Johns Hopkins University, Dept. of Biostatistics Working Papers

9-1-2005

Comparison of Affymetrix GeneChip Expression Measures

Rafael A. Irizarry
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rafa@jhu.edu

Zhijin Wu
Center for Statistical Sciences, Department of Community Health, Brown University, Zhijin_Wu@brown.edu

Harris A. Jaffee
Johns Hopkins University, hjaffee@jhmi.edu

BIOINFORMATICS
Vol. 1 no. 1 2005
Pages 1–7

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, U.S.A. and [2]Center for Statistical Sciences, Department of Community Health, Brown University, 167 Angell Street, BOX G-H, Providence, RI, 02912, U.S.A.

ABSTRACT

Motivation: Affymetrix GeneChip expression array technology has become a standard tool in medical science and basic biology research. In this system, preprocessing occurs before one obtains expression level measurements. Because the number of competing preprocessing methods was large and growing, in the summer of 2003 we developed a benchmark to help users of the technology identify the best method for their application. In conjunction with the release of a Bioconductor R package (`affycomp`), a webtool was made available for developers of preprocessing methods to submit them to a benchmark for comparison. There have now been over 30 methods compared via the webtool.

Results: Background correction, one of the main step 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 accuracy and precision. A handful of methods stand out as maintaining a useful balance.

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

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 over-all 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 have submitted more than 30 methods. In this paper we summarize the comparison of these methods, identify the most discriminating characteristics, and describe enhancements to the original benchmark that improve the ability to compare methods. We assume that the reader is familiar with the original benchmark, Affymetrix probe-level terminology, and the basic issues of preprocessing. See Cope et al. (2004) for a summary of all these subjects.

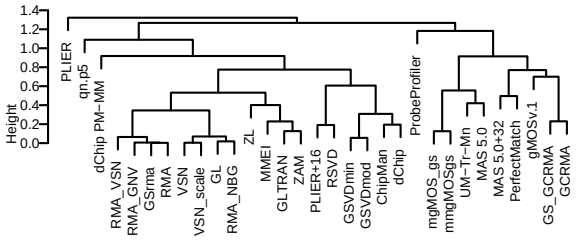

Fig. 1. Deprogram showing the results of hierarchical clustering applied to the log expression data obtained from each method when applied to the HGU-95 spike-in data. The y-axis represents the clustering height. Correlation was used as a similarity metric. We used the median correlation to summarize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

METHODS

In this section we give a very brief overview of the methods being compared. We include references that provide further details. Throughout the paper we will be using nicknames (in bold) to denote the different methods.

Affymetrix has submitted various entries. The Affymetrix default algorithm (Affymetrix, 2002) is

denoted with **MAS 5.0**. A version of MAS 5.0 that adds 32 to the expression measurements was also submitted and denoted with **MAS5+32**. Affymetrix also submitted their new algorithm (**PLIER**) and a version that adds 16 to the expression values (**PLIER+16**).

Two of the algorithms implemented by the dChip software (http://www.dchip.org) have been submitted. **dChip** denotes the PM-only version of the algorithm described by Li and Wong (2001). **dChip PM-MM** is the background adjusted version that uses the $PM - MM$.

Various version of the **RMA** methodology (Irizarry et al., 2003) have been submitted as well. RMA performs background correction, normalization, and summarization in a modular way. The different versions explore changes to these components. **RMA_NBG** is a version that does no background subtraction. **RMA_VSN** uses the variance stabilizing normalization, described by Huber et al. (2002), instead of the default quantile normalization (Bolstad et al., 2003). The method denoted **VSN** is like RMA_VSN except the RMA background correction is not applied. Notice that the procedure described in Huber et al. (2002) tries to account for background. **VSN_scale** is a version **VSN** that has been transformed by a shift and re-scale of the log expression level data. Notice that this shifted version will result in identical values for many of the assessment summaries. The method denoted with **qn.p5** is a version of RMA that arbitrarily uses only the 5-th probe in the probeset as a summary. **GS_RMA** and **RMA_GNV** are implementations of RMA that give practically equivalent results to **RMA**. The latter is GeneSpring's implementation. **GCRMA** is a version of **RMA** with a background correction component that makes use of probe sequence information (Wu et al., 2004). **GS_GCRMA** is GeneSpring's implementation of **GCRMA** .

Other methods that have been submitted are: **ChipMan** (Lauren, 2003), **GL** (Freudenberg, 2005), **GLTRAN** and **ZL** (Zhou and Rocke, 2005), **GSVDmin** and **GSVDmod** (Zuzan, 2003), **MMEI** (Deng et al., 2005), **ProbeProfiler** (http://www.corimbia.com), **gMOS_v.1**, **mgMOS_gs** and **mmgMOSgs** (Liu et al., 2005), **RSVD** (Liu et al., 2003), **UM-Tr-Mn** (Giordano et al., 2001), **PerfectMatch** (Zhang et al., 2003), and **ZAM** (Åstrand, 2003).

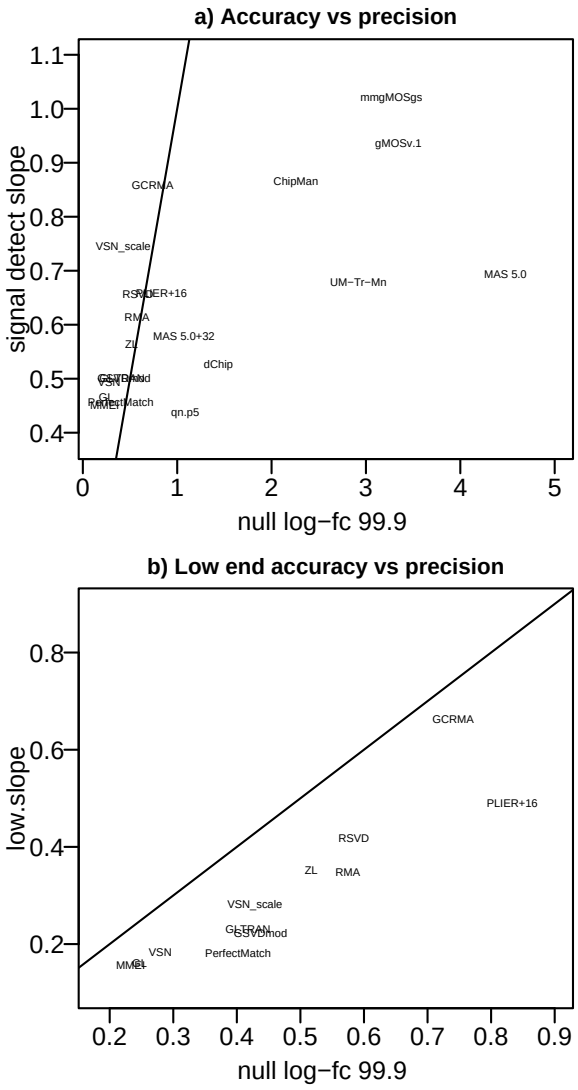

Fig. 2. Accuracy versus precision plots. The solid line is the identity line. A) A slope estimate that represents the expected log-fold-change of a gene with a fold-change of 2 is plotted against the 99.9th percentile of log-fold-change among genes that are not differentially expressed. Notice that in an array with 10,000 genes we expect 100 genes to reach this level. PM-MM dChip, PLIER, and ProbeProfi ler 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.

MOTIVATION

Figure 1 shows the results of hierarchical clustering of all the above mentioned methods. Figure 1 helps us ascertain three important facts: The first is that there are groups of methods that result in practically identical measures. Notice in particular the group of eight with close to 0 distance. The second is that clustering is mainly driven by the type of background correction. Methods that do not, or hardly, correct for background, cluster together, and most of the methods that background correct using the mismatch probes or sequence information form another cluster. The third fact is that the non background correction methods cluster more tightly than those that do. Note that RMA_NBG and VSN have a correlation of 0.998. These two methods differ only in the normalization

step. By comparing different version of RMA, Cope et al. (2004) demonstrated that normalization and summarization have slight effect compared to the differences between RMA and MAS 5.0. This suggest that background correction is the main factor that explains differences between methods.

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). This fact probably led Affymetrix to submit entries that add a constant to the expression data. Figure 2a, which plots benchmark assessments of over-all accuracy and precision against each other, provides empirical corroboration. 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 due to lack of background correction is greater for low-expressed genes (Wu et al., 2004). Figure 3a (Figure 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), moderate background correction (RMA, RSVD), and more vigorous background correction (PLIER+16, GCMRA).

To better understand the relationship between bias/variance and overall expression we have extended some of the current assessment measures and plots. In the next section we describe these extensions.

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).

Accuracy

Because accuracy depends on the overall expression of genes, we separated the main accuracy assessment, (*Signal detect slope* (row 6 in Table 1 in Cope et al. (2004)) into three components. To do this, we stratified the spiked-in genes into low expressed (nominal concentration less than 4 picoMolar), medium expressed (nominal concentration between 4 and 32) and high expressed (nominal concentration larger than 32). For each of these subgroups we followed the same procedure used to compute the *signal detect slope*. Specifically, observed log expression values were plotted against nominal concentration for each spiked-in gene, a regression line fitted, and the slope estimate recorded as the assessment measure. The new assessment measures are referred to as *low*, *med*, and *high slopes*.

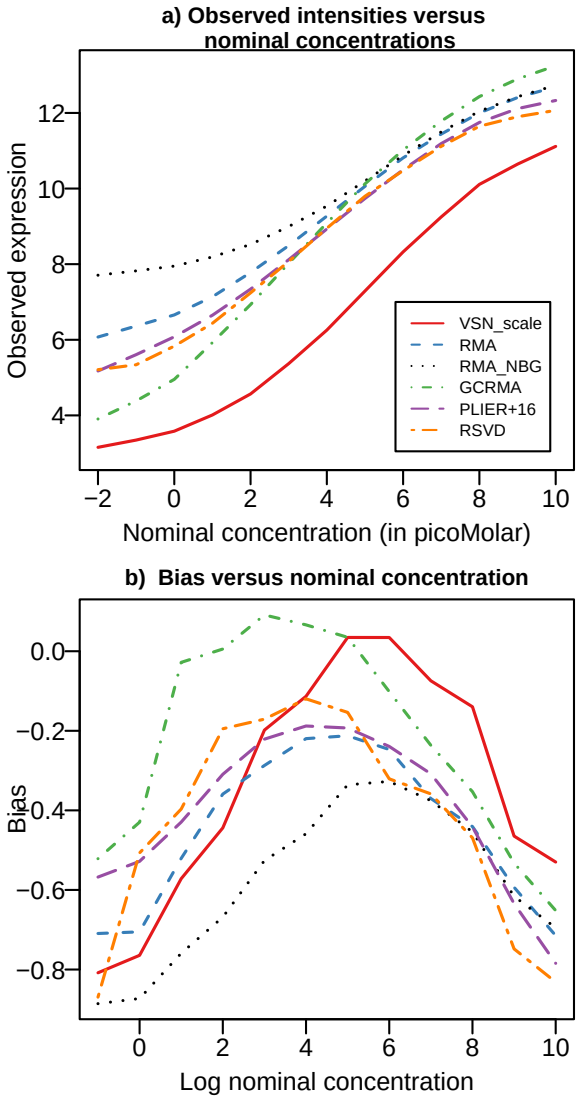

Fig. 3. A) Observed log (base 2) expression versus nominal log concentration (in picoMolar). B) The difference between one (the desired value) and local slopes, or bias, versus nominal log concentration (in picoMolar).

To better assess the concentration dependent bias, we added the plot shown in Figure 2b 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 using log base 2, thus all these slopes should be one (when nominal concentration doubles so should the observed concentrations).

Precision

To provide a more practical context for the new accuracy assessment measures, we defined the *null log-fc 99.9%* statistic. Row 6 in Table 1 of Cope et al. (2004) presented the inter-quartile 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 1 of 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.

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 (Figure 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 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 (see Table 1). 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 smaller than 4 was also added.

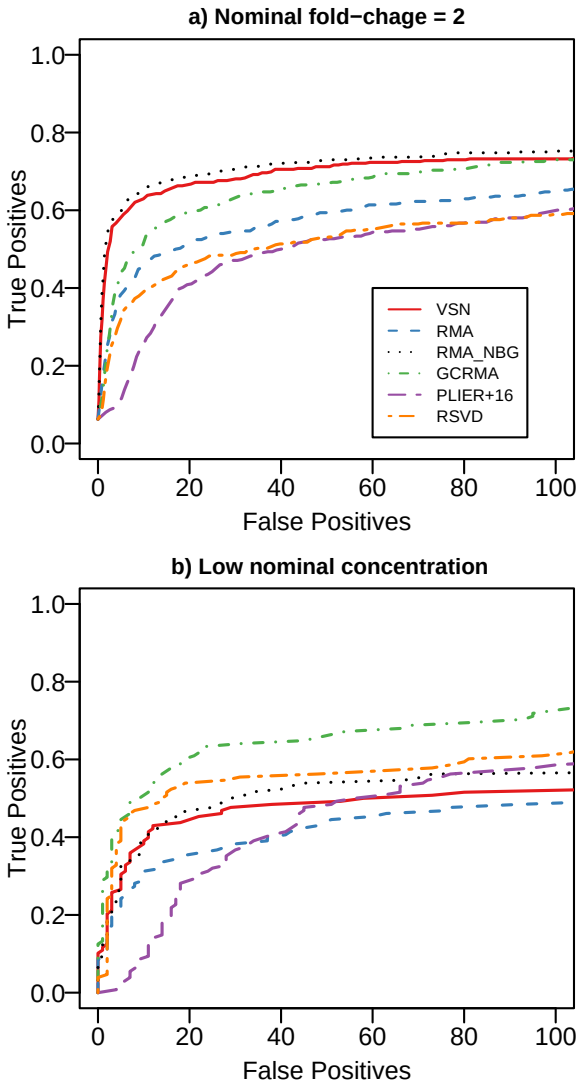

Fig. 4. A typical identifi cation rule for differential expression fi lters genes with fold change exceeding a given threshold. This fi gure shows average ROC curves which offer a graphical representation of both specifi city and sensitivity for such a detection rule. a) Average ROC curves based on comparisons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

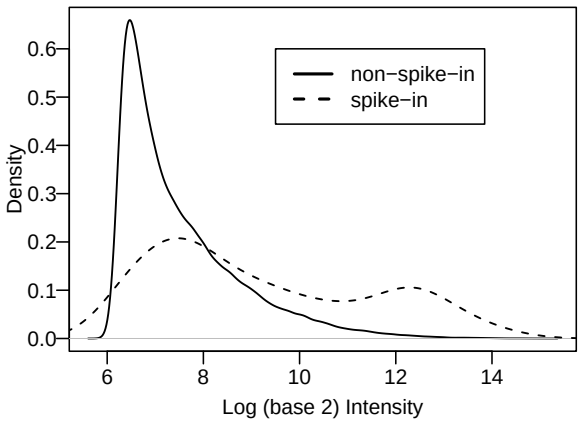

Fig. 5. Empirical density estimates of the distribution of log expression for non-spiked-in genes and spiked-in genes.

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. Notice that in a microarray with 10,000 genes, 100 false positives are expected to surpass the value represented in the x-axis. 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 likely 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 4b, 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 the Accuracy Section, against the 99.9 percentile seen in Figure 2a. Notice that some of the method 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 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

Table 1. Table showing the new assessment summary statistics described in the text. The methods are ordered by their performance in the weighted average AUC value.

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	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
ProbeProfi ler	0.31	18.75	1.61	1.57	0.39	0.03
dChip PM-MM	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
UM-Tr-Mn	0.32	2.92	0.58	0.83	0.42	0.00

perfect accuracy without taking precision into account. Other 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.

CONCLUSION

In this paper we described some enhancements to the benchmark assessment plots and summaries that further elucidate these differences. In the Discussion 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.

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.

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 the bias/variance trade-off drive 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 precession and accuracy.

REFERENCES

Affymetrix (2002). Statistical algorithms description document. Technical report. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

Åstrand, M. (2003). Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology 10*(1), 95–102.

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

Cope, L., R. Irizarry, J. H.A., Z. Wu, and S. T.P. (2004). A benchmark for affymetrix genechip expression measures. *Bioinformatics 20*(3), 323–331.

Deng, S., T.-M. Chu, and R. Wolfinger (2005). A mixed model expression index to summarize affymetrix genechip probe level data. *Mathematical Subject Classification 62-07*, 62P10. http://math.bnu.edu.cn/statprob/CSPS-IMS2005/Abstracts/ShibingDeng.pdf.

Durbin, B. P., J. S. Hardin, D. M. Hawkins, and D. M. Rocke (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. http://www.izbi.uni-leipzig.de/izbi/Working%20Paper/2005/03dipl.pdf.

Giordano, T., K. Shedden, D. Schwartz, R. Kuick, J. Taylor, N. Lee, D. Misek, J. Greenson, S. Kardia, D. Beer, G. Rennert, K. Cho, S. Gruber, E. Fearon, and S. Hanash (2001). Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. *American Journal of Pathology 159*, 1231–1238.

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

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.

Lauren, P. D. (2003, SEPTEMBER). Algorithm to model gene expression on affymetrix chips without the use of mm cells. *IEEE TRANSACTIONS ON NANOBIOSCIENCE 2*(3).

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.

Liu, L., D. M. Hawkins, S. Ghosh, and S. S. Young (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences 100*(23), 13167–13172.

Liu, X., M. Milo, N. Lawrence, and M. Rattray (2005). A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics To appear*.

Wu, Z., R. Irizarry, R. Gentlemen, F. Martinez-Murillo, and F. Spencer (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association 99*(468), 909–917.

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

Zhou, L. and D. M. Rocke (2005). An expression index for affymetrix genechips based on the generalized logarithm. *Bioinformatics*. Under review.

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.