Biostatistics (2003), **4**, 2, *pp.* 249–264
Printed in Great Britain

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

RAFAEL A. IRIZARRY*
Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205, USA
rafa@jhu.edu

BRIDGET HOBBS
Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia

FRANCOIS COLLIN
Gene Logic Inc., Berkeley, CA, USA

YASMIN D. BEAZER-BARCLAY, KRISTEN J. ANTONELLIS, UWE SCHERF
Gene Logic Inc., Gaithersburg, MD, USA

TERENCE P. SPEED
Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia. Department of Statistics, University of California at Berkeley

SUMMARY

In this paper we report exploratory analyses of high-density oligonucleotide array data from the Affymetrix GeneChip® system with the objective of improving upon currently used measures of gene expression. Our analyses make use of three data sets: a small experimental study consisting of five MGU74A mouse GeneChip® arrays, part of the data from an extensive spike-in study conducted by Gene Logic and Wyeth's Genetics Institute involving 95 HG-U95A human GeneChip® arrays; and part of a dilution study conducted by Gene Logic involving 75 HG-U95A GeneChip® arrays. We display some familiar features of the perfect match and mismatch probe (PM and MM) values of these data, and examine the variance–mean relationship with probe-level data from probes believed to be defective, and so delivering noise only. We explain why we need to normalize the arrays to one another using probe level intensities. We then examine the behavior of the PM and MM using spike-in data and assess three commonly used summary measures: Affymetrix's (i) average difference (AvDiff) and (ii) MAS 5.0 signal, and (iii) the Li and Wong multiplicative model-based expression index (MBEI). The exploratory data analyses of the probe level data motivate a new summary measure that is a robust multi-array average (RMA) of background-adjusted, normalized, and log-transformed PM values. We evaluate the four expression summary measures using the dilution study data, assessing their behavior in terms of bias, variance and (for MBEI and RMA) model fit. Finally, we evaluate the algorithms in terms of their ability to detect known levels of differential expression using the spike-in data. We conclude that there is no obvious downside to using RMA and attaching a standard error (SE) to this quantity using a linear model which removes probe-specific affinities.

*To whom correspondence should be addressed

An R package with the functions used for the analyses in this paper is part of the Bioconductor project and can be downloaded (http://www.bioconductor.org). Supplemental material, such as color versions of the figures, is available on the web (http://www.biostat.jhsph.edu/∼ririzarr/affy).

1. INTRODUCTION

High-density oligonucleotide expression array technology is now widely used in many areas of biomedical research. The system (Lockhart *et al.*, 1996) uses oligonucleotides with length of 25 base pairs that are used to probe genes. Typically, each gene will be represented by 16–20 pairs of oligonucleotides referred to as *probe sets*. The first component of these pairs is referred to as a perfect match (PM) probe. Each PM probe is paired with a mismatch (MM) probe that is created by changing the middle (13th) base with the intention of measuring non-specific binding. The PM and MM are referred to as a *probe pair*. See the Affymetrix Microarray Suite User Guide (1999) for details. RNA samples are prepared, labeled and hybridized with arrays. Arrays are scanned and images are produced and analysed to obtain an intensity value for each probe. These intensities represent how much hybridization occurred for each oligonucleotide probe. Of interest is finding a way to combine the 16–20 probe pair intensities for a given gene to define a measure of expression that represents the amount of the corresponding mRNA species.

We denote the intensities obtained for each probe as

$$PM_{ijn} \text{ and } MM_{ijn}, i = 1, \ldots, I, \; j = 1, \ldots, J_n, \text{ and } n = 1, \ldots, N$$

with n representing the different genes, i representing different RNA samples, and j representing the probe pair number (this number is related to the physical position of the oligonucleotide in the gene). The number of genes N usually ranges from 8 000 to 20 000, the number of arrays I ranges from one to hundreds, and the number of probe pairs within each gene J_n usually ranges from 16 to 20. Throughout the text indices are suppressed when there is no ambiguity.

Section 2 describes the three data sets used in this paper. Section 3 explores various interesting features of the data with the objective of defining an effective measure of gene expression using the probe level data. Section 4 describes normalization. Some expression measures, for example AvDiff and Li and Wong's MBEI, are based on $PM - MM$. Other measures, for example Affymetrix's Average Log Ratio, are based on $\log(PM/MM)$. In Sections 3 and 4 we also explore the behavior of these quantities. Section 5 describes four measures of expression. Section 6 assesses the four expression measures in terms of bias, variance, and model fit. Section 7 examines the ability of the four methods at detecting differentially expressed probe sets. Section 8 presents our conclusions.

2. DESCRIPTION OF DATA

To properly compare summary measures of expression in terms of bias, variance, sensitivity, and specificity, data for which we know the 'truth' is required. In this paper we examine three data sets for which assessments can be performed where specific results are expected. Data set A provides probes for which we can assume the measurements are entirely due to non-specific binding. This permits us to study the variance–mean relationship for intensity measures. Data set B provides the results of a spike-in experiment where gene fragments have been added at known concentrations. These data can be used to assess bias, sensitivity and specificity. Data set C provides the results from a study in which samples were hybridized at different dilutions. This permits us to assess bias and variance in a more 'realistic' scenario than with data set B.

Data sets B and C are available from the web at http://qolotus02.genelogic.com/datasets.nsf/. In this section we describe them in detail for readers interested in using them. We also explain which specific subsets of the data were used for the analyses presented in this paper.

2.1 *Mouse data set A*

Data set A comes from an experiment where five MG-U74A mouse GeneChip® arrays were used. These were hybridized with samples of lung tissue mRNA obtained from five mice exposed to different experimental conditions. About 1/5 of the probe pairs in the MG-U74A array were incorrectly sequenced. We therefore assume that the measurements read for most of these probes are entirely due to non-specific binding.

2.2 *Spike-in data sets B*

Data set B consists of experiments where 11 different cRNA fragments were added to the hybridization mixture of the GeneChip® arrays at different picomolar (pM) concentrations. The 11 control cRNAs were BioB-5, BioB-M, BioB-3, BioC-5, BioC-3, BioDn-5 (all *E. coli*), CreX-5, CreX-3 (phage P1), and DapX-5, DapX-M, DapX-3 (*B. subtilis*) (Hill *et al.*, 2000, 2001; Baugh *et al.*, 2001). The cRNA were chosen to match the target sequence for each of the Affymetrix control probe sets. For example, for DapX (a *B. subtilis* gene), the $5'$, middle and $3'$ target sequences (identified by DapX-5, DapX-M, DapX-3) were each synthesized separately and spiked-in at a specific concentration. Thus, for example, on one of the arrays DapX-3 target sequence was added to the total hybridization solution of 200 μl to give a final concentration of 0.5 pM.

There are two series of spike-in experiments. The experiments were originally carried out for the development of normalization procedures (Hill *et al.*, 2001). In this paper we use the data in a different way, mainly for the comparison of expression measures.

2.2.1 *The varying concentration series data set, B1.*

For an individual array, all of the 11 control cRNAs were spiked-in at the same concentration and this concentration was varied across arrays, taking the values 0.0, 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, and 150 pM. For example, array 1 had all control cRNAs spiked with 0.0 pM and array 2 had all control cRNAs spiked with 0.5 pM, etc. Of these 12 concentrations, 0, 0.5, 0.75, 1, 1.5, 2, 3 were represented on just one array, 5 and 100 on two arrays, and the rest were in triplicate, i.e. on three arrays for a total of 27 arrays. All arrays have a common background cRNA from an acute myeloid leukemia (AML) tumor cell line. In this paper we use only 12 arrays, one replicate for each of the 12 concentrations. One of the probe set spike-in combinations (CreX-3) failed to respond adequately, and data from that probe set is entirely omitted from the analysis. Thus we analyse data from 10 spiked-in probe-sets.

2.2.2 *Latin square series data set, B2.*

In this series each of the 11 control cRNAs were spiked-in at a different concentration on each array (apart from replicates). The 12 concentrations used were 0.5, 1, 1.5, 2, 3, 5, 12.5, 25, 37.5, 50, 75, and 100 pM, and these were arranged in a 12×12 cyclic Latin square, with each concentration appearing once in each row and column. The 12 combinations of concentrations used on the arrays were taken from the first 11 entries of the 12 rows of this Latin square. Of the 12 combinations used, 11 were done on three arrays and one on just one array. All of these arrays had the same AML background as in data set B1.

The analysis in this paper makes use of data from six arrays that are a pair of triplicates. The spike-in concentrations for each of the 11 control RNAs on the two sets of triplicates is shown in Table 1.

Table 1. *Concentrations and observed ranks of each spiked-in gene in a comparison of two sets of triplicates from the Latin square series spike-in data set*

	Concentration		Expected Rank	Observed Rank			
Probe set	Set of triplicates 1	set of triplicates 2		AvDiff	MAS 5.0	Li & Wong	RMA
BioB-5	100.0	0.5	1	6	2	1	1
BioB-3	0.5	25.0	2	16	1	3	2
BioC-5	2.0	75.0	4	74	6	2	3
BioB-M	1.0	37.5	4	30	3	7	5
BioDn-3	1.5	50.0	5	44	5	6	4
DapX-3	35.7	3.0	6	239	24	24	7
CreX-3	50.0	5.0	7	333	73	36	9
CreX-5	12.5	2.0	8	3276	33	3128	8
BioC-3	25.0	100.0	9	2709	8579	681	6431
DapX-5	5.0	1.5	10	4598	102	12203	10
DapX-M	3.0	1.0	11	165	19	13	6

Notice that relative concentrations of the spike-ins are three fold or more, which permits us to check the sensitivity of expression indices.

2.3 *Dilution data set C*

Two sources of cRNA, A (human liver tissue) and B (central nervous system cell line), were hybridized to human array (HG-U95A) in a range of proportions and dilutions. In this publication, we study data from arrays hybridized to source A starting with 1.25 μg cRNA, and rising through 2.5, 5.0, 7.5, 10.0 to 20.0 μg. There were five replicate arrays for each tissue: that is, each generated cRNA was hybridized on five HG-U95 GeneChip® arrays. Five scanners were used in this study. Each array replicate was processed in a different scanner.

3. FEATURES OF PROBE LEVEL DATA

Figure 1(a) shows histograms of log ratio, $\log_2(PM/MM)$, stratified by quantiles of abundance, $\log_2\sqrt{PM \times MM}$, with gray scale representing height of histogram (light is high and dark is low) for one array from data set A. The histograms have been scaled so that the mode of each histogram is represented with the same gray scale. This figure shows that, in general, MM grows with PM. Furthermore, for larger values of abundance the differences have a bimodal distribution with the second mode occurring for negative differences. The same bimodal effect is seen when we stratify by $\log_2(PM)$, thus it is not an artifact of conditioning on sums. In Figures 1(b)–1(e), four histograms with a broader stratification clearly show this effect. The figure also displays (in darker grays) the histograms of the defective probes where the bimodal distribution is also seen. Notice, there are many probe pairs with $MM >> PM$. Finally, notice that for about 1/3 of the probes $MM > PM$. The number of probe pairs within probe sets for which $MM > PM$ varies from 0 to 14. The distribution across probe sets is the following:

# of times $MM > PM$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# of probe sets	7401	481	628	819	1123	1461	1759	1906	1555	1200	760	345	152	50	14

All these effects have been seen in many arrays.

The defective probes are used to assess the variance–mean relationship. Intensities obtained from probe j in arrays $i = 1, \ldots, I$, PM_{ijn}, are expected to have the same mean and variance. If standard deviations (SDs) $\sqrt{\{(I-1)^{-1}\sum(PM_{ijn}-\bar{PM}_{\cdot jn})^2\}}$ and averages $\bar{PM}_{\cdot jn} = I^{-1}\sum_i PM_{ijn}$ are computed

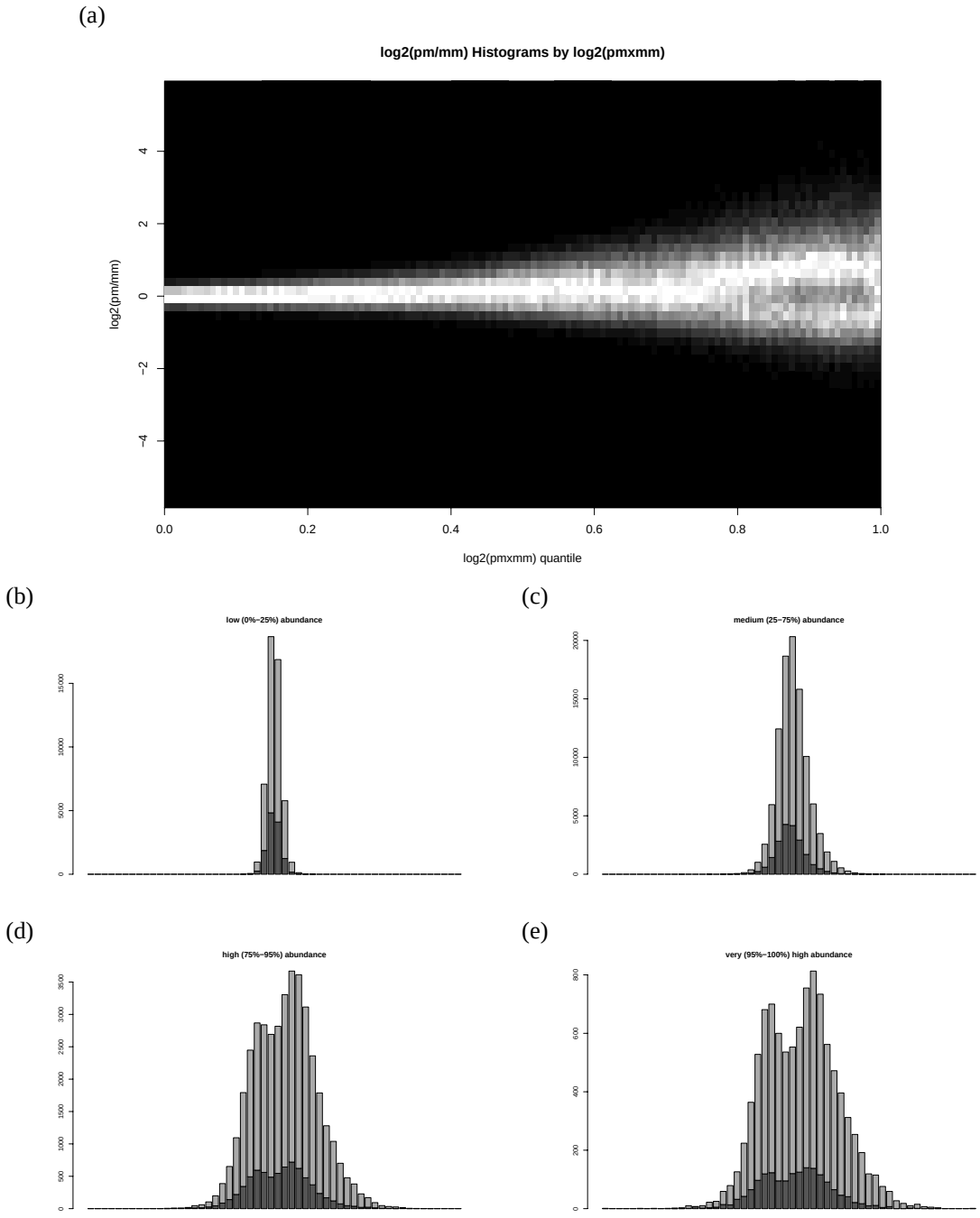

Fig. 1. (a) Histograms of log ratio $\log_2(PM/MM)$, stratified by quantiles of abundance, $\log_2\sqrt{PM \times MM}$, with gray scale representing height of histogram (light grays are high and dark grays are low) for one array from the mouse data set. The histograms have been scaled so that the mode of each histogram is represented with the same gray scale. (b) Histogram of log ratios for first quartile of abundance with the histogram for the defective probes represented by a darker gray. (c) Like (b) for abundance values between first and third quartile. (d) Like (b) for abundance values in the last quartile excluding the highest 5 percent. (e) Like (b) for the highest 5% of abundance.

for a random sample of 2000 defective probe sets, the SD increases from roughly 50 to 5000, a factor of 100-fold, as the average increases on its entire range. After a log transformation of the PM intensities there is only a 1.5-fold increase.

4. 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. These sources of variation are referred to as *interesting variation*. Observed expression levels also include variation introduced during the sample preparation, manufacture of the arrays, and the processing of the arrays (labeling, hybridization, and scanning). These are referred to as sources of 'obscuring variation'. See (Hartemink *et al.*, 2001) for a more detailed discussion. The obscuring sources of variation can have many different effects on data. Unless arrays are appropriatly *normalized*, comparing data from different arrays can lead to misleading results.

Dudoit *et al.* (2001) describe the need for normalization procedures for cDNA microarray data. Similar issues are present with GeneChip® arrays. Figures 2(a) and 2(b) show box plots of $\log_2(PM)$ and $PM - MM$ for data set C. The different gray scales represent the six different sets of five replicates processed on scanners 1 to 5. The scanner effect is clearly seen in Figure 2. For example, note that the $\log_2(PM)$ boxplot intensities obtained using scanner/fluidic station 1 were in general higher than those obtained from scanner/fluidic station 5. For the replicate arrays we expect no genes to be differentially expressed. This figure shows direct array to array comparison of PM values warrants normalization. Figure 2(b) boxplot shows that further normalization is needed for the $PM - MM$ as well.

Figures 3(a) and 3(b) show log ratios, $M = \log_2(y/x)$ versus abundance $A = \log_2 \sqrt{x \times y}$, (MVA) plots for $x = PM_1, y = PM_2$ and $x = PM_1 - MM_1, y = PM_2 - MM_2$ for two arrays (denoted with 1 and 2) in which the BioDn-3 gene has been spiked at 5 pM and 2 pM respectively. These plots have been used by, for example, Dudoit *et al.* (2002) to explore intensity related biases. Because the same RNA background was hybridized to arrays 1 and 2, we do not expect any of the non-spiked-in genes to be differentially expressed and therefore these plots to scatter around 0. It is clear from Figure 3 that these data need normalization.

For cDNA arrays the normalization procedure presented in Dudoit *et al.* (2002) has worked well in practice. For each array, a loess curve is fitted to the MVA plot of intensities of the red and green labels and the residuals are considered the normalized log ratios. However, this approach is not appropriate for GeneChip® arrays because only one sample is hybridized to each array instead of two (red and green). A procedure that normalizes each array against all others is needed.

Various methods have been proposed for normalizing GeneChip® arrays. Bolstad *et al.* (2002) present a review of these methods and find *quantile normalization* to perform best. The goal of quantile normalization is to make the distribution of probe intensities the same for arrays $i = 1, \ldots, I$. The normalization maps probe level data from all arrays, $i = 1, \ldots, I$, so that an I-dimensional quantile–quantile plot follows the I-dimensional identity line. A possible problem with this approach is that we risk removing some of the signal in the tails. However, empirical evidence suggest this is not a problem in practice: see Bolstad *et al.* (2002) for details.

In Figures 3(c) and 3(d) the MVA plots of the normalized arrays are shown. Notice how the normalization has removed the bias seen in Figures 3(a) and 3(b). The large points represent the 20 spiked-in probes and the small black dots represent a random sample of non-spiked-in probes. Notice that in all plots, normalization helps identify the spiked-in probes as differentially expressed. The benefits of this normalization at the probe level are also seen in Figures 2(c) and 2(d).

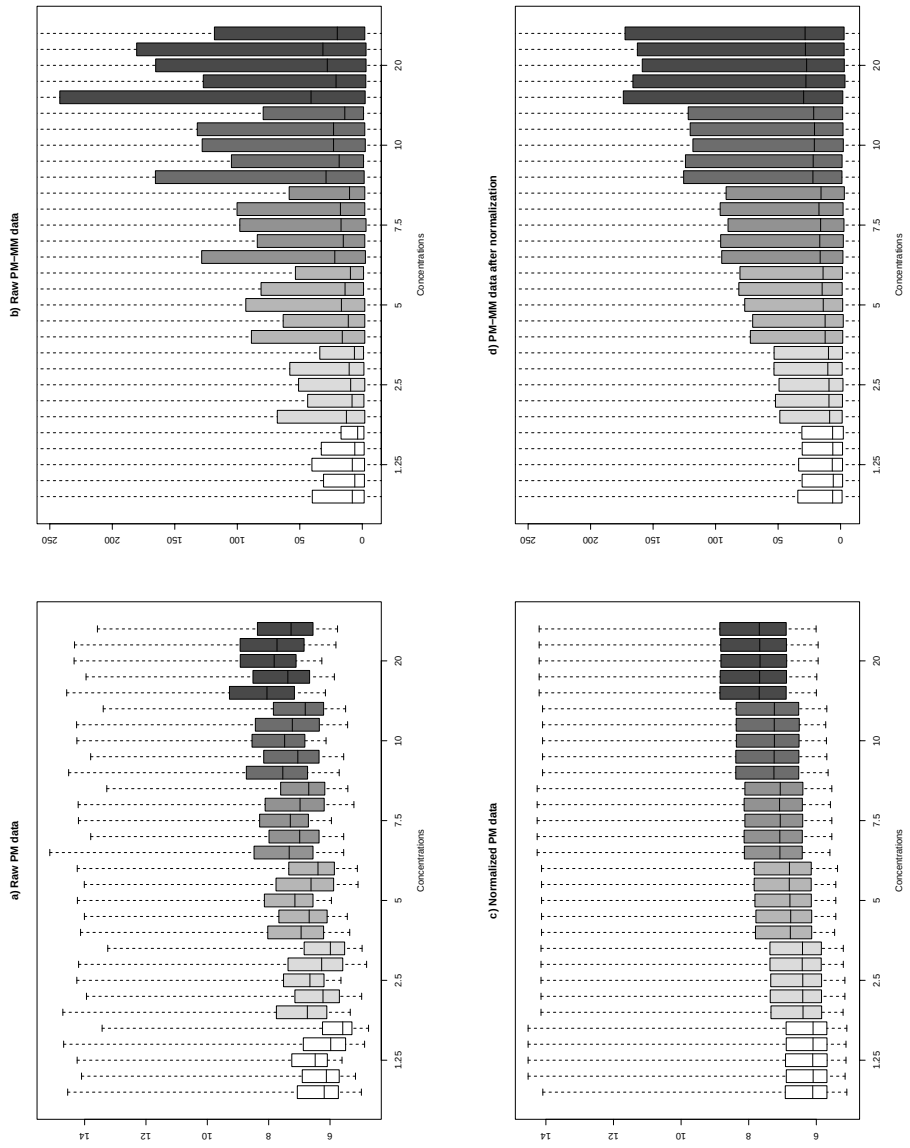

Fig. 2. Boxplots of $\log_2(PM)$ and $PM - MM$ for the 30 arrays from data set C. Because $PM - MM$ values are usually between -2000 and 10 000, a reduced range is used to get a better view of the interquartile range. The bottom row are the after quantile normalization boxplots. The y-axis scale can be deduced from the plot titles.

5. MEASURES OF EXPRESSION

Various measures of expression have been proposed: for example see Li and Wong (2001), Naef *et al.* (2001), and Holder *et al.* (2001). The most commonly used (at the time this paper was written) is AvDiff,

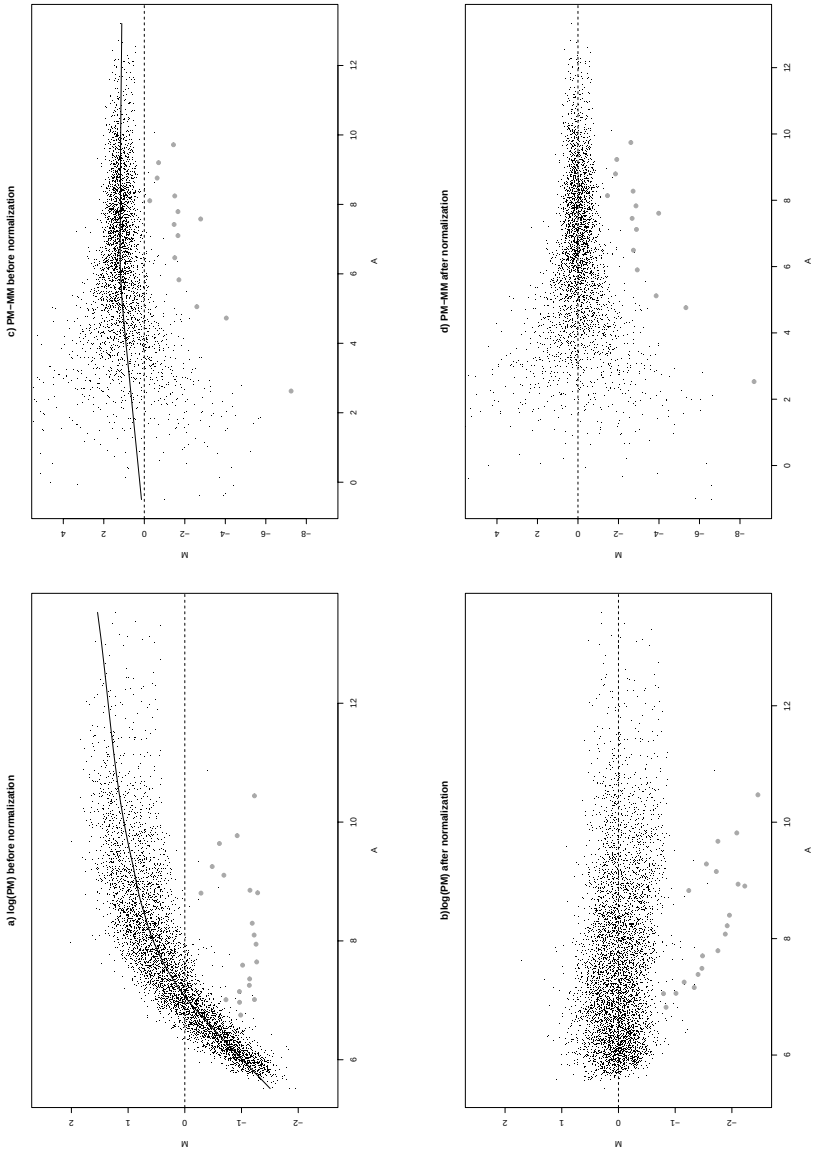

Fig. 3. MVA plots (described in text) of $\log_2(PM)$ and $\log_2(PM - MM)$ for two arrays in which the BioDn-3 gene has been spiked at 5 pM and 2 pM respectively. The large points represent the 20 spiked-in probes and the small black dots represent a random sample of non-spiked-in probes. (a) and (c) are before normalization and (b) and (d) are after quantile normalization.

the Affymetrix default. For each probe set n on each array i, AvDiff is defined by

$$\text{AvDiff} = \frac{1}{\#A} \sum_{j \in A} (PM_j - MM_j)$$

with A the subset of probes for which $d_j = PM_j - MM_j$ are within 3 SDs away from the average of $d_{(2)}, .., d_{(J-1)}$ with $d_{(j)}$ the jth smallest difference. $\#A$ represents the cardinality of A. Many of the other expression measures are versions of AvDiff with different ways of removing outliers and different ways of dealing with small values.

We have observed that linear scale measures, such as AvDiff, are not optimal. Li and Wong (2001) observed this and proposed an alternative model based expression index. For each probe set n, Li and Wong's measure is defined as the maximum likelihood estimates of the $\theta_i, i = 1, \ldots, I$ obtained from fitting

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij} \tag{1}$$

with ϕ_j representing probe-specific affinities and the ϵ_{ijn} are assumed to be independent normally distributed errors. The estimation procedure includes rules for outlier removal.

Affymetrix also appears to have noticed that the linear scale is not appropriate and, in the new version of their analysis algorithm MAS 5.0, are now using a log scale measure. Specifically the MAS 5.0 signal (measure) is defined as

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

with CT_j a quantity derived from the MMs that is never bigger than its PM pair. See Hubbell (2001) for more details.

Each of these measures rely upon the difference $PM - MM$ with the intention of correcting for non-specific binding. However, the exploratory analysis presented in Section 3 suggests that the MM may be detecting signal as well as non-specific binding. Some researchers (Naef *et al.*, 2001) propose expression measures based only on the PM. In Figure 4 we show the PM, MM, PM/MM and $PM - MM$ values for each of the 20 probes representing BioB-5 in the 12 spiked-in arrays, from data set B1, plotted against spike-in concentration. The 20 different probe pairs are represented with different symbols and line types. 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 there is a strong additive (in the log scale) probe-specific effect. As evident in Figure 4(c), the additive probe-specific effect is also detected by the MM motivating their subtraction from the PM. However, in Figure 4(d) the parallel lines are still seen in $PM - MM$, demonstrating that subtracting is not enough to remove the probe effect. The fact that parallel lines are not as obvious in Figure 4(c) shows that dividing by MM removes, to some degree, the probe effect. However, the MM also grow with concentrations, because they detect signal as well as non-specific binding, hence the signal in PM/MM is attenuated. Notice, in particular, that PM/MM is unable to distinguish between concentrations of 25 and 150. Since subtracting probe-specific MM adds noise with no obvious gain in bias and because PM/MM results in a biased signal, in this paper we propose an alternative measure to those based on $PM - MM$ or PM/MM.

Figure 4(a) shows that on a log scale (i) the PMs grow roughly linearly with respect to concentrations, (ii) the variances are roughly constant and (iii) the probe-specific affinity is approximately additive. This suggests an additive linear model for the probe set data and the average $J^{-1}\sum_{j=1}^{J}\log(PM_{ij})$ as a log scale measure of expression. However, this measure does not account for non-specific binding. Because, in Figure 4, the log-scale slope of the PM is less than 1, particularly for small concentrations, the PM values should be adjusted to account for non-specific binding. To see this consider a hypothetical case with two arrays where the signal 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 $\log_2(2s) - \log_2(s)$. For small values of s the incorrect difference would be close to 0 instead of 1.

Figure 5 shows histograms of $\log_2(MM)$ for an array in which no probe-set was spiked along with the three arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 pM (from data set

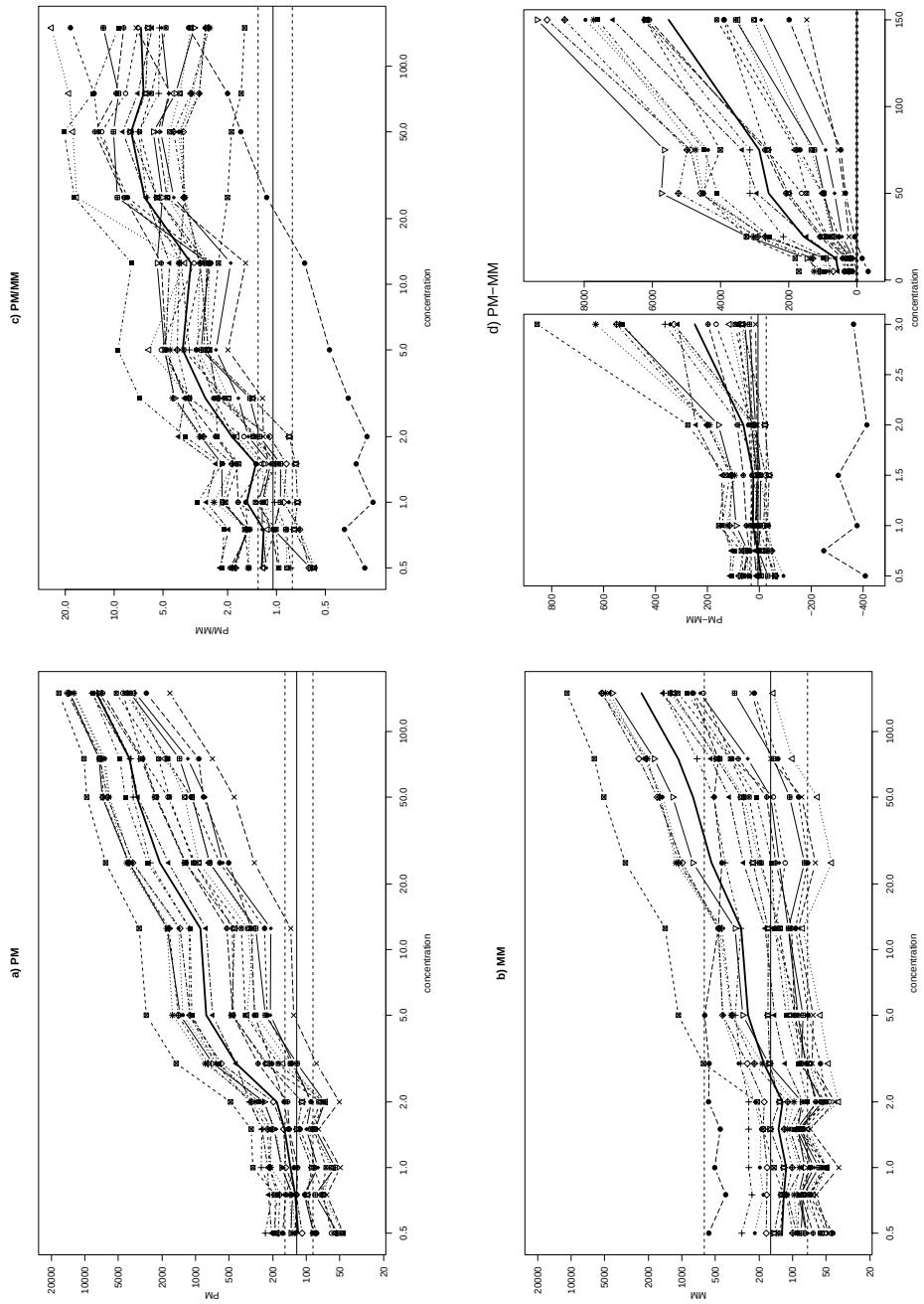

Fig. 4. PM, MM, PM/MM, and $PM - MM$ values for each of the 20 probes representing BioB-5 (with the exception of CreX-3, all other spike-in genes behaved similarly to BioB-5) in the 12 spiked-in arrays from the varying concentration experiment plotted against concentration. The different probes are represented by the different line types and symbols. The horizontal line represents the median of the 20 BioB-5 probes for the non-spiked-in array. The dashed lines are the 25th and 75th quantiles.

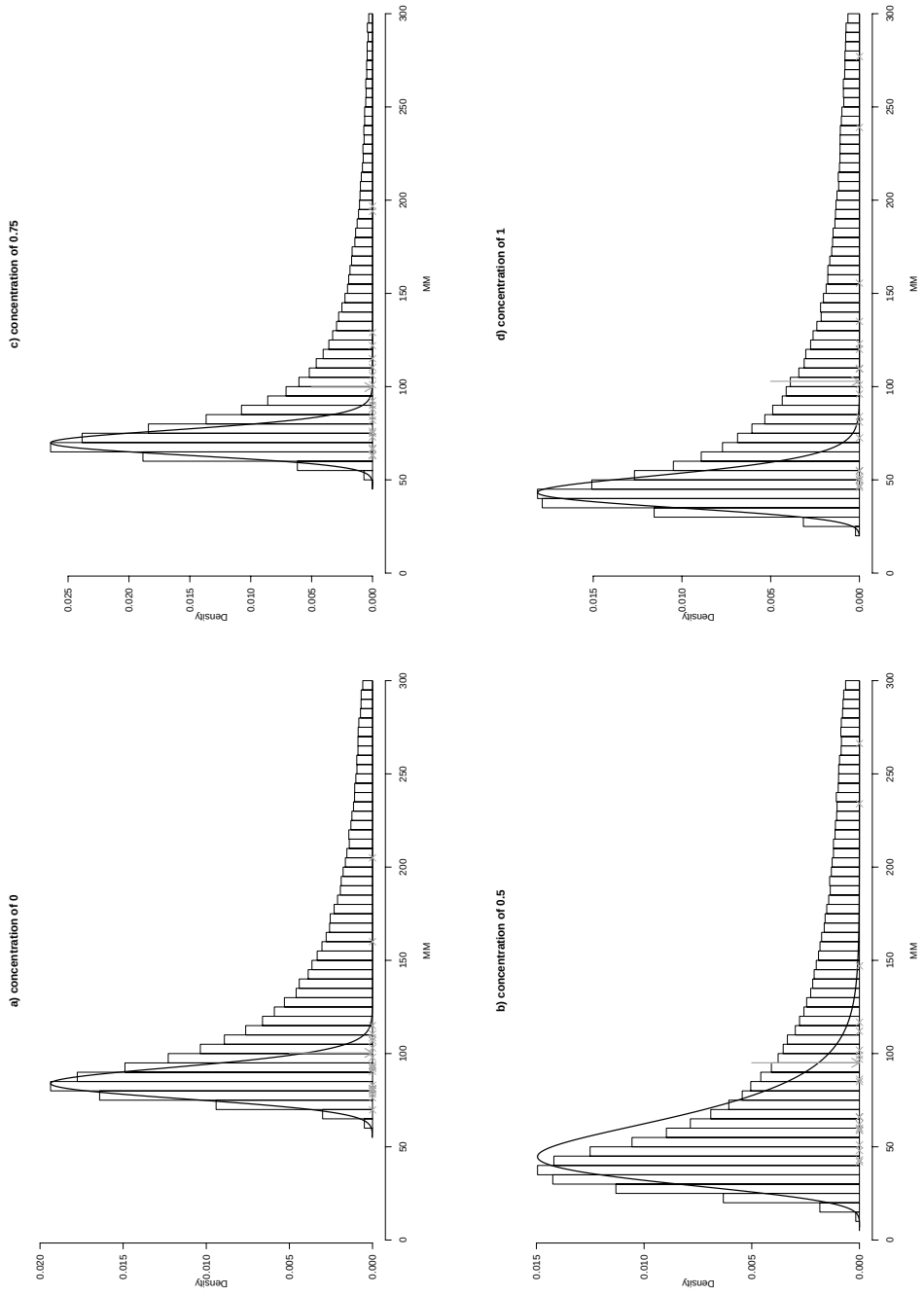

Fig. 5. Histograms of $\log_2(MM)$ for a array in which no probe-set was spiked along with the three arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 pM. The observed PM values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow. The black curve represents the log normal distribution obtained from left-of-the-mode data.

B1). All arrays in all data sets had similar shaped $\log_2(MM)$ histograms. Furthermore, the $\log_2(MM)$ histograms for the spiked-in probe set had similar histograms as well. The MMs to the left of the mode of the histogram can be approximated with the left-hand tail of a log-normal distribution. This suggests that the MMs are a mixture of probes for which (i) the intensities are largely due to non-specific binding and background noise and (ii) the intensities include transcript signal just like the PMs. The mode of the histogram is a natural estimate of the mean background level. 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, judging solely on the average, a difference would be hard to detect. However, distance of the average PM from the average background noise does in fact increase with concentration.

Figure 5 motivates a background plus signal model of the form $PM_{ijn} = bg_{ijn} + s_{ijn}$. Here bg_{ijn} represents background signal in array i caused by optical noise and non-specific binding. We assume each array has a common mean background level, $\mathrm{E}(bg_{ijn}) = \beta_i$. We want to adjust the PM intensities to remove the background effect. A naive approach is to consider $PM_{ijn} - \hat{\beta}_i$, with $\log_2(\hat{\beta}_i)$ the mode of the $\log_2(MM)$ distribution. An estimate of this distribution can be obtained using a density kernel estimate. In practice, a problem with this measure is that for a small percentage of probes $PM_{ijn} \leqslant \hat{\beta}_i$ and log transforming $PM_{ijn} - \hat{\beta}_i$ becomes a problem. An alternative background correction is to consider $B(PM_{ijn}) \equiv \mathrm{E}(s_{ijn}|PM_{ijn})$. If we impose a strictly positive distribution on s_{ijn}, then $B(PM_{ijn}) > 0$. To obtain a computationally feasible $B(\cdot)$ we consider the closed-form transformation obtained when assuming s_{ijn} is exponential and bg_{ijn} is normal. Although the data suggest that this model can be improved, the results obtained using $B(\cdot)$ work well in practice, as is demonstrated in the next section.

To obtain an expression measure we assume that for each probe set n, the background-adjusted, normalized, and log-transformed PM intensities, denoted with Y, follow a linear additive model

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \epsilon_{ijn}, i = 1, \ldots, I, j = 1, \ldots, J, n = 1, \ldots, n \tag{2}$$

with α_j a probe affinity effect, μ_i representing the log scale expression level for array i, 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 probe sets. This assumption is saying that Affymetrix technology has chosen probes with intensities that on average are representative of the associated genes expression. The estimate of μ_i gives the expression measures for probe set n on array i.

To summarize, in this paper we consider a new expression measure that (i) background-corrects the arrays using the transformation $B(\cdot)$, (ii) normalizes the arrays using quantile normalization, and (iii) for each probe set n, fits a linear model (2) to the background-corrected, normalized and log (base 2) transformed probe intensities denoted here with $Y_{ij}, i = 1, \ldots, I, j = 1, \ldots, J$. To protect against outlier probes we use a robust procedure, such as median polish (Holder *et al.*, 2001), to estimate model parameters. We use the estimate of μ_i as the log scale measure of expression which we refer to as robust multi-array average (RMA).

6. BIAS, VARIANCE, AND GOODNESS OF FIT COMPARISONS

Plots of log observed expression versus known concentration (not shown) demonstrate that the expression measures perform similarly in detecting the spiked-in probe sets. However, for the highest concentration, AvDiff and MBEI sometimes underestimate the predicted value from the known concentrations. This results from the attenuation caused by subtracting MM. We also notice that RMA is less noisy than all other measures at lower concentrations.

It is possible that the control genes used in data set B1 provide a stronger than usual signal. Therefore, a comparison based on all probe sets of the HG-U95A arrays is conducted using data set C. For these data

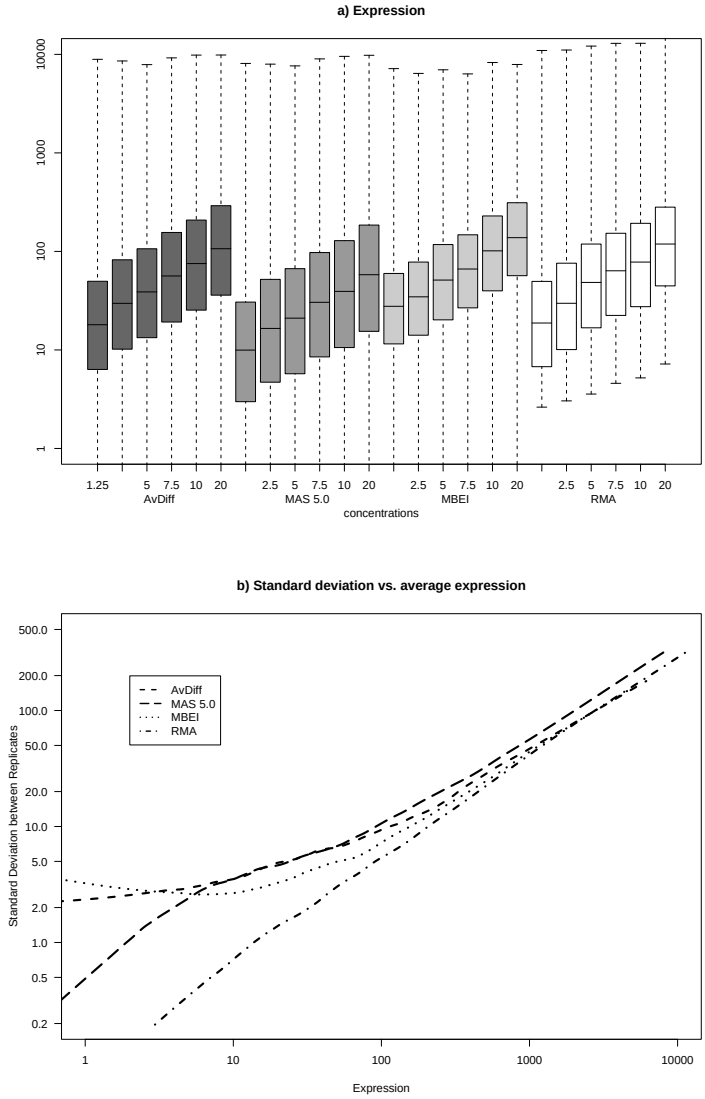

Fig. 6. Data set C boxplots. (a) Averages over replicates for each gene in (b). (b) Loess curves fitted to standard deviation versus average expression scatter-plots.

the amount of hybridization of probe sets representing expressed genes is expected to double when the amount of RNA hybridized to the array is double. Furthermore, the difference in gene expression across replicate arrays should be small.

For each of the four measures, we denote the expression values with $E_{ik}, i = 1, \ldots, 6, k = 1, \ldots, 5$ with i representing the dilution concentration level and k the replicate (which also identifies scanner). The averages are denoted with $E_{i\cdot} = (1/5)\sum_{k=1}^{5} E_{ik}$ and the SDs with $SD_i = \sqrt{(1/4)\sum_{k=1}^{5}(E_{ik} - E_{i\cdot})^2}$. Figure 6(a) shows boxplots of the $E_{i\cdot}$ for each dilution concentration i. Notice that all measures have

roughly the same ability to detect signal. Figure 6(b) shows loess curves fitted to the scatter plot (on the log scale) of SD_i vs $E_{i\cdot}$. Clearly, RMA has the smallest SD across replicates. The advantage of RMA is especially noticeable in the low expression values where the SD is 10 times smaller than the other measures.

Li and Wong's method 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 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 the SD_i defined above. If the model is close to the actual mechanism giving rise to the data, the nominal and observed SE should agree. Plots of nominal to observed SE log ratios versus expression (not shown) show that in general, RMA is closer to 0 than Li and Wong's MBEI showing that the observed and nominal standard error methods are, in general, closer when using (2) instead of (1).

7. DETECTION OF DIFFERENTIAL EXPRESSION

Data set B2 was used to assess how well the different expression measures perform at detecting differentially expressed probe sets. For each of the six arrays studied expression measures $E_{11n}, E_{12n}, E_{13n}, E_{21n}, E_{22n}, E_{23n}$ were obtained in their respective scale (log for MAS 5.0 and RMA) for each probe set $n = 1, \ldots, N$. We then computed the averages over triplicates $E_{i\cdot n} = (1/3)\sum_{k=1}^{3} E_{ikn}, i = 1, 2, n = 1, \ldots, N$. For the probe sets representing spike-in RNAs the observed ratios or 'fold changes' ($E_{2\cdot n}/E_{1\cdot n}$ for AvDiff and MBEI or $2^{E_{1\cdot n}-E_{2\cdot n}}$ for MAS 5.0 and RMA) should coincide with the true ratio of the spike-in concentrations shown in Table 1. Recall that apart from the spiked-in probe sets, the background samples hybridized to the six arrays are the same. We therefore expect only the 11 probe sets shown in Table 1 to be differentially expressed. In the left side of Figure 7 MVA plots of the average expressions obtained are shown. Probe sets with negative expression measures were left out for AvDiff and Li and Wong's MBEI. Notice that all measures separate 10 out of the 11 spiked-in probe sets from the cloud of points. However, the cloud of points for probe sets with small total intensity has a much larger spread for AvDiff, MBEI, and MAS 5.0 than for RMA. For this reason, many of the probe sets with high differential expressions for AvDiff, MBEI, and MAS 5.0 are not actually the spiked-in probe-sets. The smaller spread of RMA results in better detection of differentially expressed probe-sets. In the right side of Figure 7, quantile–quantile plots of the observed ratios are shown. RMA is the only measure to perfectly differentiate the spiked-in probe sets (with the exception BioC-3, which no measure was able to detect) from the rest. Table 1 shows the observed rank of the spiked-in probe sets.

8. CONCLUSION

In this paper we have developed a novel measure of gene expression and compared it to other standard measures. Through the analyses of three data sets, we have shown that expression is better measured using log-transformed PM values, after carrying out a global background adjustment and across-array normalization. We studied the performance of a version of the Affymetrix summary measures AvDiff and MAS 5.0, the Li and Wong model-based expression index, and the new measure RMA. We evaluated the four expression summary measures using spike-in and dilution study data, assessing their behavior in terms of bias, variance, the ability to detect known differential expression levels, and (for MBEI and RMA) model fit. We conclude that there is no obvious downside to summarizing the expression level of a probe set with RMA, and attaching an SE to this quantity using a linear model that removes probe-specific affinities. The greater sensitivity and specificity of RMA in detection of differential expression provides a useful improvement for researchers using the GeneChip® technology. We expect marginal though worthwhile gains to be achievable by using a more carefully designed and tested background correction procedure.

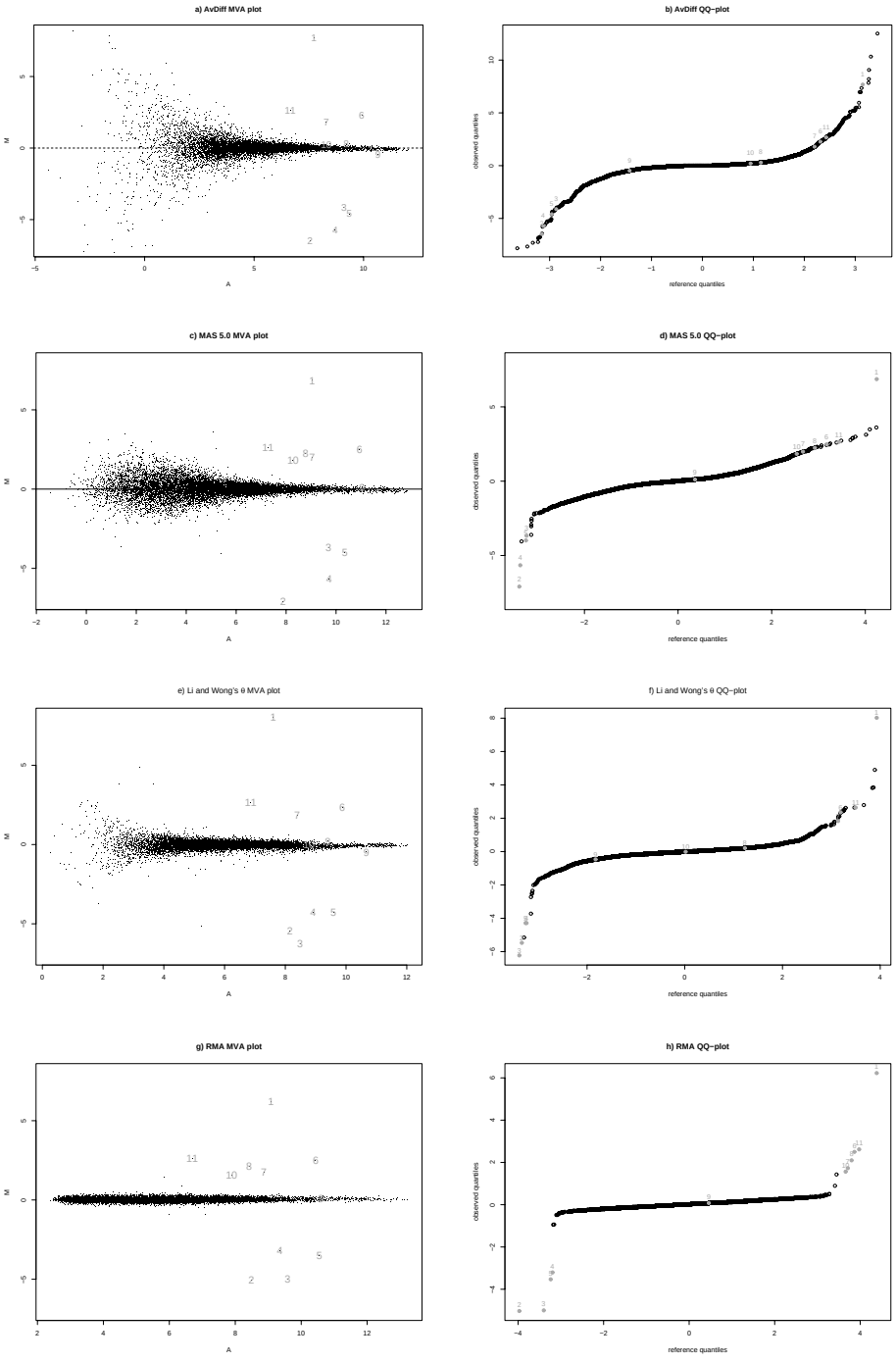

Fig. 7. MVA (described in text) and qq-plots indicating the positions of differentially expressed genes ranked by their absolute log relative expression values.

ACKNOWLEDGEMENTS

The work of Rafael A. Irizarry was supported by the PGA U01 HL66583. We would like to thank Skip Garcia, Tom Cappola and Joshua M. Hare from Johns Hopkins University for the mouse data and Gene Brown's group at Wyeth/Genetics Institute for helpful suggestions in the design of the spike-in experiment. We would like to thank Rehannah Borup and Eric Hoffman from the Children's National Medical Center Microarray Center for help obtaining the raw mouse data. We would like to thank Laurent Gautier from the Technical University of Denmark, Ben Bolstad from UC Berkeley and Magnus strand from Astra Zeneca Mölndal for developing and coding up the normalization routines. Finally, we thank Earl Hubbell (Affymetrix), Cheng Li (Harvard), the Associate Editor, and the referee for suggestions that have improved this paper.

REFERENCES

AFFYMETRIX (1999). *Affymetrix Microarray Suite User Guide, version 4 edition*. Santa Clara, CA: Affymetrix.

BAUGH, L., HILL, A., BROWN, E. AND HUNTER, C. P. (2001). Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Research* **29**, 1–9.

BOLSTAD, B., IRIZARRY, R., STRAND, M. AND SPEED, T. (2002). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, to appear.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2001). Statistical methods for identifying genes with differentialexpression in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

HARTEMINK, A. J., GIFFORD, D. K., JAAKOLA, T. S. AND YOUNG, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BiOS*.

HILL, A., HUNTER, C., TSUNG, B., TUCKER-KELLOGG, G. AND BROWN, E. (2000). Genomic analysis of gene expression in c. elegans. *Science* **290**, 809–812.

HILL, A. A., BROWN, E. L., WHITLEY, M. Z., TUCKER-KELLOGG, G., HUNTER, C. P. AND SLONIM, D. K. (2001). Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genomebiology* **2**, 1–13.

HOLDER, D., RAUBERTAS, R. F., PIKOUNIS, V. B., SVETNIK, V. AND SOPER, K. (2001). Statistical analysis of high density oligonucleotide arrays: a SAFER approach. *Proceedings of the ASA Annual Meeting 2001*. Atlanta, GA.

HUBBELL, E. (2001). Estimating signal with next generation Affymetrix software. *Gene Logic Workshop on Low Level Analysis of AffymetrixGeneChip® data.*
http://www.stat.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html.

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.

LOCKHART, D. J., DONG, H., BYRNE, M. C., FOLLETTIE, M. T., GALLO, M. V., CHEE, M. S., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H. AND BROWN, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675–1680.

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.

[*Received June 3, 2002; revised July 8, 2002; accepted for publication July 22, 2002*]