Measures of Gene Expression for Affymetrix High Density Oligonucleotide Arrays

Rafael A. Irizarry

Abstract

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. In Affymetrix GeneChip array technology, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, separately referred to as probes, arrayed on a silicon chip. After chip measurements are preprocessed, a fluorescence intensity value for each probe is obtained. A necessary step for defining a measure of expression (*ME*) is to summarize the probe intensities for a given gene. In this paper, we review the ideas that motivate a summary statistic, referred to as the robust multi-array average (*RMA*), that improves the default Affymetrix approach and provides substantial benefits to users of the GeneChip technology.

Keywords: Affymetrix GeneChip arrays; background correction; gene expression; normalization; summary measure

1 Introduction

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. Affymetrix GeneChip arrays use oligonucleotides of length 25 base pairs to probe genes. In this technology, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, separately referred to as *probes*, arrayed on a silicon chip. Details of this array technology are described by [1] and [10]. Briefly, though, RNA samples are prepared according to a specific protocol. A fluorescently labeled RNA sample is hybridized to probes on the chip. After some processing steps, the array is scanned with a laser. This scan produces an image that is analyzed to produce an intensity value for each probe (see [9] for more details). These intensities quantify the extent of the hybridization between the labeled target sample and the oligonucleotide probe. A final step to obtain a measure of gene expression (*ME*) is to summarize the intensities for a given gene in order to quantify the amount of corresponding mRNA species in the sample. The intensities obtained for each probe are denoted by PM_{ijn} and MM_{ijn}, $i = 1,\ldots,I$, $j = 1,\ldots,J_n$, and $n = 1,\ldots,N$, with i representing different RNA samples, j representing the probe pair number (this number is related to the physical position of

the oligonucleotide in the gene), and n representing the different genes. The number of genes N usually ranges from 8,000 to 20,000, the number of arrays I is usually small but may be as large as a few hundred, and the number of probe pairs within each gene J_n usually ranges from 11 to 20. Throughout the text, indices are suppressed when there is no ambiguity.

Several researchers have found problems with the *ME* provided by the first version of the Affymetrix system [1] and have suggested alternatives, the most cited example being that of Li and Wong [7]. In their most recent version, Affymetrix provides an alternative as well [2]. There are papers in the literature that compare *ME* by assessing variances, see [8] for an example. Typically, *ME* are obtained from arrays hybridized to RNA aliquots (technical replicates). Throughout the text, we denote the *ME* obtained for a given gene by E_i, with $i = 1,\ldots,I$ representing arrays. When there are replicate arrays, we define the sample variance as $\hat{\sigma}^2 = \sum_{i=1}^{I}(E_i - \bar{E})^2$ with $\bar{E}$ representing the average. *ME* that, in general, have smaller $\hat{\sigma}$ are considered better. However, without an accompanying assessment of the ability to detect signal (which can be thought of as assessing bias), this could produce misleading results. For example, a *ME* that is always $E_k = 0$ cannot be considered appropriate because of its small variance.

Irizarry *et al.* [6] carried out a comparison study of *ME* using two data sets: (*i*) part of the data from an extensive spike-in study conducted by GeneLogic and the Genetics Institute involving about 95 HGU95A human GeneChip arrays, and (*ii*) part of a dilution study conducted by GeneLogic involving 75 HGU95A GeneChip arrays. Four *ME* are compared: (*i*) the Affymetrix commercial software MicroArray Suite MAS 4.0 default (AvDiff) (*ii*) their updated software MAS 5.0 default, (*iii*) the Li and Wong [7] multiplicative model-based *ME*, and (*iv*) a summary based on a log-scale additive model, referred to as the log-scale robust multi-array average (*RMA*). This study seems to be the first to compare *ME* and also to check the reliability of the technology with data for which both bias and variance can be assessed. They find that in general the technology works well, and also that *RMA* outperforms the other three *ME*. In this paper, we give a brief overview of these findings, propose a statistical framework for data using these arrays, and demonstrate with an example why *RMA* works better.

2 Methods

2.1 Background Correction

Several processes can affect the intensities read from each probe. Apart from the specific hybridization directly related to the quantity to be measured, there is also background (or optical noise), nonspecific hybridization, and cross-hybridization. The Affymetrix strategy for extracting the signal of interest from the observed *PM* (perfect match) intensity is to subtract the corresponding *MM* (mismatch) probe intensity. In MAS 4.0, an *ME* for a gene is formed by considering the average difference (AvDiff) of the *PM* and *MM* in the probe set. More precisely, an *ME* for a gene is formed by

defining

$$\text{AvDiff} = N_A^{-1} \sum_{j \in A} (PM_j - MM_j) \tag{1}$$

with A the subset of probes for which $d_j = PM_j - MM_j$ are within 3 SDs away from the average of $d_{(2)}, .., d_{(J-1)}$, where $d_{(j)}$ is the j^{th} smallest difference. N_A represents the number of probes in A. The MAS ME and the ME from the Li and Wong reduced model, discussed in more detail in Section 2.4, are also based on $PM - MM$. Dividing instead of subtracting, *i.e.* using PM/MM, has also been suggested.

The rationale for using the $PM - MM$ quantities is that they correct the effects that bias the PM quantities. Another measure offered in MAS 4.0 software is an average based on the log of ratios PM/MM. There may be biological or physical motivation for considering differences (or ratios). We believe, though, that it is important to corroborate such assumptions empirically.

Figure 1 shows intensities of the PM, MM, PM/MM and $PM - MM$ values for each of the 20 probes representing the BioB-5 probe set in a set of 12 arrays. BioB-5 has been spiked-in on the 12 different arrays at concentrations of 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, and 150 picoMolar. All arrays had a common background cRNA from an acute myeloid leukemia (AML) tumor cell line. All plots in Figure 1 are on the log scale except for 1c. The low values of the $PM - MM$ are plotted on a linear scale because there are several negative values (in fact about 1/3 of the non-spiked in probes have $PM - MM < 0$). The 20 different probe pairs are represented with different symbols and colors. As expected, the PM values are growing in proportion to the concentration. Notice also that the lines representing the 20 probes are close to being parallel showing that there is a strong additive (in the log scale) probe-specific effect. The fact, seen in Figure 1b, that the additive probe-specific effect is also detected by the MM provides motivation for subtracting these values from the PM. However, in Figures 1c and 1d the parallel lines are still seen in $PM - MM$, demonstrating that subtracting is not enough to remove the probe effect. The lack of parallel lines in Figure 1e shows that dividing by MM removes, to some degree, the probe effect. However, since the MM also grow with concentration, and therefore detect signal as well as non-specific binding, results in an attenuated signal. Notice in particular that using PM/MM would make concentrations of 25 and 150, a six-fold difference, indistinguishable. The $PM - MM$ demonstrate some attenuation for the high concentration spike-ins but clearly not as much as PM/MM. Since subtracting probe-specific MM adds noise with no obvious gains in signal detection, and because PM/MM results in a biased signal, [6] propose background correction approaches which are different from subtracting or dividing by MM. We now give a brief review.

The horizontal lines in Figure 1 represent the median intensity obtained from an array for which no spike-in for BioB-5 was added. The dashed lines represent the first and third quartiles. For the lower concentrations, it is hard to distinguish the measured intensities from this median value. Notice also that the signal is attenuated for the lower concentrations. A possible explanation is that background correction is needed.

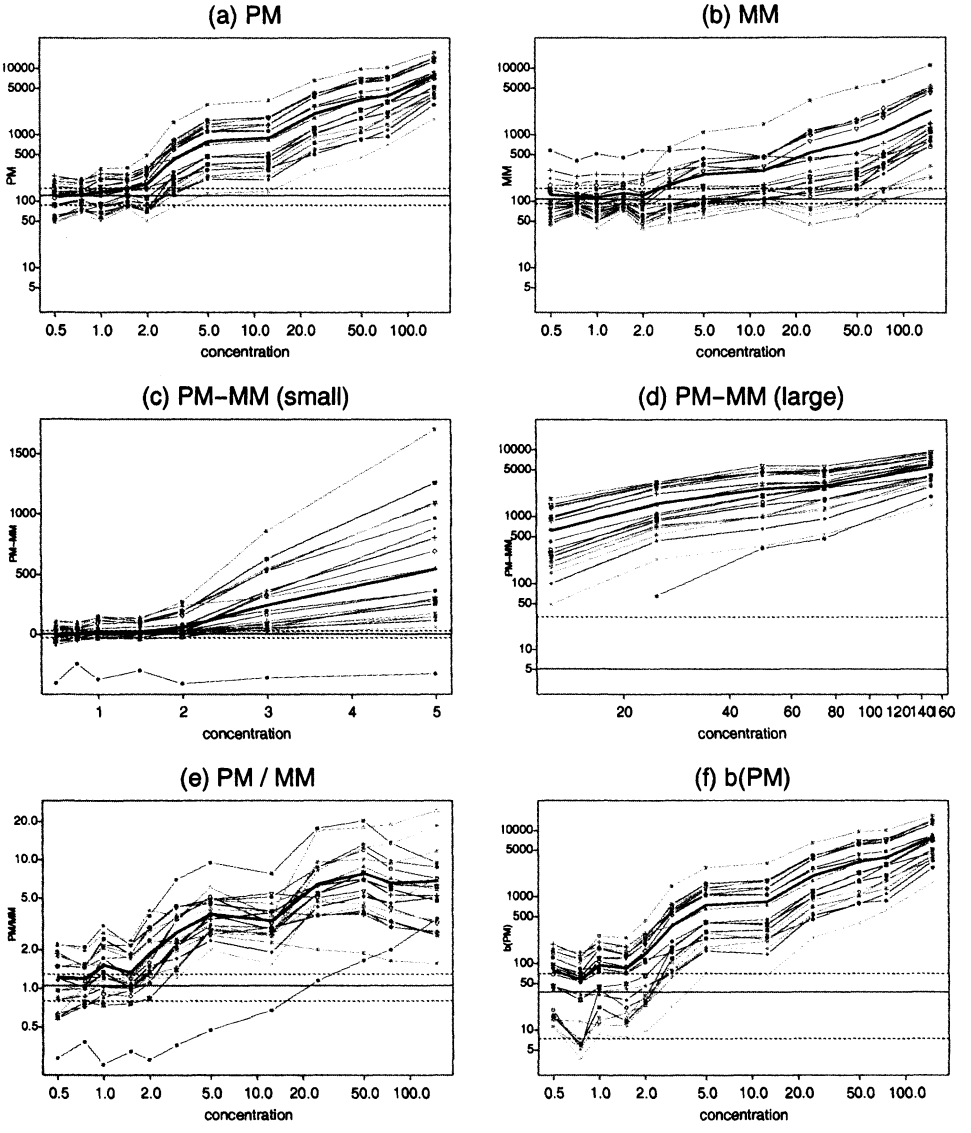

Figure 1: *PM*, *MM*, *PM*/*MM*, $PM - MM$, and $b(PM)$ intensities, for each of the 20 probes representing BioB-5 in 12 arrays where the probe set has been spiked-in, plotted against concentration. Except for 1(c), axes are on the log scale. Different probes are represented by the different colors and symbols. The horizontal line represents the median of the 20 BioB-5 probes from an array where no spike-in was added. The dashed lines are at the 25th and 75th quantiles.

To see this, consider a hypothetical case with two arrays where the signals of a probe set is twice as big in one of the arrays, but an additive signal of 100 units occurs due

to non-specific binding and/or background noise in both arrays. In this case, the observed difference in the signals would be about $\log_2(100+2s) - \log_2(100+s)$ instead of $\log_2(2s) - \log_2(s) = 1$. For small values of s, the incorrect difference would instead be close to 0.

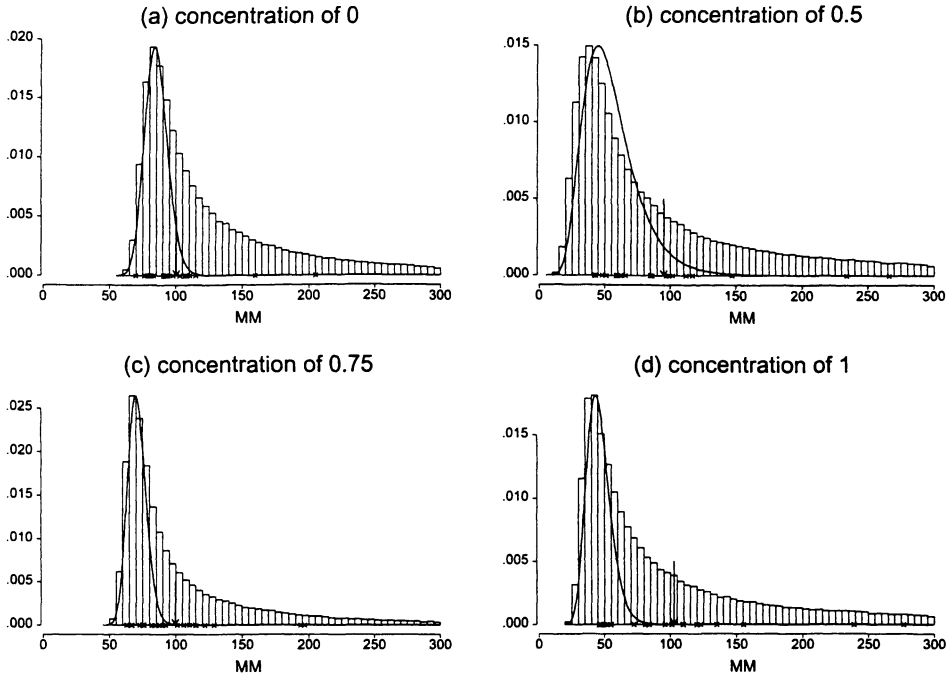

Figure 2: Histograms (density scale) of $\log_2(MM)$ for an array in which no probe set was spiked along with the 3 arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 picoMolar. The observed *PM* values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow.

Figure 2 shows histograms of *MM* for an array in which no probe set was spiked, along with the 3 arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 picoMolar. The observed *PM* values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow. All the average *PM* values are close to 100. Thus, based solely on the average, a difference would be hard to detect. Figures 2 and 3 suggest that the *MM* to the left of the mode of the histogram are similar to the left half of a normal distribution. This suggests that the *MM* are a mixture of (*i*) probes for which an intensity is read due to non-specific binding and background noise and (*ii*) probes detecting transcript signal (cross-hybridization) just like the *PM*. The distance of the average *PM* from the average background noise does in fact increase with concentration. This suggests that background correction of the data is necessary. As noted earlier, $PM - MM$ is not a solution we recommend.

The approach suggested by [6] is to use a global, instead of probe specific, back-

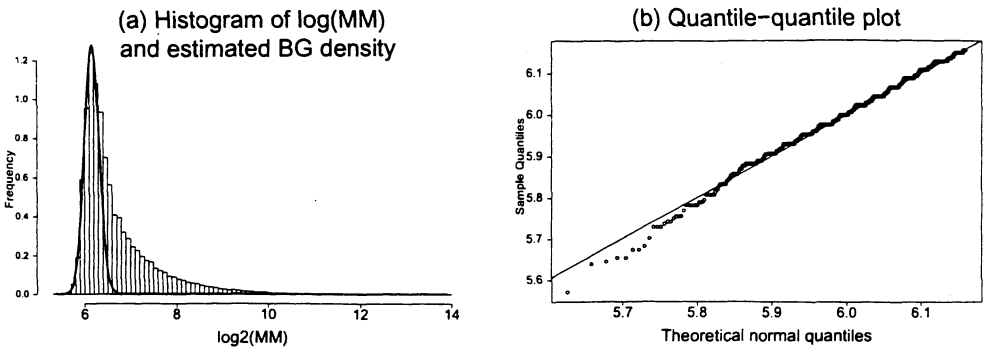

Figure 3: (a) Histogram of $\log_2(MM)$ for spike-in concentration 12.5 picoMolar array in the varying concentration series. (b) QQ plot of the MM left of the mode of histogram (a) compared to a log-normal distribution with mean and SD estimated from the data.

ground correction. We assume that the observed intensity for each PM probe is the sum of a specific binding component and a background component (which may include non-specific binding). Denote these by $PM = S + B$. Because we are interested in S, we use $b(PM) = \mathrm{E}[S|PM]$. We refer to b as a background correcting transformation. Irizarry *et al.* [6] assume B is normally distributed and that S follows an exponential distribution. This assumption is convenient because in this case there is a closed-form solution to $\mathrm{E}[S|PM]$. The solution depends on the mean and variance of the normal distribution and the rate of the exponential distribution. These parameters can be estimated from the PM and MM probe level data. Figure 1f shows the background-corrected PM for the BioB-5 probes. After background transformation, the low concentration values can be distinguished from the values obtained for the array with no spike-in (represented by the horizontal line). In addition, the fact that the slope is larger for the low concentrations in Figure 1f than in Figure 1a demonstrates that the signal is less attenuated for low intensities. However, the intensity values for PM and $b(PM)$ do not grow as a straight line (in the log scale). Further improvements may be obtained with array normalization.

2.2 Normalization

In many of the applications of high density oligonucleotide arrays, the goal is to learn how RNA populations differ in expression in response to genetic and environmental differences. For example, large expression of a particular gene or genes may cause an illness resulting in variation between diseased and normal tissue. Observed expression levels also include variation introduced during sample preparation and array manufacture and processing. Unless arrays are appropriately normalized, comparisons of data from different arrays can lead to misleading results. One approach is *quantile normalization* [6], which forces the empirical distributions of probe intensities from all arrays

to be equal. The approach works well in practice, see [3] for details.

2.3 Statistical Models

Figure 1f demonstrates that the background corrected probe intensities follow an additive model in the log scale. Irizarry *et al.* [6] propose the following model for each probe set

$$\log_2\{b(PM_{ij})\} = \mu_i + \alpha_j + \varepsilon_{ij}, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J, \tag{2}$$

with μ_i representing the log scale *ME* for array i, α_j a probe affinity effect, and ε_{ij} representing an independent identically distributed error term with mean 0. For identifiability of the parameters, we assume that $\sum_j \alpha_j = 0$ for all n. This assumption is equivalent to saying that Affymetrix technology has chosen probes with expected intensities that on average are representative of the associated gene expression.

Under model (2), an unbiased estimate of μ_i, the log scale *ME* for each array, can be obtained using the average

$$\hat{\mu}_i = J^{-1} \sum_{j=1}^{J} \log_2\{b(PM_{ij})\}. \tag{3}$$

Model (2) lends itself to various practical extensions. For example, to compare two populations of RNA species for which there are technical replicates assumed to have the same expected RNA expression, we can write

$$\log_2\{b(PM_{ij}^a)\} = \mu_i^a + \alpha_j + \varepsilon_{ij}^a, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J, \quad a = 1,2.$$

Here i denotes replicate and a the population. The natural estimate of μ^a would be based on I times more data than (3). If instead of technical replicates there were biological replicates, a term Z_{ij}, representing a random effect, could be added to the model.

Li and Wong [7] demonstrate that estimation procedures that remove outliers reduce the variance of *ME* estimates. Model (2) can be easily extended to a context that motivates robust estimates of μ. We refer to the *ME* obtained from estimating μ in model (2) using a robust method, such as the median polish approach used by [4] or robust linear regression, as *RMA* (robust multi-array average).

2.4 Measures of Expression

Figure 4 shows a standard deviation versus average probe intensities scatter-plot from a random sample of *PM* and *MM* obtained from five replicate arrays. Figure 4a shows that the SD increases from roughly 50 to 5000, a factor of 100 fold, as the average increases on its entire range. Figure 4b shows that after a log transformation of the intensities there is only a 1.5 fold increase. This makes the log scale a more natural scale for operations such as averaging. Apparently Affymetrix has also noticed this and,

unlike the MAS 4.0 *ME* AvDiff, their MAS 5.0 *ME* is based on a log scale average. Specifically, for each probe set the MAS 5.0 signal (measure) is defined as

$$\text{signal} = \exp\{\text{Tukey Biweight}(\log(PM_j - CT_j))\}$$

with $CT_j = MM_j$ if $PM_j > MM_j$; if $PM_j < MM_j$, then CT_j is a quantity derived from the *MM* that is never bigger than its *PM* pair. See [5] for more details.

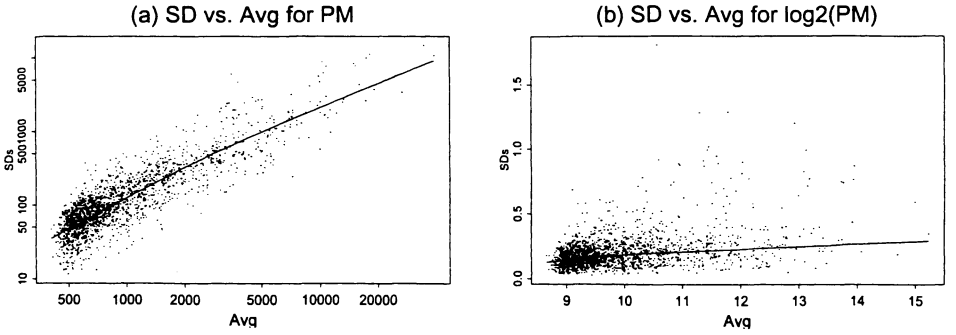

Figure 4: Standard deviations (SDs) plotted against averages from 5 MGU74A mouse arrays for a random sample of 2000 defective probe sets for (a) *PM* and (b) $\log 2(PM)$. The curves are loess fits.

Li and Wong [7] propose using the following model to obtain *ME*:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \tag{4}$$

with ϕ_j representing the probe-specific affinities and independent identically distributed mean 0 normally distributed errors ε_{ij}. For each probe set, an *ME* is defined as the maximum likelihood estimate of θ_i, $i = 1, \ldots, I$ obtained from fitting the multiplicative model. The estimation procedure includes rules for outlier removal. For computational speed, Li and Wong [7] use an iterative procedure that leads to estimates of the form

$$\hat{\theta}_i = \frac{\sum_{j=1}^{J} (PM_{ij} - MM_{ij})\hat{\phi}_j}{\sum_{j=1}^{J} \hat{\phi}_j}, \tag{5}$$

which is basically a weighted version of (1), although their algorithm does remove outliers. This means that probes that are in general high will have a larger influence on $\hat{\theta}_i$. If in fact (2) is a better approximation than (4), then (5) leads to an expression measure with larger variance than *RMA* (see [6]).

3 Results and Discussion

There is no gold standard to compare and test summaries of probe level data. For this reason, data from spike-in experiments have been used to assess the technology

and to motivate normalization procedures. In a similar way, [6] used data from spike-in and dilution experiments to assess the MAS 4.0, MAS 5.0, Li and Wong [7], and *RMA* expression measures. These data are especially useful because here there is an expected result. Irizarry *et al.* [6] demonstrate through examples that *RMA* provides more precise estimates of expression, as well as better specificity and sensitivity for detection of differential expression, than the other three measures. In this section, we give some specific examples that demonstrate why *RMA* performs better.

Figure 5 shows MVA plots: log ratios (or log fold changes) $M_n = \log(E_{1n}/E_{2n})$ versus average expression $A_n = \log(\sqrt{E_{1n}E_{2n}}) = (\log E_{1n} + \log E_{2n})/2$ for *ME* E_{1n} and E_{2n} for all genes, n=1,...,N on two arrays. The arrays being compared here are part of the spike-in experiment described in [6]. We show MVA plots for *ME* obtained using MAS 5.0, Li and Wong [7], and *RMA*. To be able to fit the Li and Wong model and to use a median polish for *RMA*, we compute *ME* using all 33 arrays that were part of the experiment. Because MAS 5.0 is an improved version of MAS 4.0 [2, 6], MAS 4.0 is not shown in Figure 5. The two arrays have 11 control genes spiked-in at different concentrations, but for illustrative purposes we show only DapX-M, which has been spiked in at concentrations of 2 picoMolar and 1 picoMolar on the two arrays respectively. The log ratio for DapX-M should be about 1, corresponding to a fold change of about 2. All other genes represented in the MVA plots should have log ratios of 0 (fold changes of 1, or equal expression) because the samples hybridized to the arrays represent the same biological assay. In the figures, genes having bigger observed fold changes than DapX-M (false positives) are represented with big dots. Only *RMA* has no false positives here. All measures result in an observed log fold change for DapX-M of over 2, which is quite different from 1. Error associated with adding the spike-in to the hybridization sample may account for this difference.

The barplots in Figure 5 show the *PM* and *MM* values for DapX-M and for two other genes that produce false results. One had a large fold change (false positive) estimated from the Li and Wong model (4), the other had a large fold change estimated from MAS 5.0. The barplots show why subtracting the *MM* can cause problems. Notice in particular the 11th probe in DapX-M, where the *MM* are several times higher than the *PM*. They also demonstrate why giving large weight to probes with high values can produce misleading results. For example, probe 13 in the set 33007_at, which is not called an outlier by the Li and Wong algorithm, will have a large weight. Numerical results obtained from these genes are given in Table 1. Table 1 shows that different results can be obtained by using the different *ME*. The values shown in the barplot for probe set 33658_at suggest that there is no fold change occurring for that gene. However, the MAS 5.0 *ME* gives a log ratio of 1/40. The variance added by subtracting the *MM* values causes MAS 5.0 to incorrectly assign a large fold change to this gene.

A possible explanation for why *RMA* outperforms the Li and Wong model is that model (2) fits the data better than (4). The following example supports this explanation. The method of Li and Wong provides not only an estimate of θ_i but a nominal SE for this estimate, denoted here with $\hat{\sigma}_i$. Under (2), one can obtain a naive nominal estimate

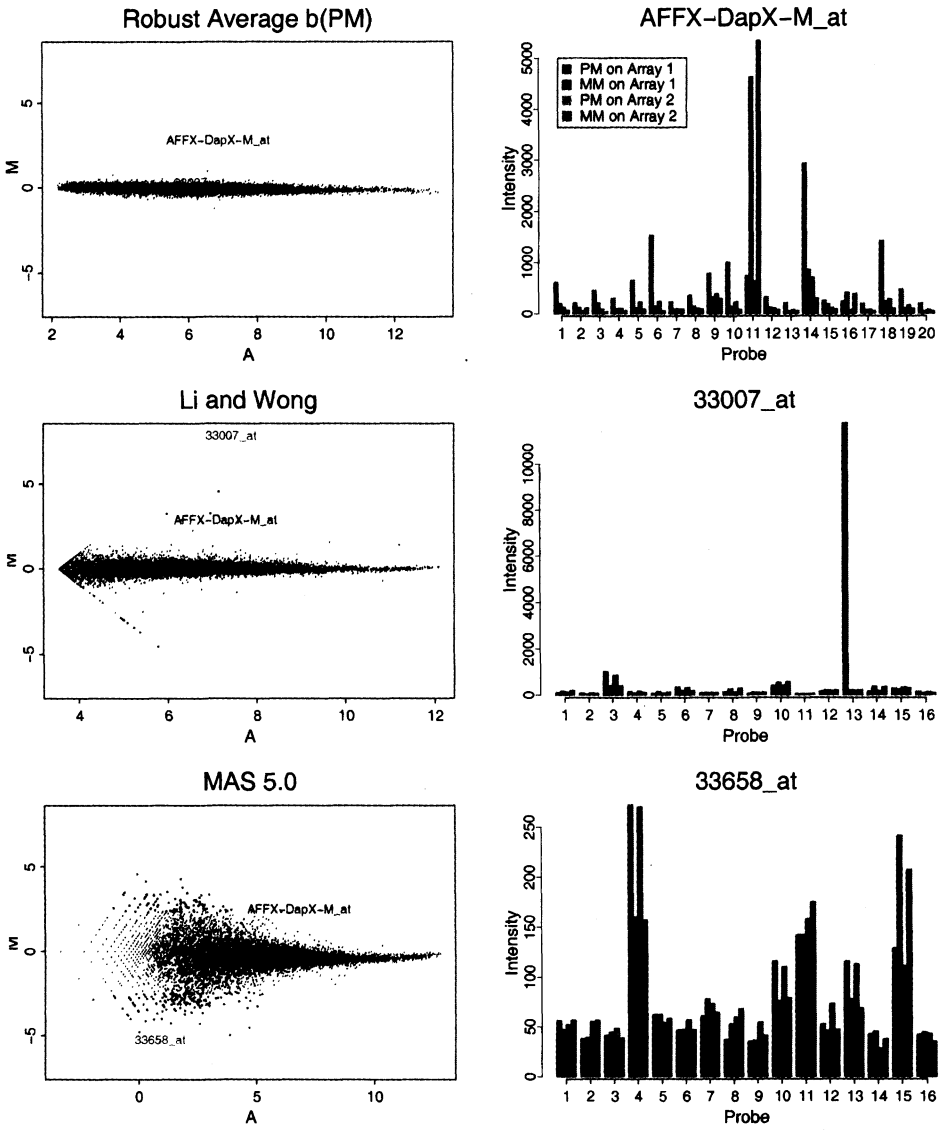

Figure 5: MVA plots indicating the position of the DapX-M which was spiked in at a concentration of 2:1. Barplot for the three genes highlighted in the MVA-plots.

for the SE of $\hat{\mu}$ using an analysis of variance approach. Because there are five replicates, one can also obtain an observed SE of any estimate by simply considering SD_i. If the model is close to the actual mechanism giving rise to the data, the nominal and observed SE should agree. Figure 6 plots the log ratio of nominal to observed variance versus expression measure. These show that in general, the observed and nominal standard errors are closer when using (2) instead of (4).

Table 1: *ME* obtained using *RMA*, the Li and Wong model, and MAS 5.0 for three different genes shown in Figure 5. Only DapX-M should be found to have true fold change.

Gene	*ME*	Array 1	Array 2	Obs. log_2 ratio	Obs. fold change
DapX-M	*RMA*	296.0	46.1	2.7	6.4
DapX-M	LiWong	414.1	61.1	2.8	6.8
DapX-M	MAS 5.0	256.9	49.8	2.4	5.2
33007_at	*RMA*	85.4	74.6	0.1	1.1
33007_at	LiWong	2595.6	11.7	7.8	222.0
33007_at	MAS 5.0	8.4	3.9	1.1	2.2
33658_at	*RMA*	10.0	9.9	0.0	1.0
33658_at	LiWong	25.3	22.7	0.1	1.1
33658_at	MAS 5.0	0.3	12.0	-5.4	1/40

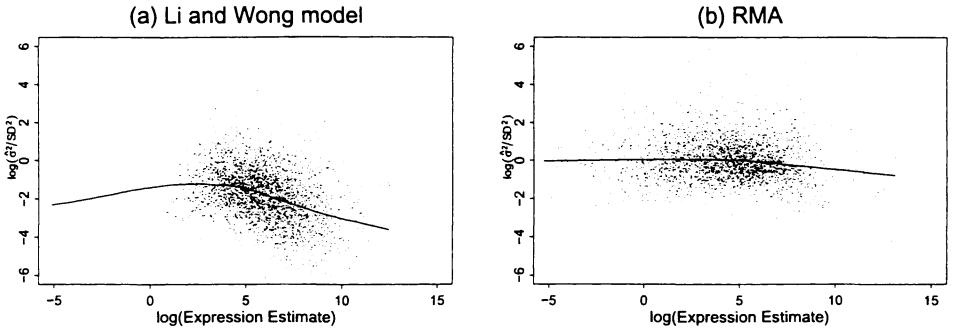

Figure 6: (a) $\log(\hat{\sigma}^2/SD^2)$ plotted against log expression of the Li and Wong *ME*; (b) $\log(\hat{\sigma}^2/SD^2)$ plotted against *RMA*

Irizarry *et al.* [6] developed *RMA*, a summary of Affymetrix GeneChip probe level data, that provides a measure of gene expression, which gives an improved measure compared to other standard measures. The above serves as a specific example demonstrating why *RMA* works better.

4 Acknowledgments

I would like to thank all of the people that have been part of this work, namely: Kristen J. Antonellis, Magnus Åstrand, Yasmin D. Beazer-Barclay, Ben Bolstad, Francois Collin, Leslie Cope, Laurent Gautier, Bridget Hobbs, and Uwe Scherf. I would also like to thank Darlene Goldstein for helpful comments. And a special thanks to Terry Speed. The availability of the dilution experiment, which Terry Speed helped to design, allowed careful assessment and comparison of the different expression measures.

These data are especially useful because we can define outcomes for which there is an expected result.

Rafael A. Irizarry, Department of Biostatistics, Johns Hopkins University, rafa@jhu.edu

References

[1] Affymetrix. *Affymetrix Microarray Suite User Guide.* Affymetrix, Santa Clara, CA, version 4 edition, 1999.

[2] Affymetrix. *Affymetrix Microarray Suite User Guide.* Affymetrix, Santa Clara, CA, version 5 edition, 2001.

[3] B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* In press.

[4] D. Holder, R. F. Raubertas, V. B. Pikounis, V. Svetnik, and K. Soper. Statistical analysis of high density oligonucleotide arrays: a SAFER approach. In *Proceedings of the ASA Annual Meeting, Atlanta, GA 2001*, 2001.

[5] E. Hubbell. Estimating signal with next generation Affymetrix software. In *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip® data*, 2001. http://www.stat.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html.

[6] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* In press.

[7] C. Li and W. H. Wong. 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, 2001.

[8] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2:1–11, 2001.

[9] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart. High density synthetic oligonucleotide arrays. *Supplement to Nature Genetics*, 21:20–24, 1999.

[10] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.