BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 7 2006, pages 789–794
doi:10.1093/bioinformatics/btk046

*Gene expression*

# Comparison of Affymetrix GeneChip expression measures

Rafael A. Irizarry[1,*], Zhijin Wu[2] and Harris A. Jaffee[1]

[1]Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA and [2]Center for Statistical Sciences, Department of Community Health, Brown University, 167 Angell Street, BOX G-H, Providence, RI 02912, USA

Received on August 25, 2005; revised on and accepted on January 5, 2006
Advance Access publication January 12, 2006
Associate Editor: Chris Stoeckert

**ABSTRACT**

**Motivation:** In the Affymetrix GeneChip system, preprocessing occurs before one obtains expression level measurements. Because the number of competing preprocessing methods was large and growing we developed a benchmark to help users identify the best method for their application. A webtool was made available for developers to benchmark their procedures. At the time of writing over 50 methods had been submitted.

**Results:** We benchmarked 31 probe set algorithms using a U95A dataset of spike in controls. Using this dataset, we found that background correction, one of the main steps in preprocessing, has the largest effect on performance. In particular, background correction appears to improve accuracy but, in general, worsen precision. The benchmark results put this balance in perspective. Furthermore, we have improved some of the original benchmark metrics to provide more detailed information regarding precision and accuracy. A handful of methods stand out as providing the best balance using spike-in data with the older U95A array, although different experiments on more current arrays may benchmark differently.

**Availability:** The `affycomp` package, now version 1.5.2, continues to be available as part of the Bioconductor project (http://www.bioconductor.org). The webtool continues to be available at http://affycomp.biostat.jhsph.edu

**Contact:** rafa@jhu.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

The development of preprocessing methodology for Affymetrix GeneChip has become an active research field. Various alternative procedures are available and new ones are being developed. Conflicting reports have been published comparing the more popular methods. Furthermore, developers of new methods usually find a way to claim overall superiority. It is common to see different papers using different assessment data and/or assessment statistics. To help users of the technology make sense of the discrepancy found in the literature and to help them identify the best method for the particular task, Cope *et al*. (2004) developed a benchmark. A webtool implementing this benchmark made it possible to compare all methods using the same assessment data and summary statistics/plots. Since its inception in the summer of 2003 developers and users have submitted more than 30 methods. Table 1 gives a brief overview of the main methods being compared. More details are available in Supplementary Table 1. Alternative versions of these methods have also been submitted and are described in Supplementary Table 2. Throughout the paper we will use the nicknames shown in the first column of this table to denote the different methods. Columns 2, 3 and 4 contain descriptions of the three main preprocessing steps: background correction, normalization and summarization. Because various methods combine two or more of these steps under one unified methodology, some of the columns are merged to described these. Notice that background correction can be global and/or probe-specific. These distinctions are made within the table cells. In column 5 we provide references containing more detailed descriptions.

In this paper we summarize the comparison of these methods, identify the most discriminating characteristics and describe current and future enhancements to the original benchmark that improve the ability to compare methods. Although, currently, the benchmark provides three assessment datasets, obtained from Affymetrix's HGU95A and HGU133 Latin square experiments and GeneLogic's dilution experiment, in this paper we demonstrate its utility using only the HGU95A data. We use this particular dataset because many developers have submitted entries based only on this dataset. However, the limited comparisons that are possible with the other two datasets lead to similar conclusions. The reader is welcomed to visit the webtool's entry comparison tool where one can create assessment plots and summaries, such as those shown in this paper, on the fly. Screen shots, included as Supplementary Material, provide some examples.

We assume that the reader is familiar with the original benchmark, Affymetrix terminology and the basic issues of preprocessing. See Cope *et al*. (2004) for a summary.

## 2 MOTIVATION

Figure 1 shows the results of hierarchical clustering of all the above mentioned methods (Supplementary Figure 1 shows similar result obtained with the dilution data). Figure 1 helps us ascertain three important facts: The first is that methods that differ only in normalization (RMA_NBG/VSN/GL and RMA/RMA_VSN) result in practically identical measures. The second is that methods that do not perform probe-specific background correction cluster together. The third fact is that methods that do perform probe-specific background do not cluster tightly. This is in agreement with the previously published results (Irizarry *et al*., 2003; Cope *et al*., 2004).

*To whom correspondence should be addressed.

 

**Table 1.** Description of methods submitted for comparison

| Method | Summarization | Background correction | Normalization | Citation |
|---|---|---|---|---|
| ChipMan | A multiplicative model similar to that of dChip is fit to the PM | | Linear transformation | (Lauren, 2003) |
| dChip | A multiplicative model is fit | MM intensities are subtracted | Spline fitted to rank invariant set | (Li and Wong, 2001) |
| GL | As RMA | None | Loess fitted to subset | (Freudenberg, 2005), http://www.izbi.uni-leipzig.de/izbi/Working%20paper/2005/03dipl.pdf |
| gMOSv.1 | Parameters from a gamma model are estimated from the PM and MM. These account for background and signal | | | (Milo *et al*., 2003) |
| GCRMA | As RMA | Based on probe sequnece | As RMA | (Wu *et al*., 2004) |
| GSVDmod | Generalized SVD is used | None | Scale normalization | (Zuzan, 2003) |
| MAS5.0 | A robust average (Tukey biweight) | Spatial effect and MM subtracted | Scale normalization | (Affymetrix, 2002), http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper%.pdf |
| MMEI | A linear mixed model is fitted | None | Linear mixed model used as well | (Deng *et al*., 2005), http://math.bnu.edu.cn/startprob/CSPS-IMS2005/Abstracts/ShibingDeng.pdf |
| PerfectMatch | Model accounts for background and signal. The non-specific and specific effects are predicted using a free energy model | | | (Zhang *et al*., 2003) |
| PLIER | A multiplicative model is fitted to PM-MM. Accounts for heteroskedacity | | As RMA | (Hubbell *et al*., 2004), http://www.affymetrix.com/community/publications/affymetrix/expr_supp.pdf |
| ProbeProfiler | Proprietary algorithm (http://www.corimbia.com) | | | |
| RMA | A robust linear model is fitted | A global correction is performed | Quantile | (Irizarry *et al*., 2003) |
| RSVD | The robust singular value decomposition methodology is applied to probe-level data | | | (Liu *et al*., 2003) |
| UMTrMn | A trimmed mean of the PM-MM is computed Negatives are truncated | | Similar to RMA | (Giordano *et al*., 2001) |
| VSN | As RMA | Generalized log transform is used to normalize and background correct | | (Huber *et al*., 2002) |
| ZAM | A robust linear model is fit | Similar to RMA | Averaged pairwise Loess | (Åstrand, 2003) |
| ZL | Model used to motivated a generalized log transform that normalizes, background corrects and summarizes. | | | (Zhou and Rocke, 2005) |

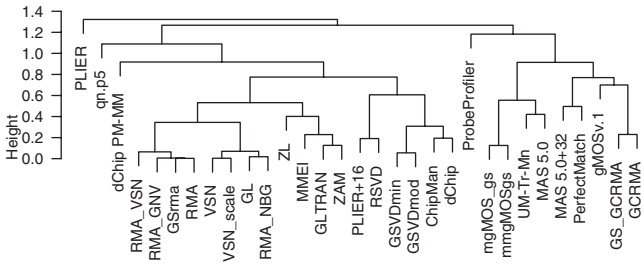


**Fig. 1.** Deprogram showing the results of hierarchical clustering applied to the log expression data obtained from each method when applied to the HGU-95 spike-in data. The *y*-axis represents the clustering height. Correlation was used as a similarity metric. We used the median correlation to summarize across the 59 arrays.

These facts suggest that background correction is the main factor that explains differences between methods.

Throughout the text we use the terms precision and accuracy. Notice that in the context of detection of differential expression these translate to more familiar terms: specificity and sensitivity. However, to provide guidance to those interested in other microarray applications we describe assessments using the more general terms.

Statistical models for probe-level data predict that no background correction leads to attenuated estimates of differential expression (bias) and that naive background correction procedures can lead to highly variable estimates of differential expression (Durbin *et al*., 2002; Huber *et al*., 2002; Wu *et al*., 2004). In particular, methods that produce expression measures close to 0 can have particularly large variance. This fact probably led Affymetrix to submit entries that add a constant to their expression measures, such as PLIER+16 (Table 1). Adding this constant greatly improves the variance of the measure because we no longer divide by numbers close to 0 when computing fold-changes.

Figure 2a, which plots benchmark assessments of overall accuracy and precision against each other, provides empirical corroboration of how background correction affects the accuracy/precision trade-off. This picture demonstrates that the most precise methods are, in general, the least accurate. Furthermore, the statistical models for probe-level data also predict that the bias owing to lack of background correction is greater for low-expressed genes (Wu *et al*., 2004). Figure 3a (Fig. 4a in the benchmark) confirms this empirically. In this figure, we included six methods as representative of methods that do no or little background correction (RMA_NBG and VSN_scale), global background correction (RMA, RSVD) and probe-specific background correction (PLIER+16, GCMRA). Notice that methods that do not background

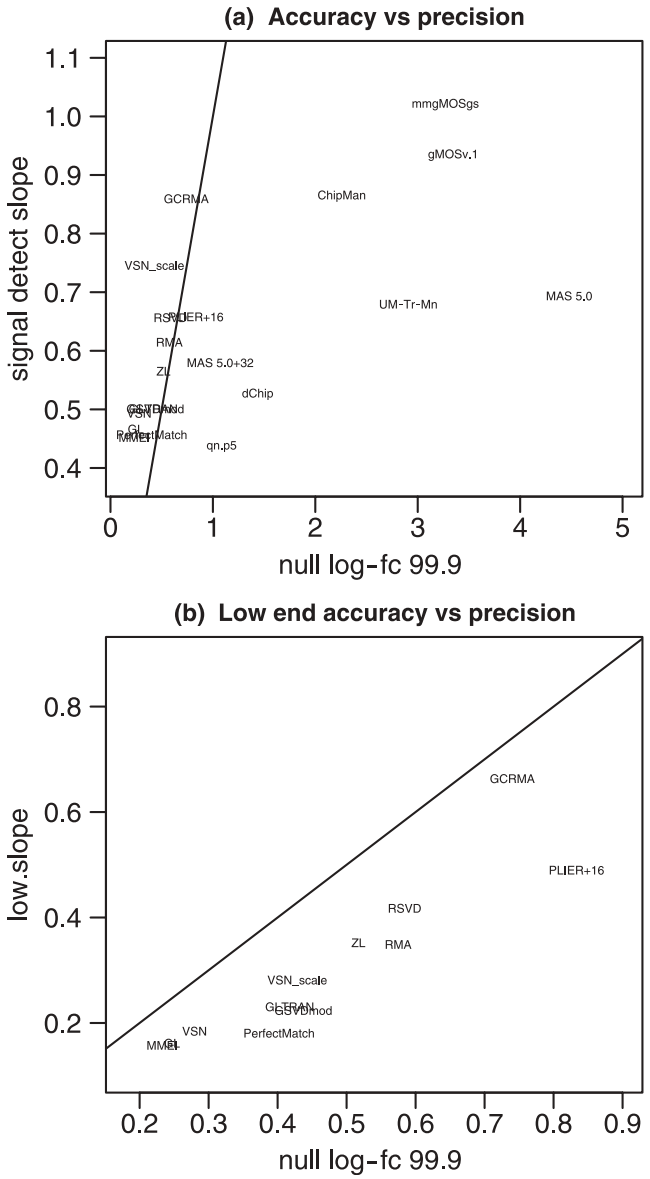


**Fig. 2.** Accuracy versus precision plots. The solid line is the identity line. (**a**) A slope estimate that represents the expected log-fold-change of a gene with a fold-change of 2 is plotted against the 99.9th percentile of log-fold-change among genes that are not differentially expressed. Notice that in a microarray with 10 000 genes, 100 false positives are expected to surpass the value represented in the *x*-axis. dChip, PLIER and ProbeProfiler are not shown because their *x*-axis values were too high (10.83, 18.75 and 123.27, respectively). (**b**) As (**a**) but the *y*-axis has the slope estimate for low expressed genes. The range of the *x*-axis has been limited to show the better performing measures.

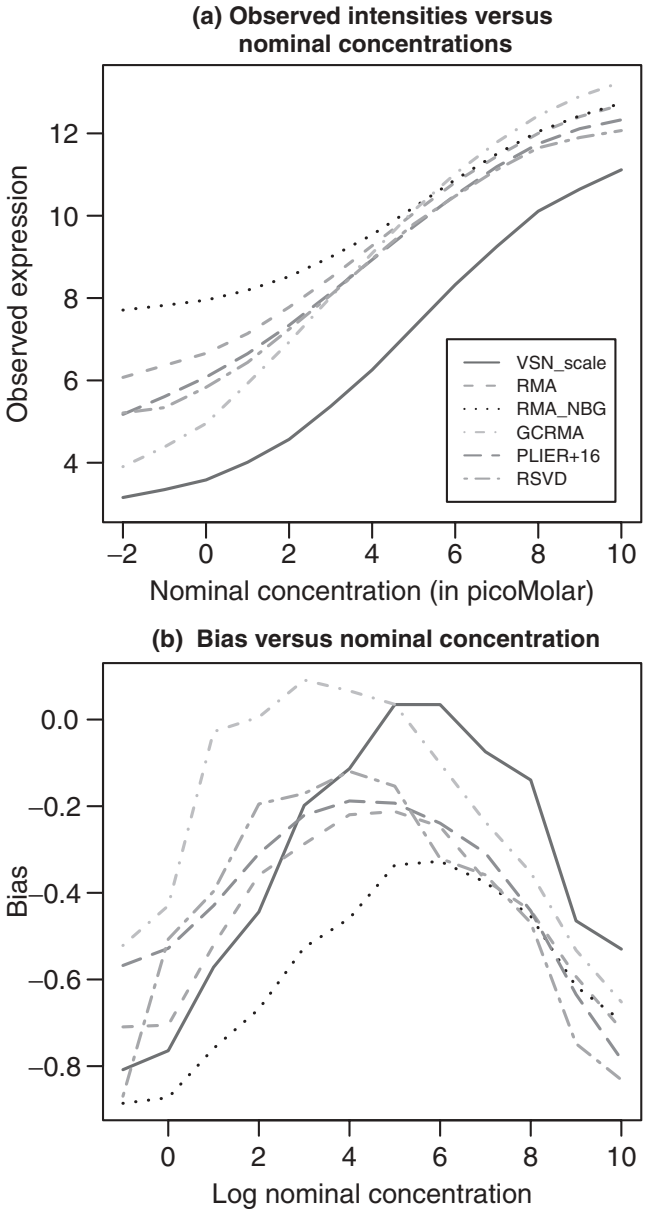


**Fig. 3.** (**a**) Observed log (base 2) expression versus nominal log concentration (in picoMolar). (**b**) The difference between one (the desired value) and local slopes, or bias, versus nominal log concentration (in picoMolar).

correct result in curves that flatten out when the nominal concentration is small. This fact is better illustrated by Figure 3b, which we describe in the next section. To better understand the relationship between accuracy/precision and overall expression we have extended some of the current assessment measures and plots. In the next section we describe these extensions as well.

## 3 ENHANCEMENTS TO BENCHMARK

For the below described measures, a 28 array subset of the HGU95 spike-in that balances concentration levels across experiment was used. This subset is described by Wu *et al*. (2004).

### 3.1 Accuracy

Because accuracy depends on the overall expression of genes, we separated the main accuracy assessment, [Signal detect slope (row 6 in Table 1 of Cope *et al*., (2004)] into three components. To do this, we stratified the spiked-in genes into low expressed (nominal concentration <4 pM), medium expressed (nominal concentration

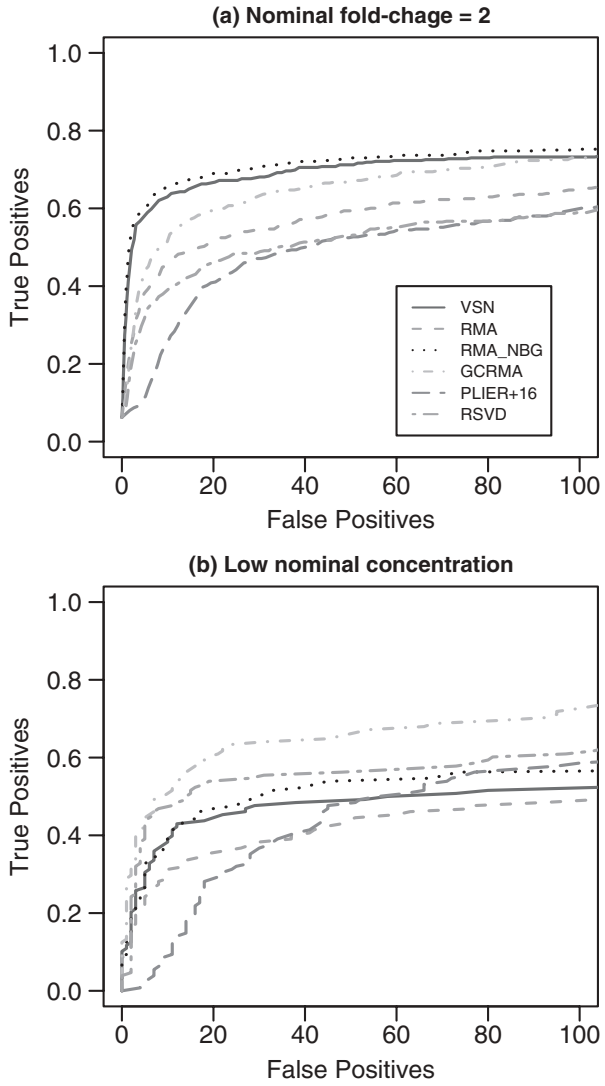


**Fig. 4.** A typical identification rule for differential expression filters genes with fold change exceeding a given threshold. This figure shows average ROC curves which offer a graphical representation of both specificity and sensitivity for such a detection rule. (**a**) Average ROC curves based on comparisons with nominal fold changes equal to 2. (**b**) As (**a**) but consider only low concentration spiked-in genes.

between 4 and 32) and high expressed (nominal concentration >32). For each of these subgroups we followed the same procedure used to compute the signal detect slope. Specifically, a regression line was fitted to the observed log expression values for the spiked-in genes using nominal concentration as the predictor, and the slope estimate recorded as the assessment measure. The new assessment measures are referred to as low, med and high slopes and are shown in Table 2.

To better assess the concentration dependent bias, we added the plot shown in Figure 3b to the benchmark. In this figure, local slopes are calculated by taking the difference between the average observed log expression values between consecutive nominal concentration levels. The difference between 1 and these local slopes are plotted against the larger of the two concentration levels. We subtract from 1 because we are in the log-scale, thus all these slopes should be 1 (when nominal concentration doubles so should the observed concentrations). Notice that these curves follow an upside-down *U* shape. This shape illustrates the fact that bias is worst for low expressed genes and high expressed genes. The low expressed genes are affected by background noise as described above. The high expressed genes are affected by scanner attenuation (not discussed in this paper).

**Table 2.** Table showing the new assessment summary statistics described in the text

| Method | SD | 99.9% | Slope Low | Slope Med | Slope High | AUC |
|---|---|---|---|---|---|---|
| GCRMA | 0.08 | 0.74 | 0.66 | 1.06 | 0.56 | 0.70 |
| GS_GCRMA | 0.10 | 0.79 | 0.62 | 1.03 | 0.55 | 0.66 |
| MMEI | 0.04 | 0.23 | 0.16 | 0.54 | 0.46 | 0.62 |
| GL | 0.05 | 0.25 | 0.16 | 0.55 | 0.46 | 0.62 |
| RMA_NBG | 0.04 | 0.24 | 0.16 | 0.56 | 0.46 | 0.61 |
| RSVD | 0.00 | 0.58 | 0.42 | 0.85 | 0.40 | 0.61 |
| ZL | 0.22 | 0.52 | 0.35 | 0.71 | 0.45 | 0.61 |
| VSN_scale | 0.09 | 0.43 | 0.28 | 0.91 | 0.70 | 0.59 |
| VSN | 0.06 | 0.28 | 0.18 | 0.6 | 0.46 | 0.59 |
| RMA_VSN | 0.09 | 0.48 | 0.31 | 0.74 | 0.46 | 0.57 |
| GLTRAN | 0.07 | 0.42 | 0.23 | 0.61 | 0.45 | 0.55 |
| ZAM | 0.09 | 0.50 | 0.30 | 0.70 | 0.47 | 0.54 |
| RMA_GNV | 0.11 | 0.58 | 0.35 | 0.76 | 0.47 | 0.52 |
| RMA | 0.11 | 0.57 | 0.35 | 0.76 | 0.47 | 0.52 |
| GSrma | 0.11 | 0.57 | 0.35 | 0.76 | 0.47 | 0.52 |
| GSVDmod | 0.07 | 0.44 | 0.22 | 0.64 | 0.42 | 0.51 |
| PerfectMatch | 0.05 | 0.40 | 0.18 | 0.56 | 0.43 | 0.50 |
| PLIER+16 | 0.13 | 0.83 | 0.49 | 0.80 | 0.46 | 0.48 |
| GSVDmin | 0.08 | 0.60 | 0.22 | 0.62 | 0.41 | 0.41 |
| MAS 5.0+32 | 0.14 | 1.07 | 0.35 | 0.71 | 0.44 | 0.12 |
| ChipMan | 0.27 | 2.26 | 0.44 | 1.11 | 0.68 | 0.12 |
| qn.p5 | 0.12 | 1.09 | 0.13 | 0.50 | 0.52 | 0.11 |
| dChip PM-only | 0.13 | 1.44 | 0.31 | 0.67 | 0.39 | 0.09 |
| mmgMOSgs | 0.40 | 3.27 | 1.34 | 1.13 | 0.45 | 0.07 |
| gMOSv.1 | 0.29 | 3.35 | 0.98 | 1.12 | 0.42 | 0.06 |
| ProbeProfiler | 0.31 | 18.75 | 1.61 | 1.57 | 0.39 | 0.03 |
| dChip | 0.23 | 14.83 | 1.40 | 0.86 | 0.35 | 0.02 |
| mgMOS_gs | 0.36 | 2.86 | 0.83 | 0.86 | 0.43 | 0.01 |
| MAS 5.0 | 0.63 | 4.48 | 0.69 | 0.81 | 0.45 | 0.00 |
| PLIER | 0.19 | 123.27 | 0.75 | 0.85 | 0.46 | 0.00 |
| UMTrMn | 0.32 | 2.92 | 0.58 | 0.83 | 0.42 | 0.00 |

The methods are ordered by their performance in the weighted average AUC value.

## 3.2 Precision

To provide a more practical context for the new accuracy assessment measures, we defined the null log-fc 99.9% statistic shown in Figure 2. Row 6 in Table 1 of Cope *et al*. (2004) presented the interquartile range (IQR) of the observed log-fold-changes among the genes that are known not to be differentially expressed. The new statistics gives the 99.9% instead of the IQR. We have also added a measure related to the Median SD represented by row 1 in Table 1of Cope *et al*. (2004). The previous measure used the dilution study data. Similarly, a spike-in experiment version of Figure 2 in the original benchmark was added.

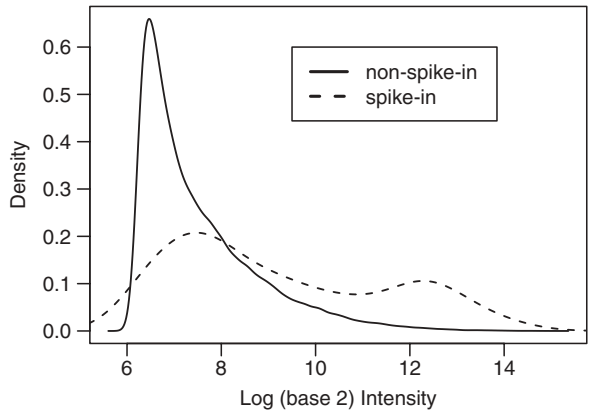


**Fig. 5.** Empirical density estimates of the distribution of log expression for non-spiked-in genes and spiked-in genes.

### 3.3 Overall detection ability

One of the chief uses of expression arrays is the identification of genes that express differently under various experimental conditions. The simplest 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. Cope *et al*. (2004) presented two ROC plots, both using log fold change as a filter. Since only spiked-in genes are actually differentially expressed in these experiments, it is easy to determine TP and FP. For the first plot every concentration pair was used to determine TP. Because many concentration pairs result in unrealistically high nominal fold-changes, a second plot used only combinations yielding fold-changes of 2 (Fig. 4a). The *x*-axis stops at 100 false positives because lists of genes with more errors are not typically useful. As summary statistics we reported the area under the curve (AUC).

According to Figure 4a, methods with no or little background correction performed best. However, many of these methods performed rather poorly in the accuracy plots seen in Figure 3. The reason for this apparent discrepancy is that in the benchmark experiment the spiked-in concentration resulted in abnormally high levels of observed expression. This is demonstrated by Figure 5 which compares the intensity distributions of the spiked-in genes and non-spiked-in genes. To allow the ROC curves to provide a more realistic summary we divided the ROC curve plots into three components. For each of the concentration groups, defined for the accuracy assessment, we created a different ROC curve and we consider only sample pairs with fold-changes equal to 2. Figure 4b shows the low intensity ROC curves for the same six methods in Figure 3. The AUC for these three ROC curves are added as summary statistics. To give a one number summary we consider a weighted average of these three AUCs (Table 2). The weights are chosen according to the percentage of genes expected to be in each concentration group.

An MA-plot that only shows the spiked-in genes in each of these concentration groups with fold-changes <4 was also added.

## 4 DISCUSSION

Figure 2a plotted the original benchmark's signal detect slope against the 99.9 percentile log-fold-change among the genes that are not differentially expressed. The value in the *y*-axis represents the expected log-fold-change of a gene with a true fold-change of 2. These two statistics give an intuitive and practical summary related to the ability to detect differentially expressed genes. In general, the higher above the identity line, the more preferable the method. Notice that various methods are well below the identity line (very large variance). This is probably explained by the use of naive background correction procedures. For most of these, a method with the same accuracy exists but with much better precision. However, there are various methods above the identity line with differences in both accuracy and precision. To compare such cases we turn our attention to Figure 3 which demonstrated that methods that do not background correct have worst bias for low expressed genes. We will focus our attention on VSN_scale and RMA_NBG, the methods that appears to perform best in 2a and 4a. In Figure 3a, we see that the curve for RMA_NBG, which does no background correction, flattens out dramatically at the low end. Notice that, except for a stretch caused by the multiplication of a constant, VSN_scale (which by definition will have an identical curve to VSN) has a similar shape to RMA_NBG. Figure 2b plots the signal detect slope obtained for genes with low expression, as described in Section 3.1, against the 99.9 percentile seen in Figure 2a. Notice that some of the methods that appeared to be performing best in Figure 2a, such as VSN_scale and RMA_NBG, are no longer performing very well. In general, the bias resulting from lack of background subtraction will be most noticeable in the summary statistics plotted in the *y*-axis of this figure. Methods such as PLIER+16 and GCRMA, which use model-based probe-specific background correction, maintain relatively good accuracy without losing much precision. RSVD maintains relatively good accuracy except for very low concentrations.

The advantage of background correcting can be seen in the ROC curves as well. Figure 4 shows ROC curves for six methods. Figure 4a shows the overall results presented in the original benchmark. Figure 4b shows the ROC curve that considers only low expressed genes. Notice that for low concentrations methods such as VSN_scale and RMA_NBG do not perform as well as GCRMA and RSVD.

Table 1 suggests that many methods are developed to perfect accuracy without taking precision into account. Others appear to be doing the opposite. In general, the latter are preferred because detection ability is much better. However, some methods such as RSVD, ZL, PLIER+16 and GCRMA appear to be finding a balance between accuracy and precision that permits them to perform well across the range of gene expression. Furthermore, we need to keep in mind that in practice it is typical to have replicate arrays which improves precision but not accuracy. For example, in experiments with many replicates a user might be willing to sacrifice accuracy for precision. In classification and clustering problems where the variance from all genes can be accumulated to degrade results, a user is likely to benefit from a very precise measure. Furthermore, some methods, such as gMOS, provide estimates of uncertainty which affycomp does not take into consideration. It is possible that these methods will perform better in comparisons of statistical tests where such estimates can be used to improve specificity.

Affymetrix's spike-in experiments have been an invaluable resource to develop and assess preprocessing methodology. However, it is important to consider other assessment datasets. For this reason, the original benchmark includes the GeneLogic dataset. Results obtained from this experiment lead to similar conclusions (see Supplementary Figure 2). Because the spiked-in-genes in the benchmark data are known, over-training is a concern. For this reason we have enhanced the benchmark web tool to accept results from an independent spike-in experiment (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). We have recently asked all submitters to make results from both experiments available. At the time of writing, most of the better performing methods had only been submitted with one dataset. Assessment plots and tables, e.g. Supplementary Figure 3, suggest that at least for the methods submitted, the conclusions are similar to those obtained with the HGU95 data.

To further improve the benchmark, we plan to add benchmark datasets, such as the one obtained from a mixture experiment described by Lemon *et al*. (2002), in the near future. Furthermore, we are currently developing a spike-in experiment that uses biological replicates instead of technical ones. Currently, there is another spike-in experiment in the public domain (Choe *et al*., 2005). Unfortunately, the design of this experiment is not compatible with the assessments we propose. Furthermore, the existence of an artifact makes it difficult to define true and false positives. We therefore have no plans to include this dataset in our benchmark. Further details are given in the supplemental material.

It is important to note that both precision and accuracy depend on the signal/noise ratios both intrinsic to the species, the experiment and the specific microarray utilized. Thus, the benchmarking of precision and accuracy we describe for the older U95A microarray and the artificial spike-in dataset may or may not be generalizable. We anticipate that the webtool will continue to be populated with additional datasets from a range of experiments and chips.

## 5 CONCLUSION

In this paper we described some enhancements to the benchmark assessment plots and summaries that further elucidate differences among existing preprocessing methods. In Section 4 we compared the methods submitted for scrutiny via the benchmark. For the sake of clarity, most of the figures in this paper compared only six methods. However, using the benchmark web tool one can compare any combination of methods via any summary statistic or plot. Beware that results for the original benchmark, as described by Cope *et al*. (2004), are available from the original assessment link on http://affycomp.biostat.jhsph.edu, while the enhancements described here are available from the new assessment link on that webpage.

The benchmark has been an invaluable tool for comparing different preprocessing methods. It has also been useful for determining the characteristics that differentiate these methods. The comparison made evident that the accuracy/precision (bias/variance) trade-off is driven mostly by background correction. It is important to note that the benchmark is not intended to be used to determine the 'best' method but rather to permit users to judge each method using scientifically meaningfully summaries. These can be used to decide the most appropriate method for their specific application. We expect this paper, along with the benchmark web tool, to help researchers continue to improve preprocessing algorithms. In particular, we have clearly laid out the importance of balancing precision and accuracy.

*Conflict of Interest*: none declared.

## REFERENCES


Affymetrix. (2002) Statistical algorithms description document. *Technical report*.

Åstrand,M. (2003) Contrast normalization of oligonucleotide arrays. *J. Comput. Biol.*, **10**, 95–102.

Choe,S.E. *et al*. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Geneome Biol*, **6**, R16.

Cope,L.M. *et al*. (2004) A benchmark for Affymetrix Genechip expression measures. *Bioinformatics*, **20**, 323–331.

Deng,S. *et al*. (2005) A mixed model expression index to summarize affymetrix geneChip probe level data. *Mathematical Subject Classification*, **62-07**, 62P10.

Durbin,B.P. *et al*. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18** (Suppl. 1), S105–S110.

Freudenberg,J.M. (2005) Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays. *Technical Report 3, Leipzig Bioinformatics Working Paper*.

Giordano,T. *et al*. (2001) Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. *Am. J. Pathol.*, **159**, 1231–1238.

Hubbell,E., Liu,W.M. and Mei,R. Supplemental data: robust estimators for expression analysis. Technical report, Affymetrix.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Irizarry,R. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Lauren,P.D. (2003) Algorithm to model gene expression on affymetrix chips without the use of mm cells. *IEEE Trans. Nanobioscience*, **2**, 163–170.

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

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Lu,L. *et al*. (2003) Robust singular value decomposition analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 13167–13172.

Milo,M. *et al*. (2003) A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochem. Soc. Trans.*, **31**, 1510–1512.

Wu,Z. *et al*. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.

Zhang,L. *et al*. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.

Zhou,L. and Rocke,D.M. (2005) An expression index for affymetrix geneChips based on the generalized logarithm. *Bioinformatics*, **21**, 3983–3989.

Zuzan,H. (2003) Generalized svd analysis for improved estimation of expression indices in the li-wong framework. *Presented in: The 2003 Affymetrix GeneChip Microarray Low-Level Workshop*.