and the fixed probes. Both [17] and [18] give excellent detailed explanations of the molecular biology pertinent to the design of microarrays.

The leading manufacturer of commercially produced microarrays is Affymetrix, Inc. Affymetrix GeneChips have a unique structure that affects the way they are analyzed [3]. Instead of using two dyes (e.g. red and green) to label two different types of cell targets on the same chip, Affymetrix employs a single-channel method. The fixed probes are of two types: perfect match (PM) and mismatch (MM). Both PM and MM probes are twenty-five nucleotides in length. PM probes are designed to be perfectly complementary to a 25 nucleotide sequence of a section of a unique gene. MM probes have the same sequence as the PM probes, except that the thirteenth base is changed to its complement (i.e. $A \iff T$, or $C \iff G$). Every PM is paired with a

MM, and the two together are called a probe pair. Clearly, a sequence of 25 nucleotides will not provide good sensitivity or specificity for a gene that is hundreds of nucleotides in length. Therefore, Affymetrix uses eleven to twenty such probe pairs, called a probe set, to interrogate each gene.

The process by which gene expression measures are determined is subject to error. First, the data are fluorescence intensities read by a scanner, which are not really the true fluorescence intensities of the material. The true fluorescence intensities are only a surrogate for the measure we really want, which is gene expression. There are also errors introduced by non-specific hybridization, cross-hybridization, quality of RNA extraction, scanner differences, *etc.* Therefore, microarray data must be background corrected and normalized before it can be used to determine differential expression of genes. This is true of both two-channel and single-channel arrays. The probe sets in Affymetrix GeneChips must also be summarized to obtain a single expression level for each gene.

The three most popular algorithms for background correcting, normalizing, and summarizing Affymetrix arrays are MAS 5.0 [1,2], dChip [14,15], and RMA [12]. All three are implemented in Bioconductor [9], a suite of packages programmed in the R language for statistical analysis [10]. Although authors of new analysis methods describe their algorithms in detail in their papers, it is sometimes difficult to understand exactly how the methods are implemented in software packages. For example, the Bioconductor implementations of MAS 5.0 and dChip give slightly different results than the original software packages, because these packages are commercially produced, and the source code is not available. Bioconductor support staff used the relevant papers to try to reproduce these methods. However, even in a well-written paper, it is some-

times difficult to understand exactly how a method might be programmed. If a package is open-source, then one can examine the source code, but this can be tedious and time-consuming. However, an understanding of the coding of the method is essential for an understanding of the statistical properties of the resulting measurements of differentially expressed genes.

In this paper, the parameter estimation method used in RMA is elucidated and examined. It is shown that the parameter estimates obtained via the current implementation are grossly inaccurate, and better estimates are devised. Section 2 gives an overview of RMA and the calculation of its parameters as currently implemented in Bioconductor. Section 3 gives the results of simulations which show that these parameter estimates are highly variable and biased, even under ideal conditions. New parameter estimates are given, and these estimates are shown to be much more stable. In section 4, microarray data, in which the true differentially expressed genes are known (so called spike-in data), are used to show that the new parameter estimates are more sensitive and specific than the current estimates. The paper concludes in section 5.

2 RMA Defined

It is biologically sound to assume that fluorescence 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 a density estimate of log base 2 PM intensities from one of the Affymetrix spike-in experiments (explained in Section 4). Given this pic-

ture, a model using the combination of a truncated normal distribution and an exponential distribution seems reasonable. These distributions have the added advantage that they are easy to manipulate mathematically, in order to calculate the conditional expectation given in (2).

The underlying assumption in RMA is that observed PM intensities are a convolution of normally distributed noise and exponentially distributed signal. More precisely,

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

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 to model that there are no negative intensity values. Then, 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. In practice, it is only necessary to compute the first terms in both the numerator and the denominator, since the other terms either cancel each other or are negligible [5].

RMA uses only the PM probes to perform its series of algorithms for background correction, normalization, and summarization of Affymetrix GeneChip data. Therefore, from this point forward, any reference to “intensities” will imply perfect-match intensities only, unless otherwise stated. Further, this paper will concentrate on the background correction method in RMA, and not on methods of normalization and summarization. However, as it was originally designed, the RMA method includes background correction using a convolution

model, normalization using quantile normalization [6,12] and summarization via median polish [5,11,19].

The parameters μ , σ and α need to be estimated from the data so that an estimate of the signal can be obtained via (2). A careful examination of the code for the background estimation in RMA showed that the following steps are used to estimate the requisite parameters (Bioconductor code is given as supplemental material).

Let $m = \max(f(x))$, where $f(x)$ is the density function of the observed intensities from the microarray data file. The value of x at which this maximum occurs, x_m , is the mode of the intensities.

- $\hat{\mu} = \arg \max[f(x|x < x_m)]$. In other words, once x_m is given, the estimate of the mean is the mode of the intensity values less than x_m .
- $\hat{\sigma}$ is the sample standard deviation of the intensity values which are less than $\hat{\mu}$, multiplied by a factor of $\sqrt{2}$.
- $\hat{\alpha}$, the rate parameter of the exponential distribution, is calculated as the inverse of the mode of the intensity values greater than x_m . In other words, $\hat{\alpha} = \arg \max[f(x|x > x_m)]^{-1}$

The rationale behind these estimates relies heavily on the validity of the convolution model (1). In general, it is thought that the smallest intensities are most affected by noise. Therefore, the smaller intensities (those less than the overall mode, x_m) are thought to be strictly from a normal noise distribution. The signal follows an exponential distribution, thus the values greater than the mode of the entire distribution should contain mostly signal. The rate parameter estimate is, in a certain sense, an estimate of where the noise distribution ends and the signal distribution begins. It is assumed that noise

permeates the intensities at every level, not just at the smaller intensities. However, some of the very largest intensities are mostly signal, meaning that any noise component is negligible. Using that rationale, the estimate of μ as currently implemented by Bioconductor is almost surely an overestimate of the mean of the noise distribution, and the estimate of α is nearly always an underestimate of the rate parameter of the signal distribution.

It is impossible to examine the validity of the convolution model analytically, since the true signal and true nature of the background noise is unknown. However, it is possible to use a simulation experiment to examine the performance of the parameter estimates in the situation where the underlying model is truly a convolution of a truncated normal distribution and an exponential distribution. The simulations are described in the next section.

3 Simulation Results

A convolution of an exponential with a truncated normal with various combinations of values for the mean, variance, and rate parameter were simulated. Specifically, values $\alpha = 250, 500,$ and 1000 for the rate parameter of the exponential distribution, $\mu = 30, 50$ and $\sigma = 5, 10$ for the mean and variance of the truncated normal, were used. These values were deemed to be reasonable values for the parameters by an exploratory examination of a collection of real microarray data generated on Affymetrix human genome (HG-U133 plus 2.0) chips.

The HG-U133 plus 2.0 chip contains fifty-four thousand probe sets, which vary in length from eleven to twenty probe pairs per set. This translates into roughly

six hundred thousand intensities. For the simulation experiment, one thousand replications with samples of size 1,000,000 were run for all combinations of μ , σ , and α .

Seven different methods to estimate μ , σ , and α , were used in the simulations. These methods are enumerated below.

M1: RMA method, which was described in Section 2.

M2: $\hat{\mu}$ and $\hat{\sigma}$ are estimated in the same manner as method 1, but either the mean, median, seventy-fifth percentile or 99.95th percentile of the data values greater than the overall mode are used to obtain $\hat{\alpha}$.

M3: $\hat{\mu} = (\mu_1 + \mu_2)/2$, where μ_1 , σ_1 , and α_1 are estimated using RMA, and μ_2 given by plugging in μ_1 , σ_1 , and α_1 into a one-step correction, given by the following formula. The mode, x_m , satisfies the equation

$$\phi\left(\frac{x_m - \mu}{\sigma} - \alpha\sigma\right) = \alpha\sigma \left[\Phi\left(\frac{x_m - \mu}{\sigma} - \alpha\sigma\right) + \Phi\left(\frac{\mu}{\sigma} + \mu\alpha\right) - 1 \right]. \quad (3)$$

In practice, the penultimate term on the right hand side of (3) is nearly equal to 1; therefore, only the first term is used. Once $\hat{\mu}$ is calculated, it is used to obtain $\hat{\sigma}$ as in M1, and $\hat{\alpha}$ as in M2.

M4: Instead of estimating $\hat{\mu}$ using the mode of the data values less than x_m , which tends to overestimate the true mean, $\hat{\mu}$ is found using the mode of the intensities less than $2x_m$. $\hat{\sigma}$ is calculated as in M1, and $\hat{\alpha}$ is given by various percentiles of the observations greater than the estimated mean (M2).

M5: $\hat{\mu}$ and $\hat{\sigma}$ are estimated as in M4, and $\hat{\alpha}$ as in M2. Then, a one-step correction values for each parameter are obtained. The corrected values are averaged with the original estimates.

M6: The overall mode (x_m) is used by itself to estimate the mean, and the

intensity values less than the overall mode are used to estimate σ . α is estimated using either the mean, median, 75th percentile or 99.95th percentile of all of the intensities (not just those whose values are larger than x_m).

M7: $\hat{\mu}$ is the average of x_m and the estimate of μ given by the one-step correction (3). σ and α are estimated as in the previous scenario.

The mean-squared errors of $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\alpha}$, for the various combinations of simulated parameter values, are given in Tables 1 and 2. The MSEs presented in the table are for M3, the one-step correction method using different parameter estimates for α (mean, median, seventy-fifth percentile, and 99.95 percentile). This method was found to give the best overall performance, in terms of MSE, across almost all simulated scenarios. For example, the MSE for $\hat{\sigma}$ when $\mu = 30$, $\sigma = 5$, and $\alpha = 250$ using RMA to estimate the parameters is 92. When using M3, in which the mean of the observations greater than $\hat{\mu}$ estimates α , the MSE for $\hat{\sigma}$ is 1.35.

Some of the numbers in the tables seem to be identical, for example, the MSE for $\hat{\mu}$ in the first row of Table 1 is listed as 2.62 three times. In reality, these values differ in the third or fourth decimal place. The same is true for other values in the tables which seem to be equal. The MSEs for $\hat{\mu}$ and $\hat{\sigma}$ tend to be particularly close across all estimates. One might expect that MSEs for $\hat{\mu}$ and $\hat{\sigma}$ would be more similar to each other than they are to the MSE for $\hat{\alpha}$, because the data used to calculate $\hat{\sigma}$ depends on the value of $\hat{\mu}$. Further, $\hat{\mu}$ and $\hat{\sigma}$ are estimated by the same method: the one-step correction method (3). In contrast, $\hat{\alpha}$ is calculated by either the mean, the median, the seventy-fifth percentile, or the 99.95 percentile of all values to the right of $\hat{\mu}$. The resulting differences among the results of these estimators overwhelms any similarity in

the estimates due to the dependency of $\hat{\alpha}$ on $\hat{\mu}$ and $\hat{\sigma}$ via (3).

In practice, M5 performed almost as well as M3; however, M2, M4, M6, and M7 did not perform as well as the others. We will not mention them further except to say that their mean-squared errors (MSE) were still many times less than those of the estimates using the current implementation of RMA.

4 Performance on Spike-In Data

To test the performance of these methods on real data where the concentrations of some genes are known, the Affymetrix spike-in data sets were used. Affymetrix has developed two Spike-In data sets, one series on the HG-U95A chip and the other on the HG-U133A chip. Both data sets, and corresponding detailed descriptions, are freely available on the Affymetrix website http://www.affymetrix.com/support/technical/sample_data/datasets.affx.

The HGU95 data set has fourteen spiked transcripts at known locations in fourteen experiments. The spiked concentrations range from 0 pM to 1024 pM. There are several replicates of each experiment, giving a total of fifty-nine arrays. The HGU133 data set consists of 3 technical replicates of 14 different experiments using forty-two spiked transcripts at known locations. The spiked concentrations range from 0 pM to 512 pM, at finer gradations than for the HGU95 data. Therefore, the HGU133 data has a larger background population and more spike-ins at a greater range than the HGU95 data set. For both data sets, the spike-ins are arranged from experiment to experiment using a Latin square design.

There is disagreement in the literature about the nature of the spiked in genes

for the HGU95 data. The authors of Affycomp [8], an online package for the evaluation of the performance of new methods on the HGU95 and HGU133 spike-in data sets, contend that transcripts “33818_at” and “546_at” should be included as spiked-in transcripts. Another group claims that “1598_g_at” and “37658_at” should also be included [20]. Further, the transcripts “407_at” and “36889_at” have been noted to have poor behavior, as noted in the description of the HGU95 data on the Affymetrix website. For these analyses, the Affycomp definition of the spiked-in transcripts for the HGU95 series is adopted.

In Tables 1 and 2, it was apparent that the best estimates of α (in the MSE sense) were the mean and the seventy-fifth percentile of the intensities greater than $\hat{\mu}$. Recall that $\hat{\mu}$ is given by $(\mu_1 + \mu_2)/2$, where μ_1 is the estimate of μ given by RMA and μ_2 is the estimate of μ from the one-step correction formula. $\sqrt{2}\hat{\sigma}$ is the standard deviation of the values less than $\hat{\mu}$. In the comparisons that follow, the results are shown for M3 only. This method found estimates of μ and σ using the one-step correction method. $\hat{\alpha}$ is given by either the mean (RMA Mean) or the seventy-fifth percentile (RMA 75) of the values greater than $\hat{\mu}$. For all three methods (RMA, RMA Mean, and RMA 75), the intensities were normalized using quantile normalization and summarized using median polish. The code for background correction using RMA-Mean and RMA-75 is given as supplemental material. Alternatively, the code is available for download at <http://faculty.smu.edu/mmcgee>.

Receiver operating characteristic curves (ROC curves) have been used in several articles for comparing the performance of methods in obtaining the “correct” answer [5,21]. Here, ROC curves are used to compare methods of detecting differentially expressed genes based on fold-change. Figure 2 shows ROC

curves for RMA Mean and RMA 75 using the HGU95 data set. For comparison, the results using the current implementation of RMA are also plotted. Note that our estimates do not perform any better or worse than does RMA.

However, there is a much larger difference among the methods in Figure 3, which shows a ROC curve calculated using the HGU133 data. Recall that these data have larger background population and more spiked transcripts. Here, both RMA Mean and RMA 75 substantially outperform the current implementation of RMA. This comparison was made using experiments one and two, whose concentrations do not differ that much; therefore, detecting true differences is difficult. On experiments where the differences in concentrations between spike-ins were much greater, the methods performed equally well (ROC curves for experiment 1 paired with all other experiments are given in supplemental materials). However, the current implementation of RMA sometimes needed a larger false positive rate to obtain the same true positive rate as with RMA Mean and RMA 75.

In addition generating ROC curves for pairs of experiments, our methods were applied to the spike-in datasets using Affycomp [8]. Affycomp uses fourteen of the fifty-nine HGU95 chips to make its comparisons. The fourteen chips are chosen so that probeset interactions were balanced. Affycomp also uses fourteen of the HGU133 chips, representing one experiment (<http://www.bioconductor.org/repository/devel/vignette/affycomp.pdf>). It is not clear which experiment is used. However, since only one experiment is chosen, all measures for the HGU133 data given by Affycomp are computed without replicates.

Affycomp returns many measurements and graphics. Among these is a ranking for overall signal to ratio assessment. RMA 75 was given a rank of 1, while

RMA Mean was given a rank of 2, implying that the performance of both methods was better overall than that of RMA. Ranks for the overall assessment for the HGU95 Spike-In data were 21 and 22, respectively. At the time RMA Mean and RMA 75 were submitted to the Affycomp contest, there were sixty-eight methods for which assessment results were available.

5 Discussion and Conclusion

RMA Mean and RMA 75 performed much better than RMA on the HGU-133 Spike-In data, but comparably on the HGU-95 data. The discrepancy can be explained on at least three fronts.

First, the HGU133 spike-in experiments are newer and use more reliable technology and methods than the HGU95 spike-in data. The discrepancy about additional spiked-in genes, mentioned in the introduction, calls into question the use of the older data as a benchmark for comparison of method performance.

Second, the convolution model may be incorrect. Separate analyses (manuscript in preparation) have revealed that the underlying background noise may, in fact, be normal, but the distribution of the signal has heavier tails than an exponential distribution. There is anecdotal evidence that RMA does not perform well on real data sets. If this assertion is true, it may be so because the convolution method is not valid. Other researchers have used the lognormal distribution [13] and the Laplace distribution [16] to model the overall signal from two-channel microarray data. Neither of these distributions has been investigated for Affymetrix chips, with or without the convolution model.

Third, not all methods can be expected to perform uniformly well across all data sets. It may be the best background correction method is platform, and possibly experiment, dependent. For example, while RMA performs well for both the HGU95 and HGU133 spike-in datasets, it has been shown that MAS 5.0 outperforms RMA on a dataset of 3860 RNA species using a DrosGenome1 GeneChip [7]. The optimal background correction algorithm may depend on the number of the spiked-in probes, the design of the spike-in experiment, and the GeneChip platform. This conjecture requires further investigation, and lends credence to the call for more publicly available spike-in datasets on a variety of platforms in order to thoroughly validate new and existing methods [4].

This paper explains the method of parameter estimation for RMA, a popular method for analysis of microarray data. It is shown that the original parameter estimates used for the RMA background correction do not perform well in simulation studies. Better results were obtained using other estimates. In addition, implementing RMA with the better parameter estimates gave better downstream results than the current implementation of RMA in the Spike-In data.

References

- [1] Affymetrix, Inc, 2001. Statistical Algorithms Reference. Data Analysis Fundamentals Technical Manual, Chapter 5.
- [2] Affymetrix, Inc, 2002. Statistical Algorithms Description Document.
- [3] Affymetrix Technical Note: Design and Performance of the GeneChip Human Genome U133 Plus 2.0 and Human Genome U133A Plus 2.0 Arrays, 2003.

- [4] Allison, D.B., Cui X., Page G.P., and Sabripour M., 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7, 55-65.
- [5] Bolstad, B.M. Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. Dissertation (Dept. of Statistics, University of California, Berkeley, 2004).
- [6] Bolstad B.M., Irizarry R.A., Astrand M., and Speed T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on the variance and bias. *Bioinformatics*, 19, 185-193.
- [7] Choe S.E., Boutros M., Michelson A.M., Church G.M., and Halfon M.S., 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6, R16.
- [8] Cope L.M., Irizarry R.A., Jaffee H.A., Wu Z., and Speed T.P., 2004. A benchmark for Affymetrix GeneChip Expression measures. *Bioinformatics*, 20, 323-331
- [9] Gentleman R.C., Carey V.J., Bates D.M., Bolstad B.M., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry RA., Cheng L., Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J.W.H., and Zhang J., 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.
- [10] Ihaka R and Gentleman RC, 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314
- [11] Irizarry R.A., Bolstad B.M., Collin F., Cope L.M., Hobbs B., and Speed T.P., 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31, 4, e15.

- [12] Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., and Speed T.P., 2003. Exploration, Normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249-264.
- [13] Konishi T., 2004. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics*, 5, 5.
- [14] Li, C. and Wong. H.W., 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98, 31-36.
- [15] Li, C. and Wong. H.W., 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 8, research0032.1-0032.11.
- [16] Purdom E. and Holmes S.P., 2005. Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4, article 16.
- [17] Simon, Richard, 2003. *Design and Analysis of DNA Microarray Investigations*. Springer-Verlag, New York, NY.
- [18] Speed, Terry (Ed), 2003. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, Boca Raton, FL.
- [19] Tukey J.W., 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 366-369.
- [20] Wolfinger, R. and Chu, T., 2003. Who are those strangers in the Latin Square? In: K.F. Johnson and S.M. Lin (Eds.) *Methods of Microarray Data Analysis III*. Springer-Verlag, New York.
- [21] Zhou L. and Rocke D.M., 2005. An expression index for Affymetrix GeneChips based on the generalized logarithm, *Bioinformatics*, 21, 3983-3989.

Captions for Tables and Figures

Table 1: MSEs of parameter estimates for $\mu = 30$, $\sigma = 5, 10$, and $\alpha = 250, 500, 1000$. Estimates were calculated $\hat{\mu} = (\mu_1 + \mu_2)/2$, where μ_1 , σ_1 , and α_1 are estimated using the Bioconductor implementation of RMA, and μ_2 given by plugging in μ_1 , σ_1 , and α_1 into a one-step correction (3). Once $\hat{\mu}$ is calculated, $\hat{\sigma}$ is given by the standard deviation of the intensities less than $\hat{\mu}$. $\hat{\alpha}$ is given by either the mean, median, 75th percentile or 99.95th percentile of the intensities values larger than the overall mode.

Table 2: MSEs of parameter estimates for $\mu = 50$, $\sigma = 5, 10$, and $\alpha = 250, 500, 1000$. Parameter estimates are given in the same way as in Table 1.

Figure 1: Density Estimates of Log Base 2 PM Intensities of one replicate from each of the 14 experiments in the HG-U133 Spike-In Data series. A convolution of a normal distribution and an exponential distribution seem reasonable for these data.

Figure 2: ROC curves of the original implementation of RMA versus two competitors: RMA where $\hat{\mu}$ and $\hat{\sigma}$ are estimated using a one-step correction, and $\hat{\alpha}$ is given by either the mean (RMA Mean) or the 75th percentile (RMA 75) of the intensities greater than $\hat{\mu}$. This comparison is done using the HG-U95 spike-in data.

Figure 3: ROC curves of RMA versus RMA Mean and RMA 75 for the HG-U133 spike-in data.

Estimation Method					
Combination	RMA	Mean	Median	75 th	99.95 th
$\mu = 30$	233	2.62	2.62	2.62	2.68
$\sigma = 5$	92	1.35	1.35	1.51	1.39
$\alpha = 250$	45135	4.70	10.2	2.56	22.1
$\mu = 30$	590	8.90	8.90	8.90	9.30
$\sigma = 10$	215	5.32	5.32	6.20	5.48
$\alpha = 250$	45028	11.85	26.59	6.51	21.27
$\mu = 30$	436	4.98	4.98	4.98	5.05
$\sigma = 5$	192	1.59	1.59	1.77	1.63
$\alpha = 500$	180781	8.07	17.2	7.02	76.72
$\mu = 30$	930	10.79	10.79	10.79	11.03
$\sigma = 10$	363	4.52	4.52	5.16	4.64
$\alpha = 500$	180508	19.7	41.0	10.61	87.29
$\mu = 30$	1094	11.61	11.62	11.61	11.86
$\sigma = 5$	552	3.38	3.38	3.82	3.48
$\alpha = 1000$	723457	16.97	33.50	20.53	333.11
$\mu = 30$	1712	18.40	18.43	18.64	18.64
$\sigma = 10$	744	5.05	5.06	5.70	5.16
$\alpha = 1000$	722989	31.15	66.89	23.70	309.1

Table 1

Estimation Method					
Combination	RMA	Mean	Median	75 th	99.95 th
$\mu = 50$	230	2.47	2.47	2.47	2.53
$\sigma = 5$	90	1.29	1.29	1.44	1.32
$\alpha = 250$	45218	4.73	10.41	2.36	21.39
$\mu = 50$	576	8.511	8.512	8.511	8.922
$\sigma = 10$	211	6.091	6.091	6.990	6.260
$\alpha = 500$	45016	12.09	26.67	6.310	20.75
$\mu = 50$	447	5.55	5.55	5.55	5.63
$\sigma = 5$	197	1.802	1.802	2.008	1.841
$\alpha = 1000$	180790	8.212	17.37	6.717	81.89
$\mu = 50$	925	11.26	11.28	11.26	11.53
$\sigma = 10$	364	5.40	5.41	6.06	5.52
$\alpha = 250$	180554	19.32	41.15	10.58	91.16
$\mu = 50$	1074	12.29	12.30	12.29	12.60
$\sigma = 10$	541	3.64	3.64	4.11	3.79
$\alpha = 500$	723356	16.83	34.08	19.30	339.8
$\mu = 50$	1731	20.26	20.39	20.36	20.86
$\sigma = 10$	761	6.575	6.578	7.786	6.773
$\alpha = 1000$	723128	37.92	80.17 ₂₀	30.09	372.67

Table 2

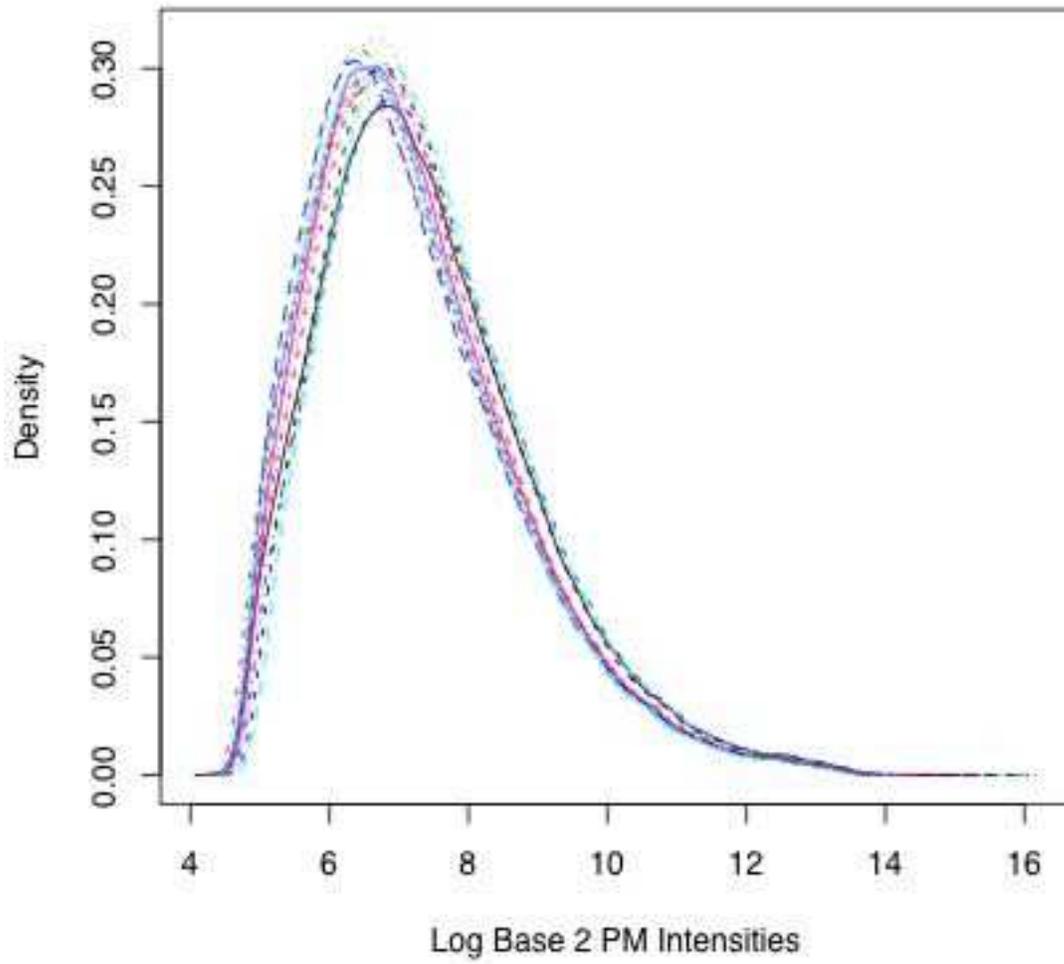


Fig. 1.

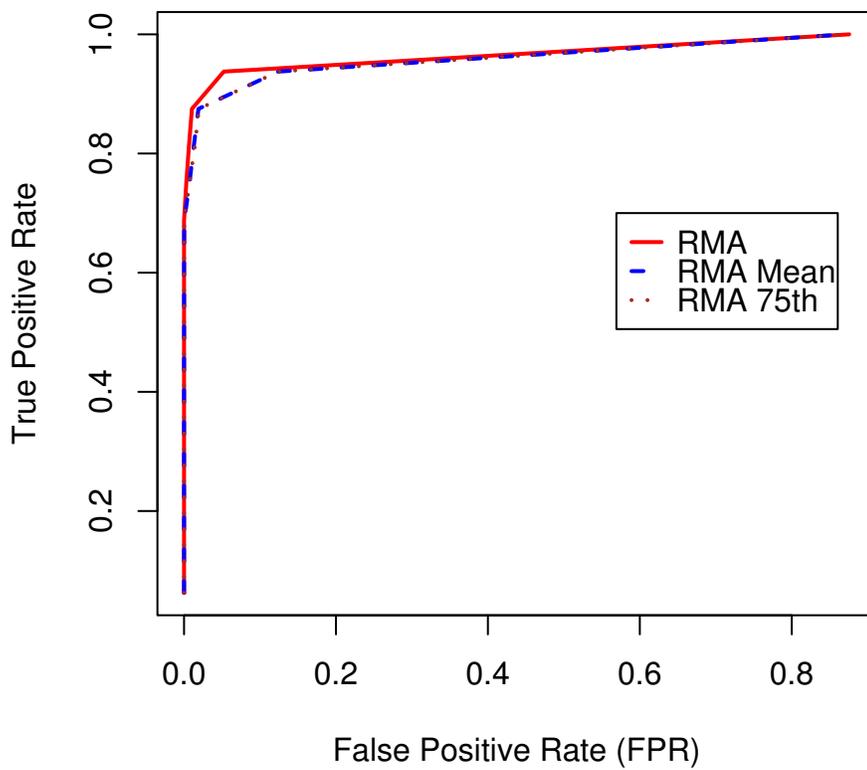


Fig. 2.

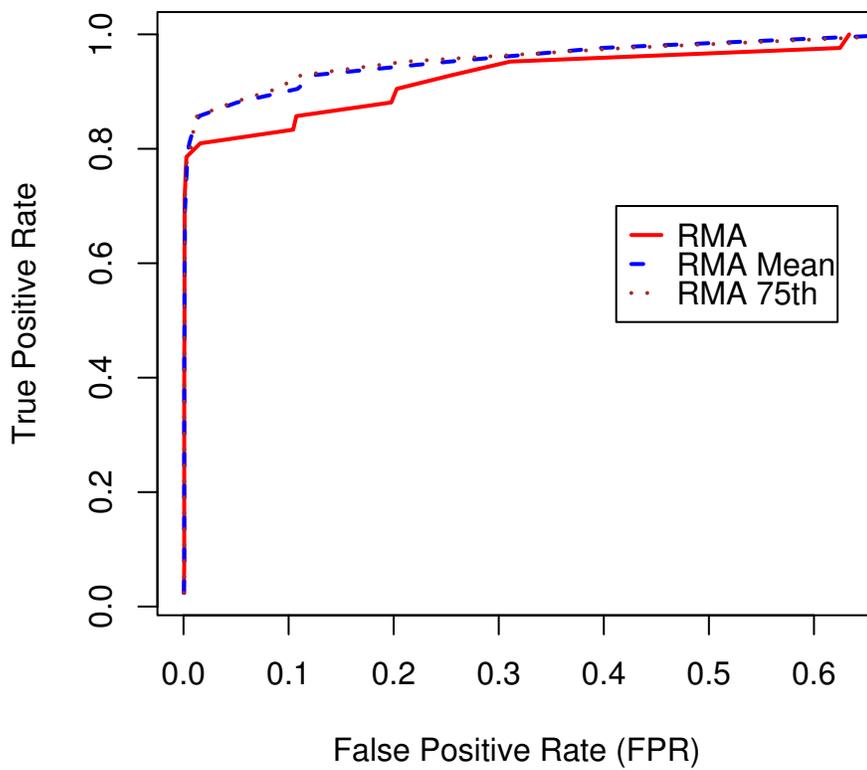


Fig. 3.