Johns Hopkins University
Johns Hopkins University, Dept. of Biostatistics Working Papers

Year 2004 *Paper* 1

A Model Based Background Adjustment for Oligonucleotide Expression Arrays

Zhijin Wu* Rafael A. Irizarry† Robert Gentleman‡
Francisco Martinez Murillo** Forrest Spencer††

*Johns Hopkins Bloomberg School of Public Health, zwu@jhsph.edu
†Johns Hopkins Bloomberg School of Public Health, rafa@jhu.edu
‡Dana-Farber Cancer Institute, rgentlem@hsph.harvard.edu
**Johns Hopkins Medical Institute, fmurill1@jhmi.edu
††Johns Hopkins Medical Institute, fspencer@jhmi.edu

A Model Based Background Adjustment for Oligonucleotide Expression Arrays

Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez Murillo, and Forrest Spencer

Abstract

High density oligonucleotide expression arrays are widely used in many areas of biomedical research. Affymetrix GeneChip arrays are the most popular. In the Affymetrix system, a fair amount of further pre-processing and data reduction occurs following the image processing step. Statistical procedures developed by academic groups have been successful at improving the default algorithms provided by the Affymetrix system. In this paper we present a solution to one of the pre-processing steps, background adjustment, based on a formal statistical framework. Our solution greatly improves the performance of the technology in various practical applications.

Affymetrix GeneChip arrays use short oligonucleotides to probe for genes in an RNA sample. Typically each gene will be represented by 11-20 pairs of oligonucleotide probes. The first component of these pairs is referred to as a perfect match probe and is designed to hybridize only with transcripts from the intended gene (specific hybridization). However, hybridization by other sequences (non-specific hybridization) is unavoidable. Furthermore, hybridization strengths are measured by a scanner that introduces optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. One approach to adjusting is to pair each perfect match probe with a mismatch probe that is designed with the intention of measuring non-specific hybridization. The default adjustment, provided as part of the Affymetrix system, is based on the difference between perfect match and mismatch probe intensities. We have found that this approach can be improved via the use of estimators derived from a statistical model that use probe sequence information. The model is based on simple hybridization theory from molecular biology and experiments specifically designed

to help develop it.

A final step in the pre-processing of these arrays is to combine the 11-20 probe pair intensities, after background adjustment and normalization, for a given gene to define a measure of expression that represents the amount of the corresponding mRNA species. In this paper we illustrate the practical consequences of not adjusting appropriately for the presence of nonspecific hybridization and provide a solution based on our background adjustment procedure. Software that computes our adjustment is available as part of the Bioconductor project (http://www.bioconductor.

A Model Based Background Adjustment for Oligonucleotide Expression Arrays

Zhijin Wu, Rafael A. Irizarry, Robert Gentleman,

Francisco Martinez-Murillo, and Forrest Spencer *

May 21, 2004

Abstract

High density oligonucleotide expression arrays are widely used in many areas of biomedical research. Affymetrix GeneChip arrays are the most popular. In the Affymetrix system, a fair amount of further pre-processing and data reduction occurs following the image processing

*Zhijin Wu is graduate student and Rafael A. Irizarry is Associate Professor of Biostatistics (E-mail: rafa@jhu.edu), Francisco Martinez Murillo is manager of JHMI Microarray Core, Forrest Spencer is Associate Professor of Medicine and Molecular Biology and Genetics and director of JHMI microarray Core, Johns Hopkins University, Baltimore, Maryland 21205. Robert Gentleman is Associate Professor of Biostatistical Science, Dana–Farber Cancer Institute, 44 Binney Street Boston, MA 02115. The authors would like to thank Terry Speed, Ben Bolstad, and Earl Hubbell for insightful discussions that helped develop our ideas. We would also like to thank Wolfgang Huber and Tom Louis for some helpful comments. The work of Rafael Irizarry is partially funded by the Hopkins PGA (http://www.hopkins-genomics.org) Administrative/Bioinformatics Component (P01 HL 66583). The work of Zhijin Wu is partially funded by the Johnson and Johnson Research Foundation.

step. Statistical procedures developed by academic groups have been successful at improving the default algorithms provided by the Affymetrix system. In this paper we present a solution to one of the pre-processing steps, background adjustment, based on a formal statistical framework. Our solution greatly improves the performance of the technology in various practical applications.

These arrays use short oligonucleotides to probe for genes in an RNA sample. Typically each gene will be represented by 11-20 pairs of oligonucleotide probes. The first component of these pairs is referred to as a *perfect match* probe and is designed to hybridize only with transcripts from the intended gene (specific hybridization). However, hybridization by other sequences (non-specific hybridization) is unavoidable. Furthermore, hybridization strengths are measured by a scanner that introduces optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. We have found that the default adhoc adjustment, provided as part of the Affymetrix system, can be improved via the use of estimators derived from a statistical model that uses probe sequence information.

A final step in pre-processing is to summarize the probe-level data for each gene to define a measure of expression that represents the amount of the corresponding mRNA species. In this paper we illustrate the practical consequences of not adjusting appropriately for the presence of non-specific hybridization and provide a solution based on our background adjustment procedure. Software that computes our adjustment is available as part of the Bioconductor project (`http://www.bioconductor.org`).

1 Introduction

Affymetrix GeneChip arrays are the most popular high density oligonucleotide expression arrays and are used by thousands of researchers worldwide. To probe genes, oligonucleotides of length 25 base pairs are used. Typically, the mRNA sequence of a gene is represented by a *probe set* composed of 11-20 *probe pairs*. Each probe pair is composed of a perfect match (PM) probe, a 25-base DNA copy of a section of the mRNA sequence of interest, and a mismatch (MM) probe, that is created by changing the middle (13th) base of the PM probe with the intention of measuring non-specific binding (NSB). See the Affymetrix Microarray Suite User Guide for details.

After RNA samples are prepared, labeled and hybridized with arrays, these are scanned and images are produced and processed to obtain an intensity value for each probe. These intensities represent the amount of hybridization for each oligonucleotide probe. However, part of the hybridization is non-specific and the intensities are affected by optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. In this paper, we refer to the part of the observed intensity due to optical noise and non-specific hybridization as *background noise*. The default background noise adjustment, provided as part of the Affymetrix system, is based on the difference $PM - MM$.

A final step is to combine the 11-20 probe pair intensities, after background adjustment and normalization, for a given gene to define a measure of expression. Various alternatives algorithms motivated by statistical models have been proposed that outperform the default algorithm in many applications, see for example Li and Wong (2001) . Irizarry et al. (2003a) find that the $PM - MM$ transformation results in expression estimates with exaggerated variance. They propose a background adjustment step that ignores the MM intensities. This approach sacrifices some accuracy

for large gains in precision. The resulting algorithm, the robust multi-array analysis (RMA), has become a popular alternative to the default algorithm provided by Affymetrix. In this paper we demonstrate that the loss of accuracy mentioned above is due to inappropriate adjustment for the presence of non-specific hybridization.

Data from our own experiments, molecular biology theory, and publicly available data sets were used to develop a statistical model that describes background noise. An empirical Bayes procedure motivated by this model results in a background adjustment procedure that improves existing approaches. In Section 2 we describe why background adjustment is important and present the data we generated to motivate the model. In Section 3 we introduce our model of the stochastic structure of the oligonucleotide array data. In section 4 we describe the practical applications. In Section 5 we show how our method provides an improvement to users of the Affymetrix GeneChip technology. In Section 6 we briefly describe the software we have developed to implement our methodology. Finally, in Section 7 we discuss our findings.

2 Motivation

The Affymetrix spike-in study[1] is a subset of the data used to develop and validate the MAS 5.0 expression measure algorithm, Affymetrix's current default (Affymetrix, 2002). For this experiment, human cRNA fragments matching 16 probe-sets on one of the Affymetrix human chips, the HGU95A GeneChip, were added to a hybridization mixture at concentrations ranging from 0 to 1024 picoMolar in a design similar to a Latin square. Apart from the spiked-in probe-sets, the same RNA mixture was hybridized to 59 arrays. Because we know the spike-in concentrations, it

[1] Available from: `http://www.affymetrix.com/analysis/download_center2.affx`

is possible to identify statistical features of the data for which the expected outcome is known in advance. This experiment is described in detail by, for example, Irizarry et al. (2003b).

Figure 1a shows a typical density estimate of *PM* intensities obtained from 14 normalized arrays from the Affymetrix spike-in experiment. For these 14 arrays the spike-in concentrations, ranging from 0 to 1024, appear exactly 16 times and exactly once for each spike-in probeset. The arrays were normalized using quantile normalization (Bolstad et al., 2003) and then the geometric averages of *PM* intensities for each spike-in concentration were computed. These averages are shown with arrows in Figure 1a and plotted against their intended or *nominal* concentrations in Figure 1b. The presence of background noise is clear from the fact that the minimum *PM* intensity is not 0 and that the geometric mean of the probesets with no spike-in is around 200 units.

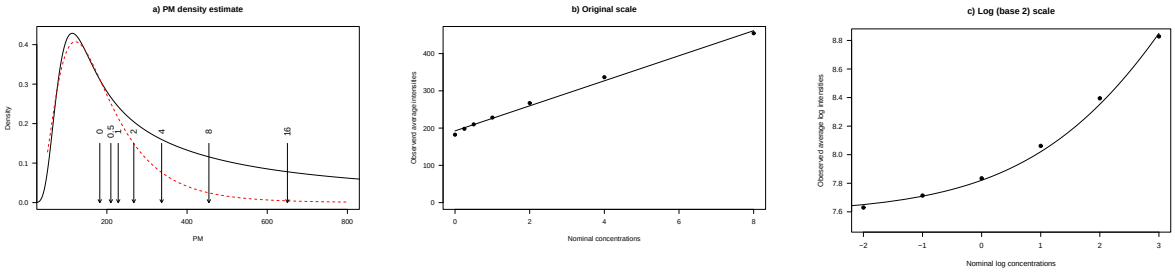

Figure 1: a) The solid lines is a density estimator of quantile normalized *PM* intensities. We only show the (55,800) range in the x-axis for illustrative purposes. The *PM* values are as high as 20000. The dotted lines are from a log-normal distribution representing background noise. The arrows denote the geometric means of the *PM* intensities of probes with the same nominal concentrations. The concentrations (in picoMolar) are shown at the top of the arrows. b) Geometric mean of observed concentrations plotted against nominal concentrations. The mean was computed for the concentration groups with 0, 1/4, 1/2, 1, 2, 4, and 8 picoMolar. The solid line represent a regression line fitted to the data. c) Same as Figure a) but the x and y axis are now in the log-scale.

By using the log-scale transformation before analyzing microarray data, investigators have, implicitly or explicitly, assumed a multiplicative measurement error model (Dudoit et al., 2002; Newton et al., 2001; Kerr et al., 2000; Wolfinger et al., 2001). The fact, seen in Figure 1, that observed

intensity increases linearly with concentration in the original scale but not in the log-scale suggests that background noise is additive with non-zero mean. Durbin et al. (2002), Huber et al. (2002), Cui, Kerr, and Churchill (2003), and Irizarry et al. (2003a) have proposed additive-background-multiplicative-measurement-error (ABME) models for intensities read from microarray scanners. Figures 1 supports this view.

2.1 Previous work

Affymetrix's first attempt at an expression measure (MAS 4.0) used the transformation $PM - MM$ to adjust for non-specific binding and background noise. In general, $MM \geq PM$ for about 1/3 of the probes on any given array (Irizarry et al., 2003a) which results in negative adjusted intensity values. This results in two obvious problem: 1) we can not use the log transformation to account for the multiplicative measurement error and 2) expression measures based on averages of the adjusted intensities are negative for about 5% of the probesets. This is in part due to the fact that the MM are somewhat sensitive to targets probed by their PM counterpart, i.e. they detect specific signal (Irizarry et al., 2003a). In the most recent version of their software, MAS 5.0, Affymetrix defines $PM - MM^*$, with MM^* a "tweaked" version of MM used to avoid adjusted values less than or equal to 0. A robust average of the log transformed adjusted PM defines the MAS 5.0 expression measure (Affymetrix, 2002).

Li and Wong (2001) were the first to propose model-based expression measures. They observed a very strong probe effect in that $PM - MM$ values, the need for non-linear normalization, and the advantages of using multi-array summaries for detection and removal of outliers. These three observation were used in the development of RMA and other popular expression measures.

Although, the expression measure presented in this paper make use of Li and Wong's discoveries, we find the ABME model is a more appropriate stochastic assumption than the one made by Li and Wong (2001). We base our approach on the model proposed by, for example, Durbin et al. (2002).

Using replicate array data, Irizarry et al. (2003b) showed that the across-replicate MAS 5.0 log expression measure standard deviation was an order of magnitude higher for probe-sets with low intensities than for probe-sets with large intensities. Under our proposed model, presented in Section 3, one can show that the variance of $\log(PM - MM^*)$ is roughly proportional to the reciprocal of the true amount of RNA. This is consistent with empirical results. Non-specific hybridization is likely to make $PM - MM^*$ a less biased estimate of a quantity proportional to the true amount of RNA than PM, but $\log(PM - MM^*)$ will generally have larger variance than $\log(PM)$, especially when the true amount of RNA is small.

The background adjustment step in RMA ignores the MM. An additive background plus specific signal with multiplicative error is proposed. They assume the specific signal follows an exponential distribution and that the background noise follows a normal distribution. After a data-driven ad-hoc procedure is used to estimate model parameters the conditional expectation of the specific signal given the observed intensity results in a closed form transformation which provides a practical solution to background adjustment. Irizarry et al. (2003a, 2003b) demonstrate, using data from spike-in and dilution experiments, that RMA outperforms other popular expression measures, including MAS 5.0, in various practical tasks. However, as demonstrated in Section 5, by only performing a global background adjustment, RMA does not adjust well for non-specific binding, resulting in attenuated correlations between nominal and observed concentrations.

2.2 Practical Considerations

One of the most popular applications of microarray technology is the identification of genes that are differentially expressed. A common parameter of interest is fold-change. Notice in Figure 1c that, on average, the observed fold-change between probes spiked-in at 0.25 and 0.50 picoMolar is approximately 1 when it should be 2. Comparisons between the higher concentrations, 4 and 8 picoMolar, for example, have observed fold changes closer to what is expected. To see how this is related to background, say that true expression values in two samples being compared are μ_1 and μ_2 picoMolar. Ideally we should observe a fold change of μ_1/μ_2. In practice, we observe intensities $PM_1 \approx k\mu_1 + B_1$ and $PM_2 \approx k\mu_2 + B_2$, with B_1 and B_2 representing background noise and k a constant to account for the change in units. Because the B's are strictly positive and, after normalization, roughly the same, the observed fold changes of the non-background-corrected values are smaller than expected. This bias will be more pronounced for smaller values of μ_1 and μ_2, as observed in Figure 1c.

Receiver operator characteristic curves presented in Irizarry et al. (2003b) demonstrate that RMA outperforms MAS 5.0 when using large absolute value of observed log fold change to define differential expression. However, as pointed out by Irizarry et al. (2003b), fold-change estimates obtained with RMA are less accurate (attenuated) than those obtained with MAS 5.0. In Section 5 demonstrate that our proposed background adjustment results in an expression almost as precise as RMA and as accurate as MAS 5.0.

2.3 Data

There have been attempts at using sequence information to predict non-specific binding (Zhang et al., 2003). These are based on deterministic models. To build appropriate stochastic models to describe how mismatch probes measure non-specific binding we have carried out three experiments. First, we obtained RNA from human embryonic kidney derived cells to create a control sample. We also used genomic DNA from yeast to create a hybridization mixture with DNA molecules non-specific to transcripts synthesized on human arrays. The processing of the sample was done following Affymetrix specifications with specific variations depending on the sample content. Three samples were prepared to study different aspects of background noise: 1) **Unlabeled** - In this sample RNA control from human embryonic kidney derived cells was not labeled, but the cocktail was created as described above, using 15 μg of amplified cRNA. Because the RNA was not labeled, the observed intensities for this hybridization will represent optical noise in the presence of biological sample. 2) **No RNA** - In this sample hybridization cocktail was hybridized with no RNA. As for the previous sample the observed intensities for this hybridization will represent optical noise in the absence of biological sample. 3) **NSB** - Yeast control RNA was hybridized to an array probing for human genes. This hybridization will represent the full component of the noise, non-specific binding (NSB) and optical noise. Data from the spike-in experiment, where arrays were hybridized to labeled human RNA as described in the introduction, will represent typical experiments where optical, NSB and specific hybridization components are all present.

Figure 2a shows a quantile-quantile plot comparing the log intensities, read from the array hybridized to the unlabeled sample, to a normal distribution. The no RNA sample data looks almost identical (data not shown). Notice that the distribution of optical noise is well approximated by a

log normal distribution. Notice also that the mean, on the log-scale, translates to an intensity of about 32 and that on the log-scale the variance is only about 0.1.

Figure 2b shows a scatter plot of the log intensities of PM, MM pairs in the NSB data. The PM and MM have been adjusted for optical noise as described in Section 4.2. This plot suggests that probe-pair NSB noise follow a bivariate log-normal distribution. The density of PM log intensities stratified by values of MM shown in Figure 2c corroborate this assertion. Notice that the log-scale variance of the $\log(PM)$ in each strata is roughly constant (it is actually increasing slightly) and about 1, much larger than for optical noise. In Section 2.4 we describe how probe sequence information can be used to describe some of the variation seen in NSB.

Figure 2d shows a log frequency versus $\log(\text{rank})$ plot for the intensities read from an array hybridized with labeled human RNA. This figure shows that when specific binding occurs the distribution has a "fat-tail". To some extent the data appear to follow a power law.

2.4 Probe sequence and affinity

The physical system producing probe intensities is a complicated one. Theoretical predictions of non-specific bindings based solely on sequence are unlikely to be as successful as empirical ones. Affymetrix technology implements an empirical approach by including MM probes and using the observed intensities to adjust for background noise. However, as mentioned above, empirical evidence suggests that simply subtracting MMs is a sub-optimal strategy. In this paper we develop a method and estimation procedures useful for background adjustment. In this Section we describe how probe sequence information can be used to motivate a useful model.

In traditional hybridizations (such as Southern blots) performed at moderate or low stringency,

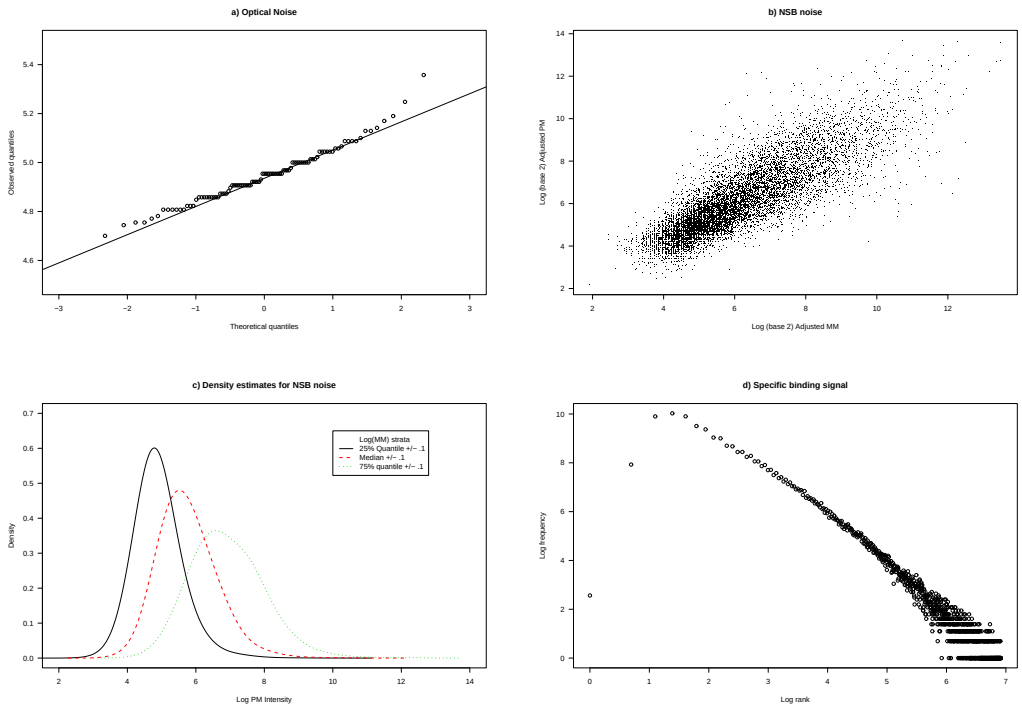

Figure 2: a) Normal quantile-quantile plot of observed log intensities from an array hybridized to the mixture with no label. b) Optical noise adjusted $\log_2(PM)$ intensity plotted against optical noise adjusted $\log_2(MM)$ intensity. c) Densities of optical noise adjusted $\log_2(PM)$ intensity within selected strata of $\log_2(MM)$ intensity. d) Log-frequency versus $\log(\text{rank})$ plot for the intensities from a typical array hybridized to human RNA.

non-specific hybridization background is often observed to be due to partial nucleic acid homology between two single strands with imperfect complementarity. This problem is closely related to the base composition of the nucleic acid molecules. G/C in sequence lead to stronger hybridization because each G-C pair forms three hydrogen bonds whereas each A-T pair forms two. On the other hand, bases U and C are labeled in amplified RNA in the above described labeling protocol, which appears to impede binding (Naef and Magnasco, 2003)

Naef and Magnasco (2003) propose a solution useful for predicting specific hybridization effects with base composition of the probes. Probe *affinity* is modeled as a sum of position-dependent

base effects:

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k=j} \text{ with } \mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l, \quad (1)$$

where $k = 1, \ldots, 25$ indicates the position along the probe, j indicates the base letter, b_k represents the base at position k, $1_{b_k=j}$ is an indicator function that is 1 when the k-th base is of type j and 0 otherwise, and $\mu_{j,k}$ represents the contribution to affinity of base j in position k. For fixed j, the effect $\mu_{j,k}$ is assumed to be a polynomial of degree 3. The model is fitted to log intensities from many arrays using least squares (Naef and Magnasco, 2003).

We adapt this idea to help describe the NSB component. We fit (1) to our NSB experiment log intensity data using a spline with 5 degrees of freedom instead of a polynomial of degree 3. The least squares estimates $\hat{\mu}_{j,k}$ are shown in Figure 3. This Figure is similar to Figure 3 in Naef and Magnasco (2003). In Section 3 we use these affinity estimates to describe NSB noise.

Zhang et al (Zhang et al., 2003) propose using a *positional-dependent-nearest-neighbor* (PDNN) model which is based on hybridization theory. This model takes into account interactions between bases that are physically close. However, Naef and Magnasco (2003) demonstrate that these interactions do not add much predictive power for specific signal probe effects. Similar results for prediction of NSB are found in Wu and Irizarry (2004). Figure 4 shows background adjusted $\log_2(PM)$ against α for the NSB data. Notice the affinities predict NSB quite well. Almost as well as the *MM* intensities. The advantage of the affinities over the MM is that they will not detect signal since they are pre-computed numbers.

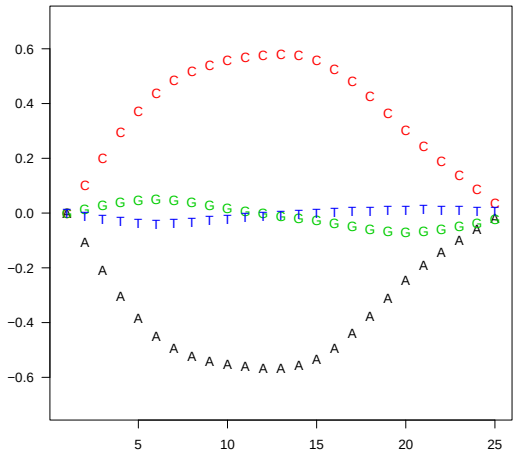

Figure 3: The effect of base A in position k, $\mu_{A,k}$, is plotted against k. Similarly for the other three bases.

3 Background Model

In this section we propose a statistical model that motivates useful background adjustments. For any particular probe-pair we assume that:

$$\begin{aligned} PM &= O_{PM} + N_{PM} + S \\ MM &= O_{MM} + N_{MM} + \phi S. \end{aligned} \tag{2}$$

Here O represents optical noise, N represents NSB noise and S is a quantity proportional to RNA expression (the quantity of interest). The parameter $0 < \phi < 1$ accounts for the fact that for some probe-pairs the MM detects signal. We assume O follows a log-normal distribution and that $\log(N_{PM})$ and $\log(N_{MM})$ follow a bivariate-normal distribution with means of μ_{PM} and μ_{MM} and the variance $\mathrm{var}[\log(N_{PM}] = \mathrm{var}[\log(N_{MM})] \equiv \sigma^2$ and correlation ρ constant across probes. We assume $\mu_P M \equiv h(\alpha_{PM})$ and $\mu_M M \equiv h(\alpha_{MM})$, with h a smooth (almost linear) function and the αs defined by (1). Because we do not expect NSB to be affected by optics we assume O and N are

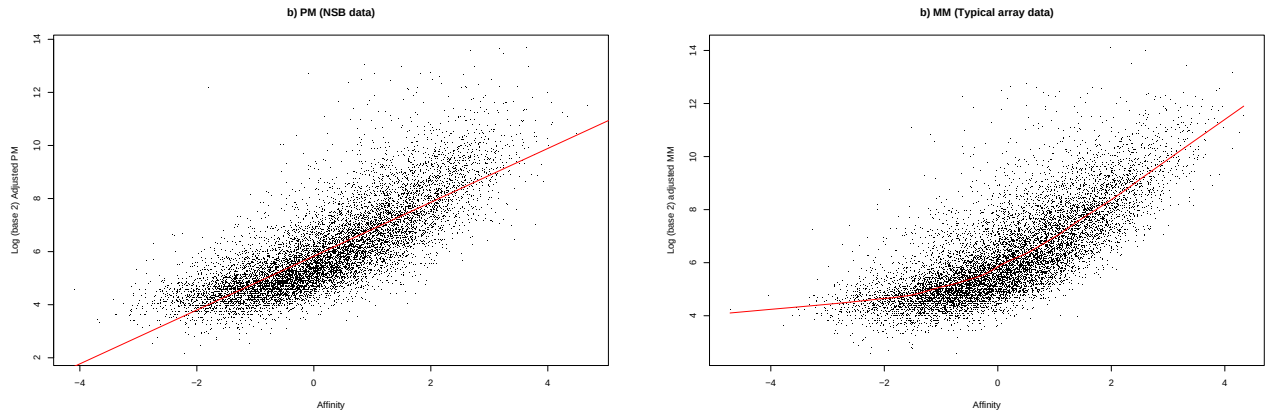

Figure 4: a) Optical noise adjusted $\log_2(PM)$ intensity plotted against affinity for NSB data. The solid line is a linear fit. b) Optical noise adjusted $\log_2(MM)$ intensity plotted against affinity for data from a typical array hybridized to human RNA. The solid line is a loess fit.

independent. Our exploratory plots support these assumptions.

The parameters μ_{PM}, μ_{MM}, ρ, and σ^2 can be estimated from data. The large amount of data results in very precise estimate of these parameters. A background adjustment procedure can then be formalized as the statistical problem of predicting S given that we observed PM and MM and assuming we know h, ρ, σ^2 and ϕ.

4 Application

4.1 Background Adjustment

In this Section we define a new version of RMA that uses background adjustment procedures motivated by (2) to improve accuracy. The normalization and summarization steps are kept the same. We use the acronym GC-RMA to denote the resulting expression measure. To facilitate the calculations of useful adjustments we make two assumptions that are worth mentioning: The

first assumption is that ϕ is 0. Although we know $\phi > 0$, in Section 5 we demonstrate that this assumption does not have a significant impact on bottom-line results. The second assumption is that O is an array-dependent constant.

4.1.1 Maximum Likelihood Estimate

Under the above described assumptions, the maximum likelihood estimate (MLE) of S can be easily shown to be $PM - O - \hat{N}_{PM}$ if $PM - \hat{N}_{PM} > m$, m otherwise. with $\hat{N}_{PM} \equiv \exp\{\rho \log(MM - O) + \mu_{PM} - \rho\mu_{MM} - (1-\rho^2)\sigma^2\}$ and m the minimum value we allow for S, typically $m = 0$. To get a numeric value of $\hat{N}$ we can use plug-in estimators of ρ, h and σ^2 described in Section 4.2.

This adjustment provides an intuitive result. Instead of subtracting MM as an adjustment, we subtract a shrunken MM quantity that has been corrected for its affinity. The MM is shrunken toward the mean of the probes with similar affinity levels and the amount of shrinkage depends on the correlation between PM and MM. The shrinking is performed in the log scale which, in practice, helps protect against extremely large values of MM resulting from the multiplicative error structure.

Although the MLE provides transformations with desirable properties, it is not necessarily a practical solution if one is interested in a task such as detecting genes that are differentially expressed. As mentioned, one of the most important application of microarray technology is estimating fold-change of expression. In this case, instead of maximizing a marginal likelihood, a more appropriate estimate is obtained from minimizing the mean squared error (MSE):

$$\mathrm{E}[\{\log(\tilde{S}/S)\}^2 1_{\{S>0\}} \,|\, PM, MM]. \tag{3}$$

Not all genes are expressed in a typical cell. Thus, we expect $S = 0$ for some probesets. If one

is eventually going to consider fold-change, via log-ratios, as the quantity of practical interest, excluding cases for which $S = 0$ from the loss function becomes important.

There are many strategies that could be devised to define appropriate estimates based on minimizing (3). For example, we could increase the value of m to reduce the MSE. We have found that decreasing and increasing m yields estimates with better accuracy and precision respectively. An empirical-Bayes-type approach is described in the next section.

4.1.2 Empirical Bayes Estimate

A practical and simple solution is provided by an empirical-Bayes-type approach. In this case we treat S as a random variable which implies minimizing (3) is equivalent to minimizing

$$\mathrm{E}[\{\log(\tilde{S}/S)\}^2 | S > 0, PM, MM]. \tag{4}$$

Thus the solution is the posterior mean estimate: $\tilde{s} = \mathrm{E}[s | S > 0, PM, MM]$, with $s \equiv \log(S)$. Here, the random variable S represents a quantity proportional to the concentration of transcripts in the hybridization mixture that are compliments to a randomly chosen PM. Assuming S is independent from N and O and with a prior distribution of $s, f(s)$, the posterior mean can be written as $\tilde{s} = \int_{-\infty}^{a} w(PM,n) \log\{PM - O - \exp(n)\}\, dn \,/\, \int_{-\infty}^{a} w(PM,n)$ with $a = \log(PM - O - m)$, $w(PM,n)$ depending on the prior f and the quantities $\mu^* = \rho(\log(MM - \hat{O}) - \mu_{MM}) + \mu_{PM}$ and $\sigma^* = \sqrt{1-\rho^2}\sigma$. The value m is the smallest value of S with positive probability. Although not as simple as the MLE, $\tilde{s}$ provide an intuitive result: it is a weighted average, over all possible background values n, of $\log\{PM - O - \exp(n)\}$ with values near μ^* (a shrunken MM) receiving larger weights and the spread of the weights depending on σ^*. As expected, values of s that are more likely to occur get more weight (equations not shown).

Imposing a uniform distribution $U[m, \log(2^{16})]$ (2^{16} is the scanner maximum) is a convenient way of specifying a heavy-tailed prior distribution for S. In Section 5 using the uniform prior provides useful results. Just like for the MLE, we consider m to be a tuning parameter. However, users may decide to use priors appropriate for their application.

4.2 Plug-in Estimates for Model Parameters

Affymetrix arrays typically have over 100000 probe pair intensities. These can be used to get stable and precise estimates of the parameters ρ, μ_{PM}, μ_{MM}, and σ^2. Because the variance in optical noise is ignorable when compared to that of the NSB, we treat O as constant and estimate it with the minimum intensity observed each array. We define $\hat{O} = \min\{\min_j PM_j, \min_j MM_j\} - 1$ with j indexing all probe pairs. We subtract 1 to avoid negatives when adjusting by subtracting $\hat{O}$. For estimating μ_{PM},μ_{MM} and σ^2 we fit a loess curve to the $\log(MM - \hat{O})$ versus α_{MM} scatter plot, as seen in Figure 4b, to obtain an estimate $\hat{h}$ of the function h. We can then use this fit to estimate μ_{PM} and μ_{MM} with $\hat{\mu}_{PM} = \hat{h}(\alpha_{PM})$ and $\hat{\mu}_{MM} = \hat{h}(\alpha_{MM})$. The correlation between $\log(B_{PM})$ and $\log(B_{MM})$ should not change from array to array. For this reason we obtain an estimate from the NSB of $\hat{\rho} = 0.7$ data and use it throughout.

4.3 Unified Model

We have used model (2) to perform background adjustment. To obtain expression measures from probe level data two more steps are involved: normalization and summarization. A potential problem with this three-step approach is that the stability of the expression measure is obtained by combining information from all arrays in a given experiment and thus all measures of expression

for a given gene become correlated across experimental units. This makes obtaining reliable estimates of uncertainty difficult. However, a multi-array versions of our model motivates a method that performs background adjustment, normalization, and summarization as part of the estimation procedure. Using this approach one can compute standard error estimates that account for the three steps.

For example, if we were comparing gene expression across different conditions, each containing various arrays, we could write the following model based on (2):

$$\begin{aligned} Y_{gij} &= O_{gij} + N_{gij} + S_{gij} && (5) \\ &= O_{gij} + \exp(\mu_{gij} + \varepsilon_{gij}) + \exp(s_g + \delta_g X_i + a_{gij} + b_i + \xi_{gij}). && (6) \end{aligned}$$

Here Y_{gij} is the *PM* intensity for the probe j in probeset g on array i, ε_{gij} is a normally distributed error that account for NSB for the same probe behaving differently in different arrays, s_g represents the baseline log expression level for probeset g, a_{gij} represents the signal detecting ability of probe j in gene g on array i, b_i is a term used to describe the need for normalization, ξ_{gij} is a normally distributed term that accounts for the multiplicative error, and δ_g is the expected differential expression for every unit difference in covariate X. Notice δ_g is the parameter of interest. As described by Naef and Magnasco (2003) a_{gj} is a function of α.

With this model in place one may obtain point estimates and standard errors for δ_g using, say, the MLE. With appropriate priors in place one could also obtain Bayesian estimates. However, both these approaches are computationally difficult. In Figure 5 we present some preliminary results obtained using generalized estimating equations to estimate δ. A difficulty with this approach is that when $S_{gij} = 0$ in (5) then (6) is not defined. However, preliminary results look promising and making this approach useful in practice is the subject of current research.

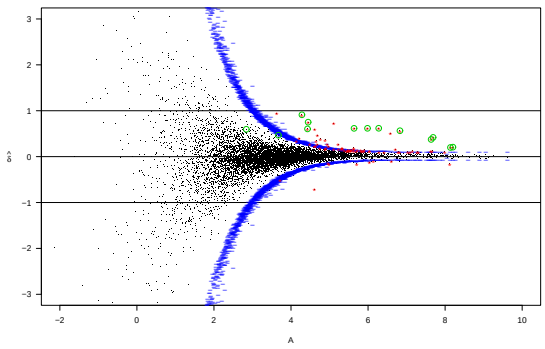

Figure 5: Estimated differential expression $\hat{\delta}_g$ plotted against average log expression level. The data comes from a typical pair of triplicate from the spike-in experiments where the nominal fold-changes for the spiked-in genes are 2. Short horizontal bars are plotted at three times the estimate of standard error of $\hat{\delta}_g$. The spiked-in genes are are shown in circles. Stars denote the non-differentially expressed genes with estimated $\hat{\delta}_g$ beyond three times standard error.

5 Results

5.1 Simulation

To assess the performance of the four discussed background adjustments we performed a simulation. We generated data using models (2) and (6). We selected $\log_2$ expression levels $0,1,\ldots,12$ for $\log_2(S)$ for ease of computation. We chose values for the background parameters based on estimates observed in practice. Specifically, we used $\mu_{PM} = \mu_{MM} = 4.6$ and ε_{gj}^{PM} and ε_{gj}^{MM} bivariate normal with mean 0, variance 1, and correlation 0.88. For each level of S we simulated 320 probe-pairs ($g = 1,\ldots,20, j = 1,\ldots,16$) and for each probe on each of 250 array ($i = 1,\ldots,250$) we generated IID $e \sim$ Normal (0,.08) so that $\varepsilon_{gij}^{PM} = \varepsilon_{gj}^{PM} + e_{gij}^{PM}$ and similarly $\varepsilon_{gij}^{MM} = \varepsilon_{gj}^{MM} + e_{gij}^{MM}$. We used these simulated values to create the *PM* and *MM* intensities.

Figure 6 shows assessments of accuracy and precision. To assess accuracy we show the average adjusted log intensity for each value of $\log_2(S)$ plotted against the true s. Notice that, as expected,

the RMA adjustment is the most biased for small values of $\log_2(S)$. MAS 5.0 and the MLE adjustment perform similarly. The empirical Bayes adjustment is less biased than the RMA adjustment but appears to over correct a bit in the $4 < s < 6$ range. In terms of precision the RMA adjustment is the best. The empirical Bayes adjustment has much better precision than MAS 5.0 and the MLE for $s < 6$. For $s > 6$ all procedures have roughly the same precision. The results presented in Figure 6 are consistent with what is observed with real data, shown in the next Section.

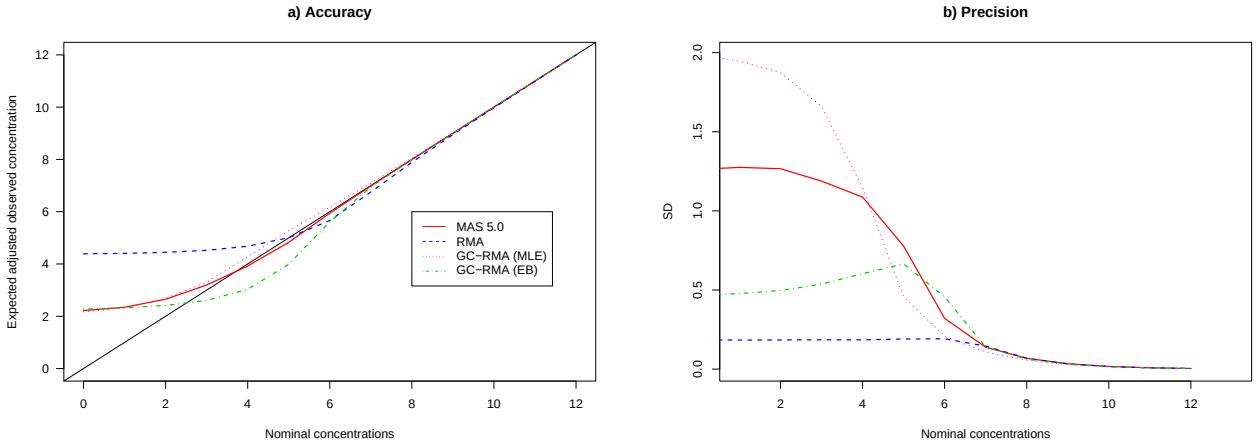

Figure 6: a) Average, over 250 simulated replicate arrays, adjusted probe log intensity plotted against "true" s values. The identity line is also shown. b) Standard deviation across 250 simulated replicate arrays of log-scale intensities plotted against "true" s values.

5.2 Detecting Differentially Expressed Genes

In this section we assess our background adjustment procedure using *bottom line* results of special scientific interest obtained on real data. Both accuracy and precision are discussed. We use 28 arrays from the spike-in data described in Section 2. The 28 arrays were chosen so that array and probe-set effects were balanced with respect to concentration. We computed expression measures for these 28 arrays using MAS 5.0, RMA, and GC-RMA. We will sometimes refer to the expression

values obtained as *observed RNA concentrations*.

For Figure 7 the observed log (base 2) concentration for each concentration group were averaged and plotted against their respective nominal log (base 2) concentration. Ideally, if the nominal concentration doubles, so should the observed concentration, and we should see a line with slope 1. We can see in Figure 7 that the empirical Bayes versions of GC-RMA outperform MAS 5.0 which in turn outperforms RMA.

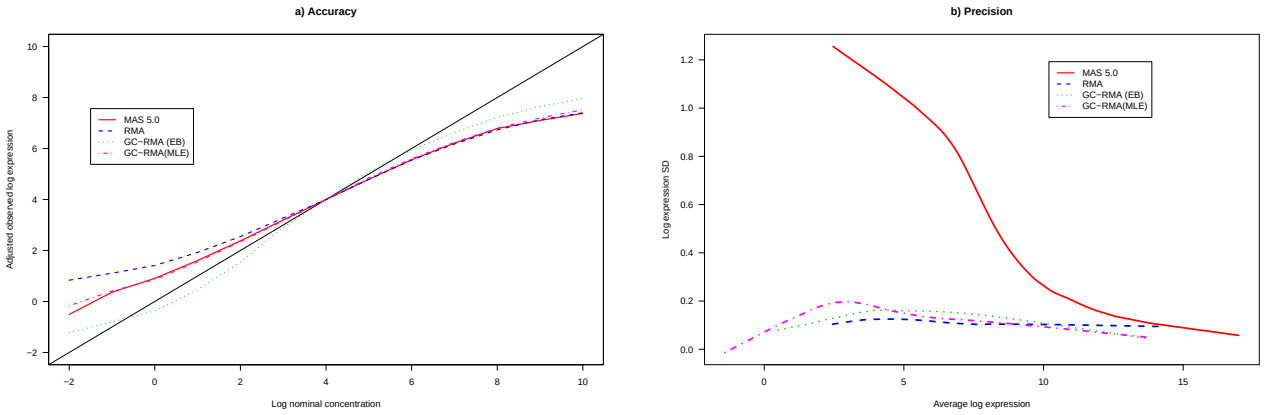

Figure 7: a) Average observed $\log_2$ intensity plotted against nominal $\log_2$ concentration for each spiked-in probeset for all 28 arrays from the Affymetrix spike-in experiment. b) For each non-spiked-in probeset we calculate the mean log expression and the observed log-scale standard deviation across all 28 arrays. The resulting scatter-plot is smoothed to generate a single curve representing mean standard deviation as a function of mean log expression.

Figure 7 also shows smooth functions fitted to the scatter plot of the standard deviation (SD) versus the average of the observed log concentration for each gene across the 28 replicate arrays. These plots show that in terms of precision, RMA and the empirical GC-RMA and RMA outperform MAS 5.0. For genes with over-all low expression RMA and the empirical Bayes GC-RMA are considerably better than MAS 5.0.

Table 1 shows the *local slopes*, observed in Figure 7, which represent the expected observed

Table 1: For concentration groups that differ by a multiple of 2 the local slope is computed. Notice that local slope is simply the difference between groups in average log observed concentration. The last three lines are versions of GCRMA. The column headers define what two concentration groups are being compared.

Method	$\frac{0.50}{0.25}$	$\frac{1}{0.5}$	$\frac{2}{1}$	$\frac{4}{2}$	$\frac{8}{4}$	$\frac{16}{8}$	$\frac{32}{16}$	$\frac{64}{32}$	$\frac{128}{64}$	$\frac{256}{128}$	$\frac{512}{256}$	$\frac{1024}{512}$
MAS 5.0	0.87	0.55	0.7	0.77	0.82	0.8	0.8	0.76	0.65	0.57	0.31	0.28
RMA	0.28	0.3	0.51	0.62	0.73	0.73	0.79	0.76	0.63	0.55	0.4	0.27
EB	0.42	0.44	0.83	1.08	1.37	1.09	1.09	0.85	0.68	0.6	0.43	0.32
MLE	0.58	0.45	0.7	0.79	0.82	0.84	0.85	0.76	0.64	0.56	0.4	0.32
PM-only EB	0.29	0.41	0.8	1.08	1.27	1.24	1.25	0.87	0.71	0.58	0.39	0.37

Table 2: As Table 1 but the local slopes are converted to local ranks out of 12626.

Method	$\frac{0.50}{0.25}$	$\frac{1}{0.5}$	$\frac{2}{1}$	$\frac{4}{2}$	$\frac{8}{4}$	$\frac{16}{8}$	$\frac{32}{16}$	$\frac{64}{32}$	$\frac{128}{64}$	$\frac{256}{128}$	$\frac{512}{256}$	$\frac{1024}{512}$
MAS 5.0	1737	2644	2147	1963	1855	1888	1907	1978	2287	2555	3781	3992
RMA	430	296	14	4	2	2	1	2	4	9	77	451
EB	148	126	2	1	1	1	1	2	8	21	136	451
MLE	253	456	149	100	89	81	78	116	189	274	599	882
PM-only EB	372	71	3	1	1	1	1	2	4	11	100	133

log fold-change for probesets with true fold-change of 1 as a function of the total nominal probeset concentration in the two samples being compared. RMA does considerably worse for the lower concentrations than for the higher. GC-RMA is generally more accurate than MAS 5.0, and MAS 5.0 is more accurate than RMA. To get a bottom-line result we can compute the expected percentile of a gene with a true-fold change of 2, at a particular nominal concentration, when compared to non-differentially expressed genes. Notice that the ideal is a percentile of 100%. To put this in the context of microarray applications we translate the percentiles to the rank among 12625 genes and present these in Table 2. The distribution of fold-changes for non-differentially expressed genes was obtained empirically, i.e. using the 28 replicate arrays. Notice that for high concentrations RMA performs best. For lower concentration, the empirical Bayes version of GC-RMA works best.

6 Software

Clearly the technology being proposed here relies on efficient access to probe-level sequence data. Since there are many different types of chips the software should be modular with respect to chip type. It is also evident that there are many other potential uses for probe-level sequence data so a second guiding software design principle is to ensure that the data are readily available for other uses. The *matchprobes* software package (Huber and Gentleman, 2004) has been created for this purpose. A package specifically for the computation of the empirical Bayes and MLE background procedures is also available. This package together with the *affy* package (Irizarry et al., 2003b) can be used to compute expression measures using these background adjustment. All these components are available as part of the Bioconductor project (`http://www.bioconductor.org`).

7 Discussion

We have presented a statistical model for background noise in Affymetrix GeneChip arrays. Our model takes advantage of sequence information to appropriately describe NSB variation. Estimation procedures motivated by the model result in practical adjustment that improves the over-all sensitivity and specificity of the current default methods. Figure 8 presents log fold change versus average log concentration (MA) plots demonstrating this.

A known problem with our approach is that we assume $\phi = 0$ in (2). However, Table 1 demonstrates that the attenuation effect of subtracting *MM* is not large. The last row in Tables 1 and 2 shows the results obtained using an empirical Bayes approach that ignores the *MM*. As expected, this *PM*-only GC-RMA performs better for highly expressed genes (less attenuation) but a bit

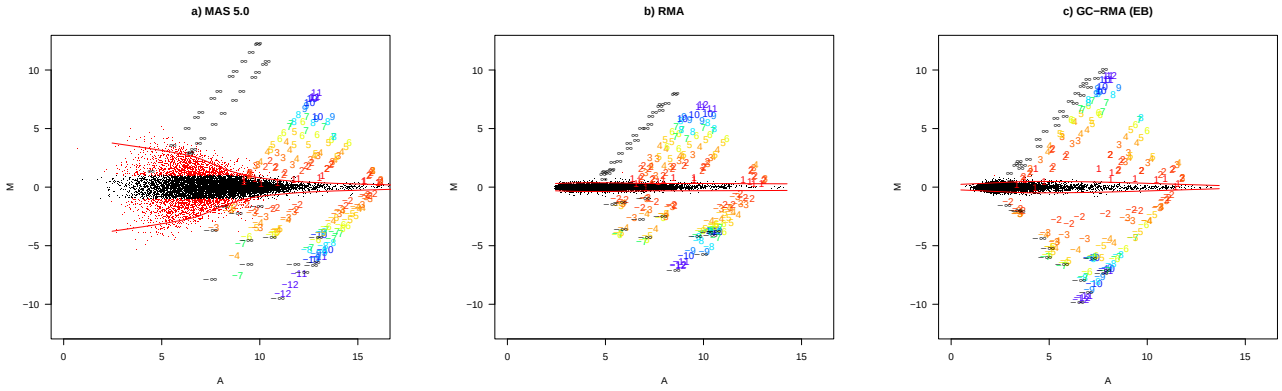

Figure 8: The MA plot shows log fold change as a function of mean log expression level. A set of 14 arrays representing a single experiment from the Affymetrix spike-in data are used for this plot. A total of 13 sets of fold changes are generated by comparing the first array in the set to each of the others. Genes are symbolized by numbers representing the nominal $\log_2$ fold change for the gene. Non-differentially expressed genes with observed fold changes larger than 2 are plotted in different color. All other probesets are represented with black dots. The smooth lines are 3SDs away with SD depending on log expression.

worse for low expressed genes. However, the differences are minor. Because this expression measure uses the *MM* only to estimate the μ_{PM} we need only, say, about 1000 *MM* probes covering the range of α. This implies Affymetrix could manufacture an array for almost half the price (1000 *MM* instead of 100000+) and provide expression measures just as accurate as with current arrays. We predict that Affymetrix will stop using paired *MM* in the not-so distant future.

Two other minor problems with our approach are that 1) we assume the variance σ^2 is independent of α and 2) the attenuation observed for higher probes, predicted by *adsorption* models (Naef et al., 2001), is not accounted for by our model. Future work is to run experiments to study the attenuation and incorporate this appropriately into our model.

References

Affymetrix (2002). *Statistical Algorithms Description Document*. Technical report, Affymetrix Inc. http://www.jax.org/staff/churchill/labsite/research/expression/Cui-Transform.pdf.

Affymetrix Manual (2001). *Affymetrix Mircoarray Suite User Guide version 5.0*. Santa Clara, CA.

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinfromatics* **19**(2), 185–193.

Cui, X., Kerr, M. K., and Churchill, G. A. (2003). Data transformations for cDNA microarray data. http://www.jax.org/staff/churchill/labsite/research/expression/Cui-Transform.pdf.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica* **12**(1), 111–139.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18**(Suppl. 1), S105–S110.

Huber, W. and Gentleman, R. (2004). Matchprobes: a bioconductor package for the sequence-matching of microarray probe elements. *Bioinformatics* To appear.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **1**, 1:9.

Irizarry, R. A., B. Hobbs, F. C., Beaxer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Irizarry, R. A., Gautier, L., and Cope, L. (2003b). An R package for analyses of affymetrix oligonucleotide arrays. In: R. I. G. Parmigiani, E.S. Garrett and S. Zeger (eds.), *The Analysis of Gene Expression Data: Methods and Software*. Springer, Berlin.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A* **98**, 31–36.

Naef, F., Lim, D. A., Patil, N., and Magnasco, M. O. (2001). From features to expression: High density oligonucleotide array analysis revisited. *Tech Report* **1**, 1–9.

Naef, F. and Magnasco, M. O. (2003). Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E* **68**, 011906.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625–637.

Wu, Z. and Irizarry, R. (2004). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. In: *Proceedings of RECOMB 2004*.

Zhang, L., Miles, M. F., and Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology* **21**(7), 818–821.