BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–10

A Benchmark for Affymetrix GeneChip Expression Measures

Leslie M. Cope[1], Rafael A. Irizarry[2], Harris A. Jaffee[2], Zhijin Wu[2] and Terence P. Speed[3]

[1]Department of Mathematical Sciences, Johns Hopkins University, 104 Whitehead Hall, 3400 North Charles Street, Baltimore, MD, 21218, U.S.A., [2]Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD, 21205, U.S.A. and [3]Department of Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA, 94720, U.S.A.

ABSTRACT

Motivation: The defining feature of oligonucleotide expression arrays is the use of several probes to assay each targeted transcript. This is a bonanza for the statistical geneticist, who can create probeset summaries with specific characteristics. There are now several methods available for summarizing probe level data from the popular Affymetrix GeneChips, but it is difficult to identify the best method for a given inquiry.

Results: We have developed a graphical tool to evaluate summaries of Affymetrix probe level data. Plots and summary statistics offer a picture of how an expression measure performs in several important areas. This picture facilitates the comparison of competing expression measures and the selection of methods suitable for a specific investigation. The key is a benchmark dataset consisting of a dilution study and a spike-in study. Because the truth is known for these data, we identify statistical features of the data for which the expected outcome is known in advance. Those features highlighted in our suite of graphs are justified by questions of biological interest and motivated by the presence of appropriate data.

Availability: In conjunction with the release of a graphics toolbox as part of the Bioconductor project (http://www.bioconductor.org), a webtool is available at http://affycomp.biostat.jhsph.edu. Supplemental material is available at http://www.biostat.jhsph.edu/~ririzarr/papers/suppaffycomp.pdf.

Contact: rafa@jhu.edu

INTRODUCTION

High density oligonucleotide array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. The defining feature of oligonucleotide expression arrays is the use of several probes to assay each targeted transcript. In order to obtain expression measures it is necessary to summarize the probe level data. This is a bonanza for the statistical geneticist, offering great opportunity to create probeset summaries with specific characteristics. On the other hand, the researcher with data in hand and a particular question in mind is not necessarily able to identify best method. Using a spike-in study prepared by Affymetrix and a dilution study by Gene Logic as benchmark data, we have developed a graphical tool for the evaluation and comparison of expression measures on the Affymetrix GeneChip platform [Lockhart et al. (1996)].

Plots and summary statistics offer a picture of how an expression measure performs in several important areas and facilitate the comparison of competing methods. The assessments evaluate performance in terms of bias (lack of accuracy) and variance (precision). The **R** package `affycomp`, available from the Bioconductor Project (www.bioconductor.org), can be used to automatically generate an *Image Report* showing all plots and summary statistics for one or several expression measures. A webtool that automatically generates Image Reports for single expression measures is also available at http://affycomp.biostat.jhsph.edu.

The benchmark data is crucial here. It is this data that turns a collection of plots and statistics into a genuine evaluative tool. The control of input in spike-in and dilution experiments makes it possible to identify features of the data for which the expected outcome is known in advance [Hill et al. (2001, 2000); Baugh et al. (2001)]. An expression measure can then be evaluated in terms of these features.

The individual plots and summary statistics included are not new. Similar or even identical descriptive methods have been used by a variety of researchers to evaluate measures of expression [Holder et al. (2001); Workman et al. (2002); Irizarry et al. (2003); Naef et al. (2001); Li and Wong (2001)]. Nor is the report comprehensive; we do not claim to have exhausted the possibilities contained even within this data and other benchmark data sets are available as well. Lemon et al. (2002), for example, use

data from mixture experiments prepared in their own lab to systematically evaluate specific characteristics of an expression measure. What is new is the proposal to standardize a subset of the commonly used evaluative tools in order to make results easier to produce and interpret. Each plot or statistic here focuses attention on a very specific problem in the analysis of expression data (e.g. measuring absolute abundance of RNA, or reducing error in replicate measurements of a single sample). The plots are intended to offer a fairly complete view of how a measure performs on these problems while summary statistics extract particular features of the plots to provide a convenient bottom-line.

First we describe the data sets and software used to create the Image Report. Then we describe the Image Report in detail. Two examples illustrate how these tools can be used: 1) We compare to MAS 5.0, the default measure available from Affymetrix and described in their manual, an expression measure developed by Li and Wong (2001) (dChip), and the robust multi-array analysis (RMA) expression measure developed by Irizarry et al. (2003). 2) We compare four versions of RMA to understand how normalization, robustness, multi-array analysis, and use of mismatch (MM) probes affect performance. We conclude with a discussion and a description of the webtool.

SYSTEMS AND METHODS

The benchmark data used is freely available to the research community. The assessment tools have been coded in the **R** statistical language [Ihaka and Gentleman (1996)], and are included in the `affycomp` package as part of the **Bioconductor Project**. The package is also the basis of a webtool that will automatically assess an expression measure and compare it to MAS 5.0.

MAS 5.0 results were obtained using Affymetrix software, and results for dChip were obtained using the official software release [Li and Wong (2001)]. All other analysis was performed in **R** using functions available in the **Bioconductor Project**.

For the dilution study by GeneLogic (http://qolotus02.genelogic.com/datasets.nsf/), two sources of cRNA, human liver tissue and central nervous system cell line (CNS), were hybridized to human arrays (HG-U95Av2) in a range of dilutions and proportions [Irizarry et al. (2003)]. We studied data from six groups of arrays that had hybridized liver and CNS cRNA at concentrations of 1.25, 2.5, 5.0, 7.5, 10.0, and 20.0 μg total cRNA. Five replicate arrays were available for each generated cRNA (n=60 total). Oligos from genes specific to foreign species were synthesized and added to each hybridization mixture at nominal amounts of 0.5, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, and 100 picoMolar. These oligos correspond to probe-sets: 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 (a B. subtilis gene). Oligos corresponding to the 3', middle and 5' end of each gene were synthesized and added separately. The same concentrations were used across all 60 arrays.

The spike-in study by Affymetrix (http://www.affymetrix.com/analysis/download_center2.affx) is a subset of the data used to develop and validate the MAS 5.0 algorithm. Human cRNA fragments matching 16 probe-sets on the HGU95A GeneChip were added to the hybridization mixture of the arrays at concentrations ranging from 0 to 1024 picoMolar. The same hybridization mixture, obtained from a common tissue source, was used for all arrays. The cRNAs were spiked-in at a different concentration on each array (apart from replicates) arranged in a cyclic Latin square design with each concentration appearing once in each row and column. The design is described in detail by Irizarry et al. (2003). A table describing the design is included in the supplemental material. A detailed tabular description is contained in the R package and in the webtool .

We are reporting 16 spiked-in probesets as opposed to the 14 originally described by Affymetrix. These extra spike-ins have been reported by various researchers, for example Wolfinger and Chu (2002). We claim that the probeset with ID 33818_at should be included as the transcript in the 12th column of Latin square design. In the excel file describing the Latin square, provided by Affymetrix, probeset 407_at, which is in the 12th column, actually has the same concentration pattern as the transcript in column 1, 37777_at. No spike-in gene given by Affymetrix has the pattern consistent with the 12th column of the Latin square design. However, expression measures obtained with MAS 5.0, dChip and RMA for probeset 33818_at follow the missing pattern. We also claim probeset 546_at should be considered with same concentration as 36202_at, because it is designed against the same target: Unigene ID Hs. 75209. The data is consistent with this as 546_at shows the same pattern as 36202_at. Wolfinger and Chu (2002) identified these two probes as well. Wolfinger and Chu (2002) claim 1598_g_at and 37658_at have the same pattern as 1597_at and that 1032_at has the same pattern as 684_at. However, for these three probesets there is no supporting evidence apart from data. Furthermore, these patterns are seen with dChip but not with RMA and MAS 5.0. We therefore do not include the these as spike-in probesets.

THE IMAGE REPORT

Three guiding principles determined what we included in the image report. 1) Each plot or statistic must be motivated by the benchmark data (i.e. must evaluate

performance on a task for which the expected outcome is known), 2) each must be justified by biological interest or statistical principle, and 3) each must facilitate comparison of competing expression measures or offer insight into appropriate analysis of data from a single given measure.

The basic Image Report includes 6 Figures, described below, and a table of summary statistics for a single measure of expression. These plots and statistics address five main issues in the analysis of expression array data 1) variability of expression across replicate arrays 2) response of expression measure to changes in abundance of RNA 3) sensitivity of fold-change measures to amount of actual RNA sample 4) accuracy of fold-change as a measure of relative expression 5) usefulness of raw fold change score for the detection of differential expression

When multiple expression measures are to be compared directly, a comparative Image Report can be generated. Small differences between the basic and comparative plots are mentioned in the descriptions below. The webtool and the R package alike can readily generate either report. Examples of these reports can be found on the webtool webpage. In the next section, we describe two applications where the image report is useful. All examples of figures included here are from those applications.

All data is plotted on the $\log_2$ scale. The log transformation is made at the beginning and all intermediate transformations are made to the logged data. Fold changes are calculated for single chip comparisons, even where replicates are available. This standardizes the procedure for those cases in which replicates are not available, and gives results that depend as little as possible on the the specific design of the benchmark datasets.

Notation

The spike-in data set encompasses more than one set of experimental conditions. Within each experiment, only the spike-in concentrations are varied; background is the same for all arrays. Fold change calculations are always made within experiment to ensure that only spiked-in genes will be differentially expressed. We refer to spike-in concentrations as nominal concentration and their ratios as nominal fold change as a reminder that small errors in the actual quantity of RNA in the sample prevent us from knowing the true concentrations. We will use M to denote log observed fold change and FC as an abbreviation of fold change. Notation for formulas is as follows:

Dilution study: Let y_{tdrg} represent $\log_2$ expression for tissue $t = 1, 2$, dilution $d = 1, \ldots, 6$, replicate $r = 1, \ldots, 5$ and gene $g = 1, \ldots, n$.

Spike-in study: A different variable is used to avoid confusion. Let x_{ecg} represent $\log_2$ expression for experiment $e = 1, 2, 3$, array type $c = 1, 2, \ldots, 20$ and gene $g = 1, \ldots, n$. For the spike-in genes only, χ_{cg} represents the $\log_2$ of the nominal concentration.

The Plots and Statistics

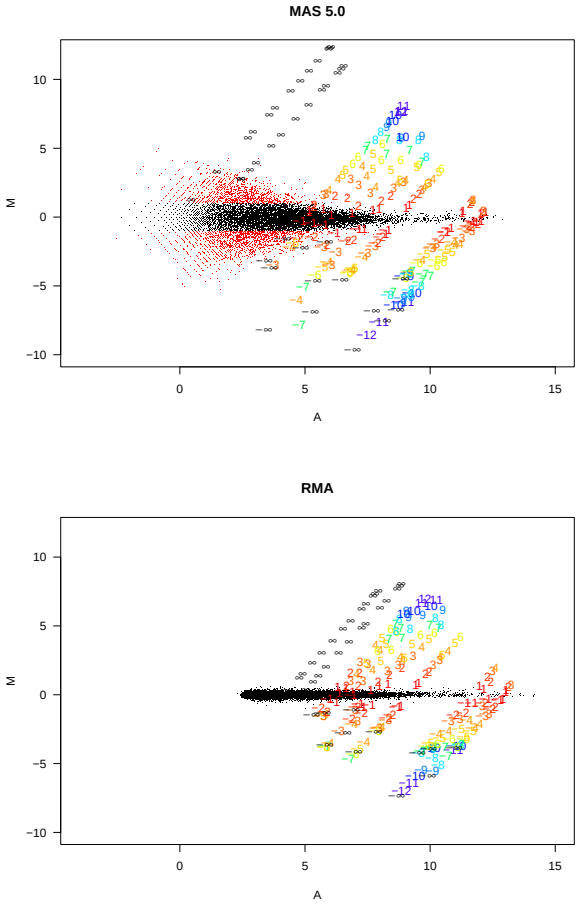

Fig. 1. The MA plot shows log fold change as a function of mean log expression level. A set of 14 arrays representing a single experiment from the Affymetrix spike-in data are used for this plot. A total of 13 sets of fold changes are generated by comparing the first array in the set to each of the others. 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 red. All other probesets are represented with black dots.

1) MA plot: The MA plot shows log fold change as a function of mean log expression level. This has come to be a fundamental graphical tool for the analysis of expression array data. A great deal of information about the distribution of observed fold changes can be read from such a plot. A set of 14 arrays representing a single experiment from the Affymetrix spike-in data are used for this plot. We choose 14 so that all possible combinations of pairs of concentrations appear once. A total of 13 sets of fold changes are generated by comparing the first array in

the set to each of the others, so that $M_{1cg} = x_{11g} - x_{1cg}$ for $c = 2, \ldots 14$. These are plotted together against the mean values $A_{1cg} = (x_{11g} + x_{1cg})/2$. To make the plot more informative, spiked-in 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 red. All other probesets are represented with black dots.

MA plots for competing methods are presented side by side on separate axes. Figure 1 shows a comparative plot for two expression measures.

2) Variance across replicates plot: The variance of an expression measure across replicate arrays should be low. The GeneLogic dilution data set includes 5 replicate arrays under each experimental condition. For each gene, and each experimental condition, we calculate the mean log expression $y_{td\cdot g}$ and plot this against the observed standard deviation $s_{tdg} = \sqrt{\sum_r (y_{tdrg} - y_{td\cdot g})^2/4}$ of the replicate arrays. The resulting scatterplot is smoothed to generate a single curve representing mean standard deviation as a function of mean log expression. Standard deviation should be low and independent of expression level.

Curves from competing methods are plotted on the same set of axes as seen in Figure 2. A summary of the information in this plot includes the median standard deviation and the average R^2 between replicates, which are the first two entries in Table 1.

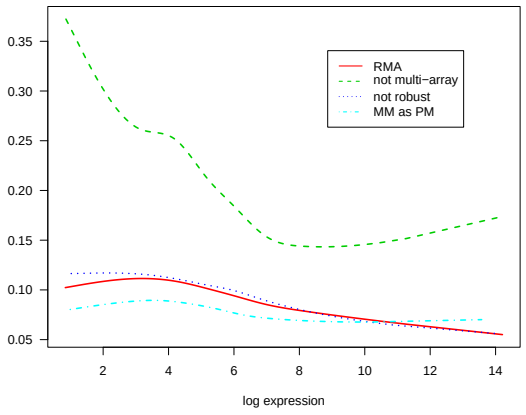

Fig. 2. For each gene, and each experimental condition, we calculate the mean log expression and the observed standard deviation across 5 replicates. The resulting scatterplot is smoothed to generate a single curve representing mean standard deviation as a function of mean log expression.

3) Sensitivity of expression ratios to total quantity of RNA plot: The total quantity of RNA in the hybridization solution will vary somewhat from experiment to experiment. Expression ratio estimates (fold changes) are relative and should not co-vary with RNA quantity. To simulate extreme variation in total quantity of RNA, the lowest concentration at $1.25\mu g$ and the highest at $20\mu g$ are used. Observed expression is first averaged across replicates to obtain a single mean value $y_{td\cdot g}$ for each tissue and dilution. The log expression ratio between liver and CNS samples $M_{dg} = y_{1d\cdot g} - y_{2d\cdot g}$ is calculated within each dilution level and for every gene on the array. Finally M_{1g} is plotted against M_{6g} to produce a scatterplot with one point for each gene. Orange and red color is used to denote genes with $M_{6g} - M_{1g}$ bigger than $\log_2(2)$ and $\log_2(3)$ respectively. The rest are denoted with black (default foreground color) points. (This plot is not shown in this paper)

The comparative plot is different. The difference between the two log expression ratios is calculated for each gene on the array and these quantities are represented in side by side boxplots, as seen in Figure 3.

Summary statistics include the correlation between fold change measurements and the total number of genes showing greater than 2 fold and 3 fold difference in log expression ratio. These are entries three to five in Table 1.

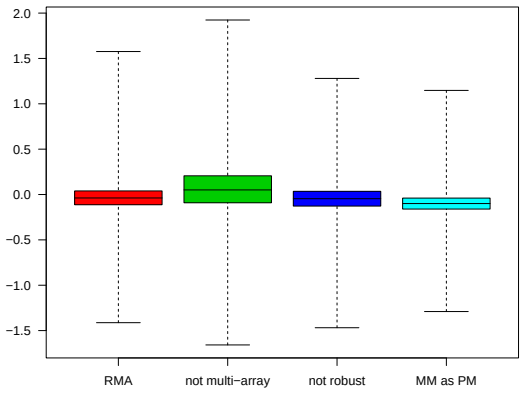

Fig. 3. This plot, using the GeneLogic dilution data, shows the sensitivity of fold change calculations to total RNA abundance. Average log fold-changes between liver and CNS for the lowest concentration and the highest in the dilution data set are computed. The boxplots show the distribution of the differences between these log fold-changes.

4) Observed expression v. nominal expression plots: The fundamental biological interpretation of expression values is that they are measures of RNA abundance. In addition to evaluating expression measures in this regard, these plots are valuable partners to the variance of replicates plot. Reduction in variance should not be obtained at the cost of a reduction in signal detection. Attenuation of signal is recognizable here.

a) Spike-in data Plot: The entire Affymetrix spike-in dataset is used for the first of these plots. Observed concentration x_{ecg} is plotted against nominal concentration =

χ_{ecg} for each spiked-in gene. The $\log_2$ scale is especially useful here. A difference of 1 on this scale represents a doubling of RNA concentration. Ideally, if the nominal concentration doubles, so should the observed concentration. On the $\log_2$ scale then, observed concentration should be linear in true concentrations with a slope of 1.

When competing methods are to be compared, the observed intensities are averaged at each nominal concentration value, resulting in a single mean curve. The curves for competing methods are plotted on a single set of axes as seen in Figure 4a.

We fit a simple linear model to the scatterplot data and report the estimated slope and R^2 coefficient in the table of statistics. These are the sixth and seventh entries in Table 1.

b) Dilution Data Plot: The second plot uses the Gene Logic dilution data. Unlike the spike-in experiment the absolute abundance of no gene is known in this study. However the relative abundance of each gene changes predictably with the dilution. If the dilution is halved then the relative abundance of each gene should double. On a $\log_2$ scale, for expressed genes, observed abundance should be a linear function of the inverse of the dilution with a slope of 1. For this plot, replicates are first averaged to obtain a mean expression value $y_{td\cdot g}$ for each gene in each dilution and tissue. These values are regressed against the inverse of dilution amount to obtain an estimate of slope S_g for each gene. The values S_g are normalized-corrected using the probesets that were spiked-in these samples (see R code for details) and plotted against the mean log expression for each gene $y_{\cdots g}$.

For comparative plots, the scatterplots are smoothed to obtain a single curve for each expression method. These are plotted together on a single set of axes. Figure 4b shows an example.

The overall median of slope values is reported in the eighth entry of Table 1.

5) ROC curves: One of the chief uses of expression arrays is the identification of genes that express differently under various experimental conditions. A typical identification rule filters genes with fold change exceeding a given threshold. Receiver Operator Characteristic (ROC) curves offer a graphical representation of both specificity and sensitivity for such a rule. ROC curves are created by plotting the true positive (TP) rate (sensitivity) against false positive (FP) rate (1-specificity) obtained at each possible threshold value. We present two ROC plots, both using log fold change as a filter. In each case the Affymetrix spike-in dataset is used. In this paper we use absolute TP and FP instead of rates because it is easier to interpret. Since only spike-in genes are actually differentially expressed in these experiments, it is easy to determine TP and FP.

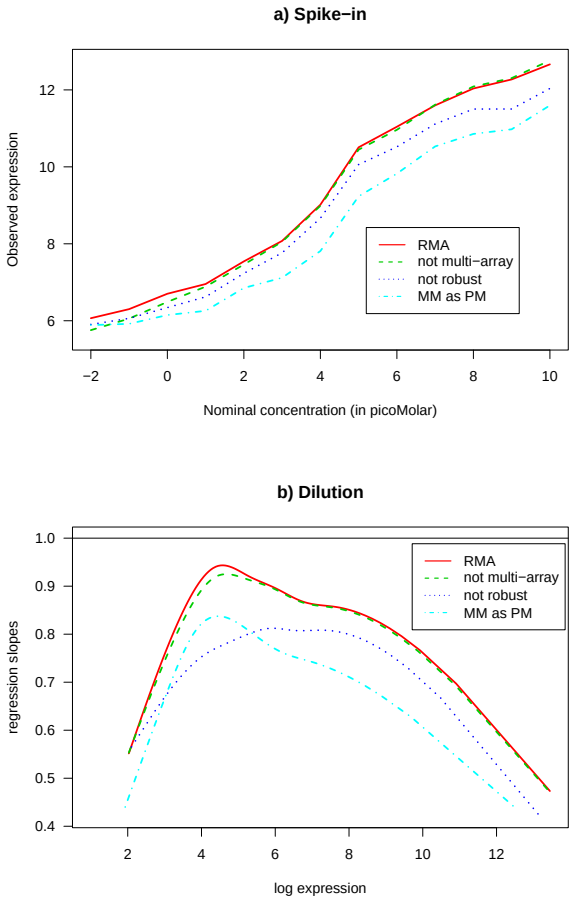

Fig. 4. a) Average observed $\log_2$ intensity plotted against nominal $\log_2$ concentration for each spiked-in gene for all arrays in Affymetrix spike-In experiment. b) For the GeneLogic dilution data, log expression values are regressed against their log nominal concentration. The resulting slope estimates are plotted against average log intensity across all concentrations. Smooth curves are fitted and shown.

a) ROC curve, general FC: For the first plot, log fold changes $M_{eijg} = x_{eig} - x_{ejg}$ are calculated for every gene and for arrays i and j where $i < j$. For every pair of arrays we order the probesets by the observed absolute value of their log ratio. We can then go through this list and count the number of TP we get for every possible value of $FP = 0, 1, 2, \ldots, 12609$ (12609 is the total number of non-spiked in probesets). Notice that with this information we can create a separate ROC curve for each pair of arrays. To form an *average* ROC curve we compute the average TP across comparisons for each FP value. This creates a single average ROC curve. Average TP is plotted against FP up to a maximum of 100 false positives.

b) ROC curve: FC=2 In this plot attention is restricted to arrays in which all nominal fold changes are equal to 2. This is the lowest nominal fold change in the dataset. Otherwise, procedures are the same as above.

m ROC curves for competing methods are plotted together on a single set of axes. Figure 5 is an example showing both general FC and FC=2 plots.

The area under the curve (AUC) is probably the most common summary statistic for a ROC curve. However, because in practice we rarely validate more the 100 genes we report as a summary statistic the AUC up to 100 FP. We standardize so that the largest possible value is 1. Filtering on a fold change of 2.0 has become almost standard in expression array analysis. As a service to the analyst, we report the average number of true and false positives obtained using this particular filter. In Table 1, these are entries 9, 10, and 11 respectively for general FC and 12, 13 and 14 for FC=2.

6) Observed fold change v. nominal fold change plots: Exploratory studies often leave an investigator with a large list of potentially interesting genes. Validation is expensive and time consuming and appropriate prioritization can reduce waste significantly. It is hoped that the largest observed fold changes indeed correspond to the largest actual fold changes.

a) FC plot, general: In the first plot, log fold changes $M_{eijg} = x_{eig} - x_{ejg}$ are calculated for each spiked-in gene and for arrays i and j where $i < j$. Likewise the nominal log fold changes $\chi_{ig} - \chi_{jg}$, are calculated and these values are plotted against one another. The resulting scatterplots should be generally linear, with low variability and a slope of 1. A unique color and symbol is used for each gene in the plot. Horizontal dashed lines show quantiles of the fold changes observed for non-differentially expressed genes.

b) FC plot, close-up of low concentration: The second plot is identical except that here we use a subset of the data in which nominal concentration is no higher than 2 picoMolar.

Plots for competing methods are presented side by side on separate axes as seen in Figure 6.

We report the interquartile range of the log fold changes observed for non-differentially expressed genes. We also fit a simple linear model to the data in each scatterplot and report the estimated slope. The IQR can be seen in entry 15 of Table 1. The slope are in entries 16 and 17 for the general assessment and low concentration close-up respectively.

An additional plot is included in the package, but not in the webtool.

7) Predicted variability plot: Probeset summary methods sometimes include an estimate of standard error for each expression value. Using the replicates from the dilution data, we calculate the mean predicted variance for each gene, tissue and dilution by squaring the estimated standard error. The usual sample variance $s^2_{tdg} = \sum_r (y_{tdrg} - y_{td\cdot g})^2/4$ are calculated as well. A boxplots of the log ratios of the predicted and observed variance is used as an assessment and can be seen in the supplemental material. The correlation between these two is computed in the `affycomp` package assessment.

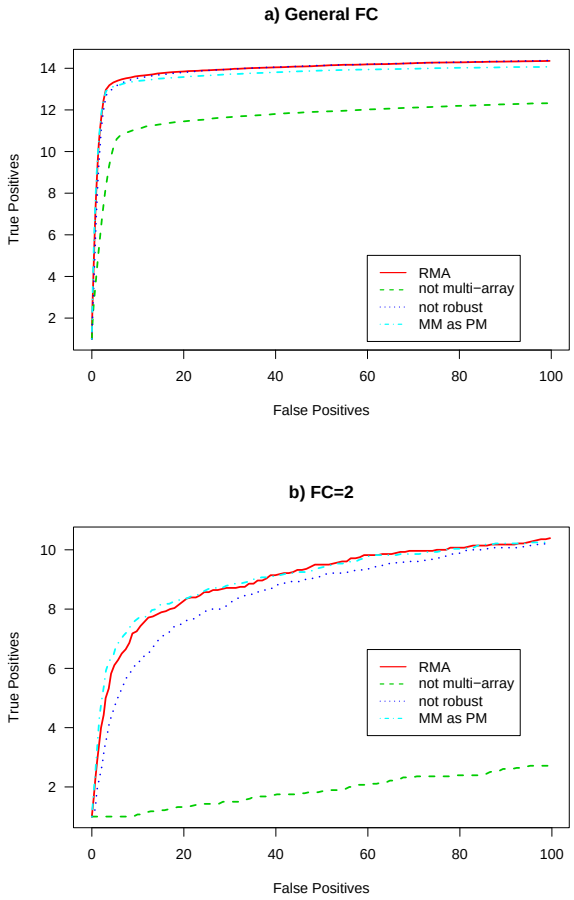

Fig. 5. A typical identification rule for differential expression filters genes with fold change exceeding a given threshold. This figure shows average ROC curves which offer a graphical representation of both specificity and sensitivity for such a detection rule. a) Average ROC curves based on comparisons with nominal fold changes ranging from 2 to 4096. b) As a) but with nominal fold changes equal to 2.

IMPLEMENTATION

The main purpose of the assessment tools is the direct comparison of expression measures. Using the R package as well as the webtool, one can readily compare two or more expression measures. These comparisons can

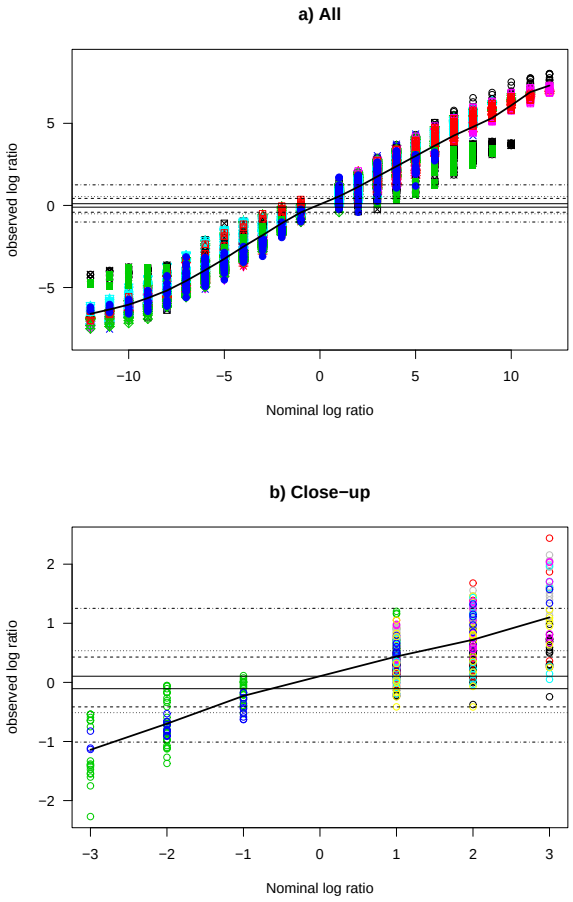

Fig. 6. a) For RMA, observed log fold changes plotted against nominal log fold changes. The dashed lines represent highest, 25th highest, 100th highest, 25th percentile, 75th percentile, smallest 100th, smallest 25th, and smallest log fold change for the genes that were not differentially expressed. b) Like a) but the observed fold changes were calculated for spiked in genes with nominal concentrations no higher than 2 picoMolar.

be used to decide which expression measure is most appropriate for a particular task. In illustration, Table 1 shows the summary statistics obtained from comparing the Affymetrix default, an expression measure developed by Li and Wong (2001) (dChip), and the robust multi-array analysis (RMA) expression measure developed by Irizarry et al. (2003). The intention in introducing these methods here is to illustrate the assessment tools and to demonstrate their uses. For a detailed comparison of these three measures see Irizarry et al. (2003).

The table shows that both RMA and dChip add bias to signal estimates compared to MAS. However, the reduction in variance obtained by either of these methods is large enough to offset the small increase in bias. Notice in particular that although the bias in log fold change estimates is about 10% greater for RMA than for MAS (statistic 16), variability of the same estimates (measured using IQR in statistic 15) is about 9 times larger for MAS than for RMA. The area under the ROC curve are much larger for dChip and RMA than MAS 5.0 showing these expression measures are superior for detection of differential expression using raw fold change. A detailed comparison of these three measures is given in Irizarry et al. (2003). MAS 5.0 does not provide a standard error to go along with their expression measure. However, dChip does. Using a simple analysis of variance approach one can also obtain standard errors for RMA. The supplemtnal material shows a comparison of dChip and RMA standard errors using the Predicted variability plot.

We now demonstrate another application of the assessment tools. Among RMA's features are multi-array normalization of probe level data, and a robust multi-array procedure for summarizing the probe level data. We are interested in determining the impact that these features have on expression values. We therefore designed two variations on RMA, in which each of these features is left out in turn. We consider a third expression measure, identical to RMA but that it treats the MM probes as if they were additional PM probes. The three measures are: 1) The *not multi-array* measure ignores the multi-array normalization and simply computes the median of the background corrected PM values on each array. Scalar normalization, like that used by MAS 5.0, is performed so that all arrays have the same median value. 2) The *not robust* measure performs the multi-array normalization and then summarizes the probe level data by simply taking the average of the $\log_2$ intensities. 3) The *MM as PM* measure background corrects and normalizes the MMs in the same way as the PMs and then includes them as if they were PMs in the robust summary step.

Table 1 demonstrates the differences between these measures. Robust and non-robust methods perform quite similarly. This is not surprising, given the large number of chips in these experiments. The robust method was developed specifically to handle much smaller experiments, and has been demonstrated to show greater advantage under those circumstances. The single array method results in the least biased measurements of RNA abundance and of log fold change. However, the variance of both observed intensity and observed fold change are higher than by any other method. The reduction in bias in not enough to offset the increase in variance and the overall performance of the method suffers. When MM probes are treated as additional PM probes, variance notably improves. There is a significant increase in bias in the estimation of RNA abundance, and results on bias in the estimation of fold change are mixed.

DISCUSSION

The techniques reviewed in this paper seek only to assess measurements of gene expression in terms of accuracy (low bias) and precision (low variance). It is straightforward to estimate the variance of a measurement process: we simply need replicate measurements on the same samples. On the other hand, to estimate bias in measurements, we need the truth, in an absolute or relative form, or at least a different, more accurate measurement of the same samples. We are fortunate to have spike-in and dilution data sets, where the truth is known to the extent that dilution errors can be ignored. In addition there are other experiments, such as the GeneLogic's mixture experiment, and Wright and Lemon's mixture experiment [Lemon et al. (2002)], which provide further opportunities to assess the bias and variance of measurement procedures.

Two important issues we don't address in this paper are the detection of gene expression, i.e. determining whether a particular mRNA transcript is present or absent in a sample of transcripts, and the detection of differential expression, i.e., determining whether or not a given transcript, or which of a given class of transcripts, is present at different concentrations in two mRNA samples. Both these issues require different types of assessments, and discussing them here would take us too far afield. We do assess the accuracy and precision of relative expression of specific genes. There is no contradiction here, as the problem of estimating relative expression is not the same as deciding whether differential expression is taking place.

Let us note that all assessments of bias include measurement variability, so that while variance can be assessed without reference to bias, we cannot assess bias without reference to variance. Furthermore, we provide different assessments of bias and variance as they relate to different applications so that users of the technology can decide which measure better suits their specific purpose.

Figure 1 is an MA plot comparing observed and predicted relative expression values, and so gives a snapshot of the combined consequences of bias and variance in estimates. The diamond shape of the points is a straightforward consequence of the fact that the measurements on each chip have lower and upper bounds, and when these are differenced and averaged, these bounds translate into the linear constraints seen. Although the points here include both bias and variance components, one can see the bias quite clearly.

By contrast, Figure 2 refers only to variances, making use of replicate data on each of many thousands of genes. Figure 3 is another look at variance, this time comparing expression measurements when there is quite a lot more or less than the recommended amount of mRNA in the experiment. Both Figures 2 and 3 are straightforward to interpret.

By making use of known mRNA concentrations of the spiked-in transcripts, Figure 4 allows us to assess bias in our measurements, although necessarily with variance present.

In Figure 5 we focus on the question of identifying differential expression. However, we only use a log fold-change rule to compare expression summaries, and do not consider more sophisticated approaches such as using t or moderated t-statistics. This is because our aim is the comparison of bias and variance of measurement processes, not algorithms for the specific task of identifying differential expression. Thus we offer two ROC curves, one comprehensive, and one focused on genes whose nominal change is 2-fold.

Finally, Figure 6 compares observed log fold change to the known values for the spiked-in genes, and is again an assessment of bias, with variance inevitably present.

The displays we offer permit both absolute and comparative analyses, although in this context the term absolute must be qualified. A hybridization-based quantification of transcript abundance can only be up to an undetermined constant. This shows up as an undetermined intercept in plots such as Figure 6.

Some specific conclusions using our graphics toolbox now follow. One is that using MM intensities as we do PM intensities in RMA leads to expression measures and fold change estimates with less variance. Also bias in fold change estimates decreases slightly , as assessed by the ROC curve and the AUC. We cannot say that MM values are of no use, although we have not yet seen a demonstration of their effectiveness in carrying out the original task of adjusting PM values. It does seem clear that RMA estimates have considerably less variance than those from MAS 5.0. We also see that robust and multi-array aspects of RMA provide important improvements in various aspects. Using quantile normalization instead of scaling normalization also proves to be important.

Where do we go next? There is clearly a need for more data sets of the kind used on this paper. An ever-present issue in this kind of algorithm development and tuning is the need to avoid over-dependence on particular data sets, a phenomenon called "over-training". We need a wide range of suitable data sets, so that training (perhaps better called calibrating in this context) and testing can be done on widely differing data. Only then can we comfortably extrapolate the conclusions from benchmark data sets to general use.

It is easy to identify places at which the assessment tools fall short of ideal.

A related need is for a framework such as ours for comparing different platforms used for measuring gene expression. These include the different short and long oligonucleotide and cDNA microarrays, SAGE, quantitative RT-PCR, and other techniques. There are already a number of

publications addressing these issues, see e.g. Yuen et al. (2002) and Barczak et al. (2002) and references therein, but we are still far from having comprehensive data (such as the spike-in data sets described here) or a framework for comparing platforms.

The Webtool

We invite all interested parties to put their probe summary methods to the test in a friendly competition. Download the benchmark data and develop one or more probe summaries. Return the expression-level data and we'll tell you how you did on this set of tasks. The goal is threefold. In addition to vetting the toolbox and competing for bragging rights, this will be an opportunity to systematically examine the strengths and weaknesses of the various approaches to probeset summary. Existing expression measures have proven very effective, but a great deal more improvement is possible.

ACKNOWLEDGEMENTS

The work of Leslie Cope, Rafael Irizarry, and Harris Jaffee is funded by the Hopkins PGA (www.hopkins-genomics.org) Administrative/Bioinformatics Component (P01 HL 66583). The authors would like to thank Muralidhar Bopparaju, Mohan Parigi, Jiong Yang, and Gerald G. Gilyeat for help with the webtool.

REFERENCES

Barczak, A., M. W. Rodriguez, K. Hanspers, L. L. Koth, Y. C. Tai, B. M. Bolstad, T. Speed, and D. J. Erie (2002). Spotted oligonucleotide arrays for human expression analysis. *Genome Research.* In press.

Baugh, L., A. Hill, E. Brown, and C. Hunter (2001). Quantitative analysis of mrna amplification by in vitro transcription. *Nucleic Acids Res 29*, 1–9.

Hill, A., E. Brown, M. Whitleyn, G. Tucker-Kellogg, C. Hunter, and D. Slonim (2001). Evaluation of normalization procedures for oligonucleotide array data based on spiked crna controls. *Genome Biology 2*.

Hill, A., C. Hunter, B. Tsung, G. Tucker-Kellogg, and E. Brown (2000). Genomic analysis of gene expression in c. elegans. *Scienc 290*, 809–812.

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

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5*, 299–314.

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

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research 31*.

Lemon, W., J. Palatini, R. Krahe, and F. Wright (2002). Theoretical and empirical comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics 18*(11), 1470–1476.

Li, C. and W. Wong (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., 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 (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology 14*, 1675–1680.

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

Wolfinger, R. and T.-M. Chu (2002). Who are those strangers in the latin square? *Critical Assessment of Microarray Data Analysis "CAMDA 02"*.

Workman, C., L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielsen, H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen (2002). A new non-linear normalization method for reducing variability in dna microarray experiments. *Genome Biology 3*.

Yuen, T., E. Wurmbach, R. Pfeffer, B. Ebersole, and S. Sealfon (2002, May). Accuracy and calibration of commercial oligonucleotide and custom cdna microarrays. *Nucleic Acids Research 30*(10), e48.

Table 1. Assessment summary statistics table. The second column denotes the Figure to which the summary statistic relates. Columns 3,4 and 5 compare MAS 5.0, dChip, and RMA. The statistics are described in the text. For each row, the best performing expression measure is denoted with a bold number. Columns 6, 7, and 8 compare RMA to alternatives based on RMA. For each row, if the best performing expression measure is not RMA it is denoted with a bold number.

Assessment	Figure	MAS 5.0	dChip	RMA	not multi-array	not robust	MM as PM
1) Median SD	2	0.29	0.089	**0.088**	0.19	0.092	**0.074**
2) R2	2	0.89	0.99	**0.99**	0.98	0.99	**0.99**
3) 1.25v20 corr	3	0.73	0.91	**0.94**	0.87	**0.94**	0.93
4) 2-fold discrepancy	3	1200	40	**21**	99	12	**6**
5) 3-fold discrepancy	3	330	8	**0**	12	**0**	**0**
6) Signal detect slope	4a	**0.71**	0.53	0.63	**0.65**	0.59	0.55
7) Signal detect R2	4a	**0.86**	0.85	0.8	**0.81**	0.76	0.72
8) Median slope	4b	0.85	0.77	**0.87**	0.86	0.79	0.76
9) AUC (FP$<$100)	5a	0.36	0.67	**0.82**	0.69	0.82	0.81
10) AFP, call if fc$>$2	5a	3100	37	**16**	220	19	**15**
11) ATP, call if fc$>$2	5a	**13**	11	12	**12**	12	11
12) FC=2, AUC (FP$<$100)	5b	0.065	0.17	**0.54**	0.12	0.52	**0.55**
13) FC=2, AFP, call if fc$>$2	5b	1400	12	**0.5**	18	**0.5**	**0.5**
14) FC=2, ATP, call if fc$>$2	5b	**3.7**	1.3	1.7	**2.3**	1.4	1.4
15) IQR	6	2.7	0.45	**0.31**	0.67	0.31	**0.25**
16) Obs-intended-fc slope	6a	**0.69**	0.52	0.61	**0.64**	0.58	0.54
17) Obs-(low)int-fc slope	6b	**0.65**	0.32	0.36	**0.45**	0.34	0.21