BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 3983–3989
doi:10.1093/bioinformatics/bti665

Gene expression

An expression index for Affymetrix GeneChips based on the generalized logarithm

Lei Zhou[1] and David M. Rocke[2,*]

[1]Department of Biostatistics, Amgen, Inc., USA and [2]Division of Biostatistics, University of California, Davis, CA, USA

Received on June 4, 2005; revised on August 26, 2005; accepted on September 5, 2005
Advance Access publication September 13, 2005

ABSTRACT

Motivation: Affymetrix GeneChip high-density oligonucleotide arrays interrogate a single transcript using multiple short 25mer probes. Usually, a necessary step in the analysis of experiments using these GeneChips is to summarize each of these probe sets into a single expression index that can then be used for determining differential expression, for classification, for clustering, and for other analyses. In this paper, we propose a new expression index that is competitive with the best existing methods, and superior in many cases. We call this expression index method GLA, for GLog Average, since after normalization at the probe level, we take the mean generalized logarithm of perfect match probes.

Results: In this paper, we use Affycomp as the primary tool to assess the weaknesses and strengths of GLA. Comparisons are made between GLA and most widely used summary methods (RMA, MAS5.0 and MBEI) in great detail. The substantial reduction in variability and increased ability to detect differential expression, together with the simplicity of implementation, make GLA a plausible candidate for analysis of Affymetrix GeneChip data.

Contact: dmrocke@ucdavis.edu

1 INTRODUCTION

High-density oligonucleotide array technology is widely used in various areas of medical and biological research projects worldwide. The Affymetrix GeneChip is the most popular single platform of this type, and has become a standard research tool. In the Affymetrix system, an mRNA molecule transcribed from a gene is represented on the GeneChip by a probe set composed of a number probe pairs, typically 10–20. Each probe pair consists of a perfect match (PM) probe and the corresponding mismatch (MM) probe. Each PM probe spot contains thousands of identical short 25mer oligonucleotide sequences that match a segment of an exon of the gene of interest (in some cases the probe may match a segment of the untranslated region, an intron, or may cross a splice joint). The MM probe contains oligonucleotide sequences identical to the PM probe except for a single nucleotide at the center of the sequence which is different and is intended to serve as an internal control of hybridization specificity. After hybridizing the array with a fluorescent-labeled cRNA sample, the array is scanned and the resulting image is processed to give an intensity value for each feature of each probe pair for each probe set.

This paper focus its attention on processing for each probe set, the probe level intensities, to give an expression measure which represents the abundance of the mRNA species in the cRNA sample that has been hybridized to the array. We call such a summary measure an expression index here. More specifically, the PM and MM intensities for each probe set are combined together in some way to produce biologically meaningful expression values. Ideally, expression indices should be both precise (low variance) and accurate (low bias). Affymetrix itself provided a summary measure of average difference (AvgDiff) in 1999. In 2001, Affymetrix developed a new summary measure based on Tukey's biweight function, called MAS5.0. Numerous alternatives for computing expression indices have been proposed in the last couple of years; for example, Model-based expression indices (MBEI) from the dChip algorithms of Li and Wong (2001a,b) and robust multichip average (RMA) derived from robust multichip analysis by Irizarry *et al*. (2003a,b). A number of algorithms, including the aforementioned, have been implemented in the Bioconductor R project (Gentleman *et al*., 2004, http://genomebiology.com/2004/5/10/R80; Ihaka and Gentleman, 1996; R Development Core Team, 2005, http://www.R-project.org). Each algorithm consists conceptually of the following components, although they may not be performed in this order, and may occur simultaneously through a common algorithm.

(1) Possible background correction, which removes background noise from signal intensities;
(2) possible normalization, which is intended to even out unwanted non-biological variation across arrays; and
(3) summarization, which gives the expression index.

An alternative MBEI estimating method is proposed here. This relatively simple method is based on the arithmetic average of the lowess-normalized, GLog-transformed, background corrected PM probes; we call this method GLA for GLog Average. This paper first describes this method in detail and then assesses the performance of the GLA method of computing expression values compared with RMA, MAS5.0 and MBEI using Affymetrix spike-in and GeneLogic dilution datasets in the Affycomp (Cope *et al*., 2004) framework.

2 THE GLA ALGORITHM

In accord with many investigators, we base our expression index on the PM probes alone, without consideration of the MM probes.

[*]To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Although it is possible that use of the MM probes may eventually prove valuable, our feeling is that there is at present too much bias introduced from specific cross-hybridization induced by the use of the MM probes to be a tolerable price for the possible correction for non-specific cross-hybridization. Evidence of this is found, for example, in the relatively large number of probes for which the MM signal exceeds the PM signal by an amount too large to be explained by chance. In our experience, there are often thousands of such probes. If at a later date, incorporation of the MM probes is desired, very similar methods to those proposed below can be used.

The main motivation for the GLA algorithm is that the mean is a summary measure that performs well when the the observations to be averaged have the same variance. Using the variation model of Rocke and Durbin (2001), we obtain a quadratic variance function for the background corrected probe intensities $z = y - \alpha$ consisting of

$$V(y - \alpha) = a^2 + b^2\mu^2,$$

where y is the uncorrected probe intensity, α is the background or an estimate of it, $z = y - \alpha$ is the background corrected probe intensity which has expectation μ, and a and b are positive constants which are assumed not to vary across a single experiment (also see Ideker *et al*., 2001). The GLog transformation

$$z = f(y) = \ln\left(y - \alpha + \sqrt{(y - \alpha)^2 + \lambda}\right)$$

approximately stabilizes the variance of the probe-level data if the background correction α and the transformation parameter λ are carefully chosen (Durbin *et al*., 2002; Hawkins, 2002; Huber *et al*., 2002; Munson, 2001). The parameters α and λ are estimated using a maximum-likelihood method while simultaneously fitting a linear model to the data, as in Durbin and Rocke (2003). This transformation converges to $\ln(2y)$ for large y and is approximately linear when $y - \alpha$ approaches 0. The log transformation for positive data is clearly a special case of the generalized log transformation family.

In the application of these concepts to expression indices, we estimate a single transformation parameter λ and either a single background parameter or a set of array-specific background parameters α_i and using an R routine available from the authors. The method is given in Durbin and Rocke (2003) except that it is applied to probe level data rather than gene-level data. We found that the array-specific background parameters were close to each other, and that it appeared not to affect the performance of the algorithm, so in this paper, we use a single background parameter α for each experiment.

This is a computationally intensive process, but needs to be done only once for each experiment. It is perhaps worth noting that another method and associated Bioconductor package vsn exists for estimating the transformation parameter (Huber *et al*., 2002, 2003). Our methodology essentially tries to make the errors in a statistical model of expression for each gene have a variance that is independent of the mean expression, whereas vsn assumes that most genes have expression that does not vary with the experimental factors and therefore tries to stabilize the variance of the data, not the errors, and uses a robust method of estimation so that the genes which are responsive may be ignored. When the assumptions of both methods are satisfied, the results should be similar.

Once the data have been background corrected and transformed, it is generally necessary to normalize the data to account for array-to-array differences. The simplest way to do this is to include a array normalization parameter in the ANOVA formulation of the experiment, and this is what was done for the parameter estimation phase. There are a number of more aggressive normalization methods, including the use of a non-parametric smoother, such as lowess (Schadt *et al*., 2001, 2002; Li and Wong, 2000), and even matching quantiles (Irizarry *et al*., 2003; Bolstad *et al*., 2003). For this study, we have used lowess normalization of the following type. For the full experiment, we order the probes by the mean across arrays of the background-corrected, GLog-transformed, probe level data. Using this ordering, we number the probes from 1 to p, and for each slide i, we perform a lowess smooth with y equal to the expression index, x equal to the rank and with span 0.1, yielding the smooth l_{ij} for array i and probe j. We then normalize the data $\{y_{ij}\}$ to $\{y_{ij} - l_{ij} + \bar{l}_j\}$, where $\bar{l}_j = p^{-1}\sum_i l_{ij}$. This is done as part of the statistical model that is fit by maximum likelihood. This method should adjust adequately for intensity-dependent, array-specific biases.[1]

The GLA expression index for a given probe set is thus the mean of the lowess-normalized, Glog-transformed, background-corrected, PM probes for that probe set.

3 RESULTS OF AFFYCOMP

There are numerous methods available for the purpose of generating probe set summaries, including RMA (Irizarry *et al*., 2003a,b), MAS5.0 (Affymetrix, 2001), MBEI (Li and Wong, 2001a,b). We use the Affycomp package as the primary tool to assess expression measures from GLA and competitors in terms of accuracy (small bias) and precision (low variance) using the Affymetrix spike-in data and GeneLogic dilution data, which are referred as benchmark datasets, and for which the truth is assumed to be known in advance. Information on both the Affycomp web tool and the benchmark datasets can be obtained at: http://affycomp.biostat.jhsph.edu/. The probe level data are transformed, normalized and then the probes in each probe set are averaged to generate the GLA expression measure. GLA expression measures are exponentiated and divided by 2 so as to be converted back to raw scale before being sent to the Affycomp web tool since Affycomp takes $\log_2$ of raw scale expression measures immediately after it takes the data in, and all assessments are conducted on $\log_2$ scale.

Table 1 lists assessment summary statistics provided by Affycomp on the GeneLogic dilution dataset. As the first check on how four the summary methods work to reduce the variability, the median standard deviation is calculated for each of the four methods. There are five replicate arrays for each combination of tissue source and concentration level. The standard deviation and mean expression are calculated for each gene on each set of replicates on the $\log_2$ scale. As a result, there are 12 by 12 626 standard deviations for each summary method. Median standard deviations are then extracted. The GLA method gives the smallest median standard deviation, which is ~50% of that of RMA and MBEI and ~15%

[1] The transformation parameter estimates are $\hat{\lambda} = 179.02$, $\hat{\alpha} = 18.03$ for the dilution data and $\hat{\lambda} = 1039.6$, $\hat{\alpha} = 70.29$ for the spike-in data. Estimates of array-specific αs are very close to the estimates of the two global αs and do not make a difference in GLA's performance and as a result, we use global $\hat{\alpha}$ for simplicity. The choice of a span of 0.1 is somewhat arbitrary, but seems to work well. We have not conducted studies to determine the optimal bandwidth, though this may be worth doing.

Table 1. Affycomp assessment summaries from the GeneLogic dilution data conducted on $\log_2$ scale

Dilution data: assessment	GLA	MAS5.0	RMA	MBEI
Median SD	0.045	0.292	0.088	0.089
R^2	0.994	0.889	0.994	0.987
1.25v20 corr	0.940	0.729	0.936	0.908
2-Fold discrepancy	2	1226	21	40
3-Fold discrepancy	0	332	0	8
Median slope	0.637	0.847	0.866	0.766

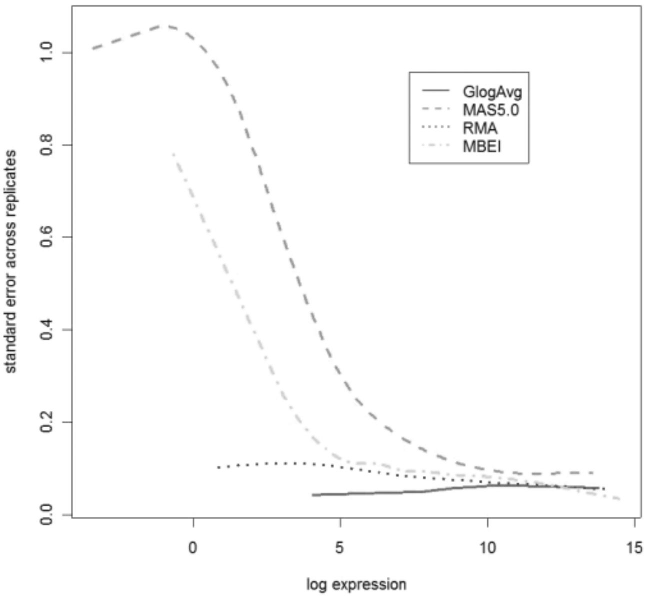

Fig. 1. Smooth curves fitted to the scatter plots of SD versus average of $\log_2$ expression for each gene for the GeneLogic dilution data. All genes for all six concentrations in liver and CNS were used.

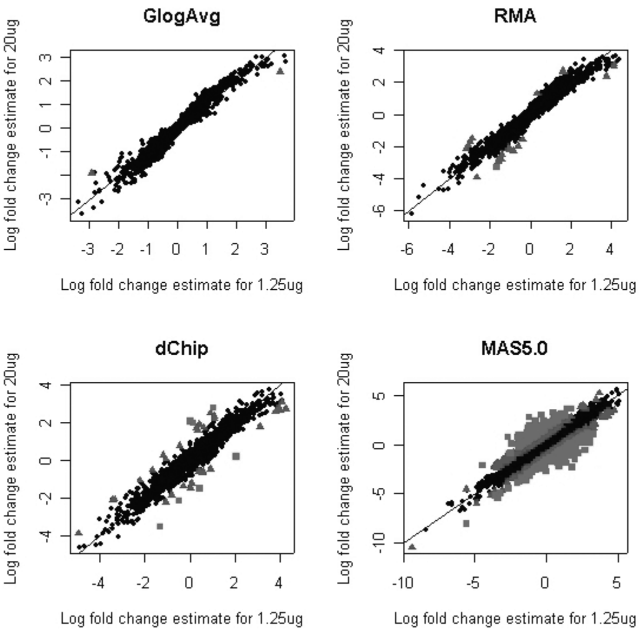

Fig. 2. Log fold change at 20 μg versus log fold change at 1.25 μg for GeneLogic dilution dataset by four methods. Genes with >2-fold discrepancies are shown in solid dark grey trianges, and those with larger than 3-fold discrepancies are shown in solid light grey squares.

of that of MAS5.0 (first row of Table 1). We can observe this phenomenon graphically by fitting a smooth curve to the scatter plot of the standard deviation versus the mean expression, as shown in Figure 1. The GLA method has the lowest standard deviation along almost the whole range of mean expression. The MAS5.0 and MBEI methods show an undesirable elevation of variance at low expression levels, which is not present in the GLA method, and less evident in the RMA method than in MBEI and MAS5.0. This plot again shows that the GLA method obtains the lowest level of variability and also obtains a desirable near independence of the standard deviation on the mean.[2]

The second row of Table 1 gives the average of the squared correlation coefficient R^2 over all pairs of array replicates. High R^2 is an indicator of high precision. R^2 of GLA is 0.994 which is as good as that of RMA and better than that of the other two summary methods. The high R^2 of GLA supports the high precision of GLA expression measure.

Fold change is a relative measure and ideally it should not vary with the amount of cRNA hybridized to the array [though fold change is a poor measure when expression under some conditions is low (Rocke, 2004)]. The four summary methods are assessed according to this data feature of which the expected outcome we know in advance. The next three lines of Table 1 are based on a comparison of estimated fold change at the highest and lowest cRNA concentration level. On the $\log_2$ scale, the expression index is averaged across each set of replicates and the expression ratio between CNS and liver samples is calculated at each concentration level of cRNA. The third row of Table 1 shows the correlation of fold change measurements at the highest (20 μg) and lowest (1.5 μg) concentration levels of cRNA. The GLA expression index gives the highest correlation among the four summary methods, although RMA is close. In Figure 2, the observed $\log_2$ fold changes between CNS and liver when the array is hybridized to cRNA at two concentration levels are plotted against each other. The number of genes that display bigger than 2- and 3-fold discrepancies are symbolized as solid dark grey triangles and solid light grey squares, respectively, and are counted and listed in the next two rows of Table 1. The GLA method scores at the top for these two measures, showing that the GLA expression index gives the fewest large fold change discrepancies.

The last row of Table 1 shows the median slope in the regression of expression index on RNA concentration. Ideally, this should be 1, though it is ~1 for none of the methods. The GLA method does not fare as well in this comparison, though the importance of this is disputable, and is addressed below.

[2] The difference in the range on the x-axis is a consequence of different methods of handling low-end data. For a simple example, if the data were transformed by $z = \log(y + c)$, then the lowest observed value would be $z = \log(y_{\min} + c)$, which can vary substantially with c.

Table 2. Affycomp assessment summaries from the Affymetrix spike-in data

Spike-in data: assessment	GLA	MAS5.0	RMA	MBEI
Mean signal detect slope	0.501	0.689	0.609	0.520
AUC (*FP* < 100)	0.840	0.356	0.821	0.674
AFP, call if *fc* > 2	1.456	3109	15.842	36.907
ATP, call if *fc* > 2	11.090	1.282	11.979	11.428
FC = 2, AUC (*FP* < 100)	0.653	0.065	0.543	0.168
FC = 2, AFP, call if *fc* > 2	0.678	3072	1.000	28.643
FC = 2, ATP, call if *fc* > 2	1.036	3.71	1.714	1.25
IQR	0.186	2.65	0.308	0.447

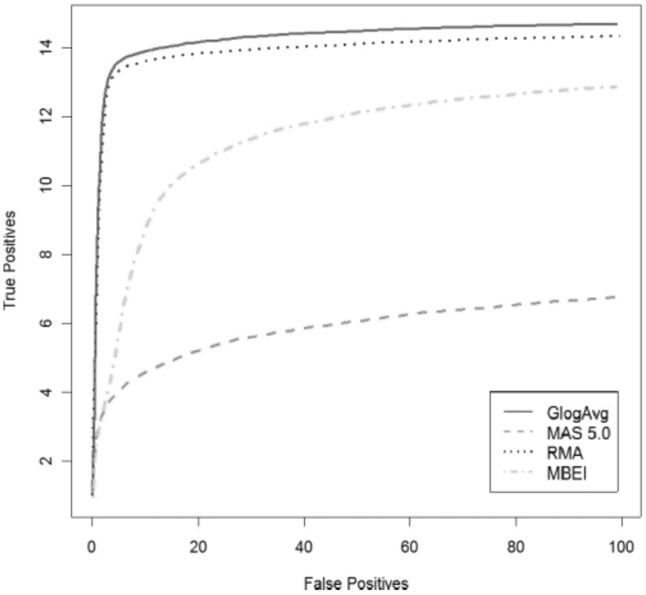

Fig. 3. Average ROC curves for the Affymetrix spike-in data based on comparisons with nominal FC ranging from 2 to 4096.

Table 2 lists Affycomp assessment summary statistics from the Affymetrix spike-in data. One of the principal aims of gene-expression array analyses is to select genes for further consideration using some statistic to measure differential expression, which ideally has large specificity and sensitivity. The ROC (receiver operator characteristic) plot is a popular graphical aid to assess the performance of different methods with respect to specificity and sensitivity in detecting differentially expressed genes. For the spike-in study, only the transcripts that have been spiked in should be differentially expressed. With the truth known in advance, it is easy to identify true positives and false positives (FPs) in order to measure sensitivity and specificity. Using fold change as the statistic (probably not the best such statistics as we will see later) to filter genes on pairs of experiments of which the nominal fold change ranging from 0 to 4096, we compute the number of FPs (detected non-spikes) and true positives (detected spikes) for a large range of cutoffs. FPs and true positives are averaged across all pairs of comparisons and ROC curves, Figure 3, plots the average number of FPs against the average number of true positives. Line 2 of Table 2 shows the area under the curve (AUC) evaluated up to 100 FPs. To calculate the AUC, we computed the number of true positives (which can range from 0 to 14, the number of spiked transcripts) for each value of the number of FPs from 0 to 100, and then averaged the 101 true positive values, and finally divided by the number of spikes. In this metric, the GLA method outperforms RMA, MBEI and MAS5.0. Also as seen in Figure 3, GLA obtains highest sensitivity and specificity among the four summary methods when using fold change as the statistic to identify differentially expressed genes.

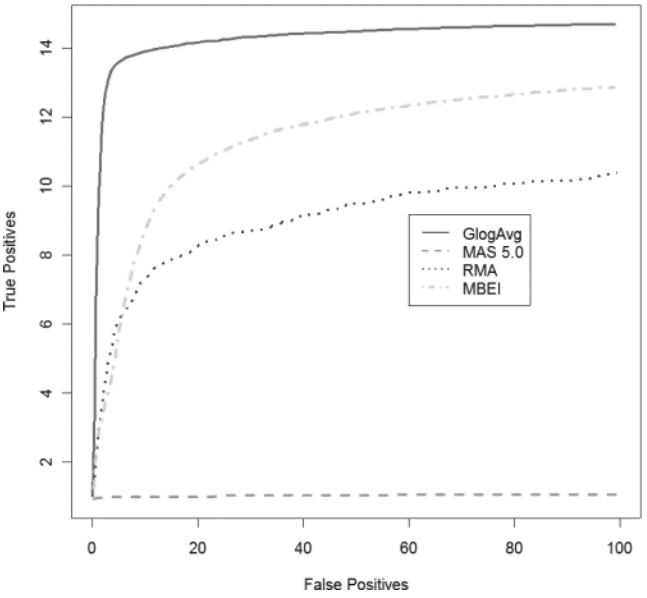

Fig. 4. Average ROC curves for the Affymetrix spike-in data based on comparisons with nominal FC = 2.

If we use observed fold change >2 as the rule to filter genes, in each comparison we compute the number of spikes with observed fold change >2 (true positives) and the number of non-spikes with observed fold change >2 (FPs) and then take average across all comparisons. The third and fourth lines of Table 2 record the results as AFP and ATP, averaged number of FPs and averaged number of true positives. This is one point on the ROC curve, so naturally shows the same ranking as is seen in line 2 of Table 2 and in Figure 3. However, the magnitude of the difference may be surprising. With approximately the same number of true positives as RMA and MBEI, the GLA method gives a tiny fraction of the number of FPs compared with the others (MAS5.0 is, as usual, not competitive).

The smallest nominal fold change available is 2 in the Affymetix U95 spike-in study. The four summary methods can be assessed on their ability to detect spikes when nominal fold change is as small as 2. In this case, we use only pairs of experiments in which the nominal fold change is equal to 2 (rather than all comparisons no matter what the true fold change), we otain the AUC, AFP and ATP values shown in lines 5–7 of Table 2. Again, the GLA method attains the greatest AUC. When using fold change >2 as the rule to decide if the gene is differentially expressed, the ATP of GLA (1.036) is smaller than that of RMA (1.714); however, the AFP of GLA (0.678) is also smaller than that of RMA (1). Figure 4 shows the averaged ROC curves obtained in this scenario, GLA apparently obtains the highest sensitivity and specificity among the

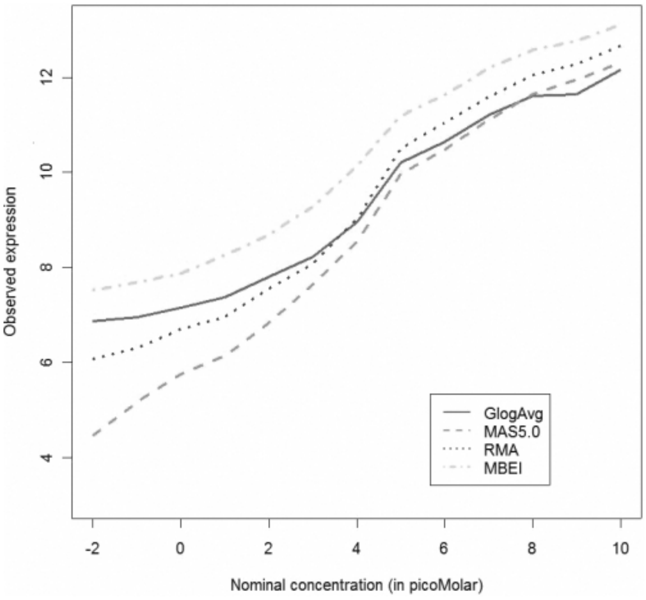

Fig. 5. Average observed $\log_2$ intensity plotted against nominal $\log_2$ concentration for each spike-in transcript for all arrays in the Affymetrix spike-in data set.

four summary methods when using fold change as the statistic to identify differentially expressed genes when the true nominal fold change is at a very low level. The GLA method detects more true positives at every level of FPs. The interquantile range (IQR) of the $\log_2$ fold changes of the non-differentially expressed genes is recorded in the last line of Table 2 as another metric to assess the four summary methods in increasing precision. The GLA expression index has the lowest IQR and again proves to have lowest variability and highest precision among the four expression indices.

Affycomp suggests that for the spike-in data, on the $\log_2$ scale, considering only spikes, if we regress observed expression against nominal concentration, estimated slopes should be distributed around 1. In Affycomp, for 16 spikes, on $\log_2$ scale, observed expression measures are averaged at each nominal expression level and are plotted against nominal expressions, slope and R^2 coefficient are estimated. However, since every gene has its own baseline expression level which might be different from that of other genes, we suggest that it may be superior to regress observed expression against nominal expression for each spiked transcript individually, and then calculate the average of slopes and R^2. These mean signal detect slopes are listed in Table 2 as the first entry. Average of R^2s is not listed here because all four expression summaries provide similarly good, >0.95 R^2 estimates. The GLA and MBEI methods have very close estimated slopes; MAS5.0 has the highest slope and RMA has a slope that falls in between. However, none of the four slopes is close to 1. Observed expression can be plotted against nominal concentration for each spike on $\log_2$ scale. When the four summary methods are compared, the observed expressions are averaged at each nominal concentration value and resulted in four smooth curve as shown in Figure 5. Graphically, we can also see that all four curves have slopes <1.

There are many reasons why this phenomenon should not be surprising. Binding type assays usually have an S-shaped response curve, so that response will flatten out at either high or low levels of concentration. The linearity range of the detector can also be a factor. Methods that work well in stabilizing the variance and allowing for effective inference generally will react approximately linearly to absolute changes at low levels, and approximately linearly to log changes at high levels, but the slopes of these responses may not be the same (the GLog transform is logartithmic at high values of the arguement, and approximately linear near 0). It appears that the GLA method produces a smaller and more variable slope in this test than the other methods, and the reasons for this are still obscure.

The question is whether this is important or not. At this time, the most important issue seems to be the detection of changes that have actually occurred, and not the quantitative estimation of the fold change. Even were the second goal useful, it is not clear that detecting fold changes that are a multiple of the true fold changes is in fact a problem, since there are few cases indeed in which there are reliable quantitative models connecting the magnitude of the fold change with the degree of the biological effect. Nonetheless, it does appear that the GLA method may sacrifice a small amount of accuracy of reproduction of fold changes in return for large gains in sensitivity.

4 THE EFFECT OF THE EXPRESSION INDEX ON STATISTICAL TESTS

Affycomp uses fold change as the statistic to study with respect to the ROC curve and to subset the data for further analysis. Nevertheless, this may not be the ideal method for several reasons: (1) observed fold change is quite noisy at low levels; (2) theoretical fold change is only the correct metric at high levels of expression; and (3) fold change does not recognize the important fact that even after stabilization of the variance as a function of the mean there are still important variance inhomogeneities (Rocke, 2004). The last point is perhaps most important. If the natural variability of a transcript without biological treatments or changes is already 2- or 4-fold, an observed change of 2-fold is probably unimportant, and a deliberate change in transcripts of 2-fold may be undetectable for very good reasons.

Perhaps a better method of assessing differential expression is to use *t*-tests, ANOVA *F*-tests, and other statistical procedures that reflect the design of the experiment. We can use such statistics to study the ROC curve and to decide by some specific cutoff, such as a particular false discovery rate (FDR), which genes are in the identified set (Benjamini and Hochberg, 1995; Reiner *et al*., 2003).

The simplest statistical method for detecting differential expression is the *t*-test, which can be used to compare two conditions when there is biological replication of samples. With more than two conditions or even multiple factors, ANOVA and more complicated linear model tests can be applied. Using the standard *t*-test and using only values from a given gene, owing to the small number of degrees of freedom of the *t*-test (4 for the equal variance *t*-test in this case, and fewer for the Welch *t*-test), the test may not be powerful and may miss some real changes. In order to account for this situation we could pool the mean square errors across genes, but this is likely to be misleading, since it is quite apparent empirically that the natural variability of different genes differs, even after variance stabilization. We can account for this variability, and increase the power of the tests, by employing an empirical Bayes analysis of a hierarchical variance model to provide an

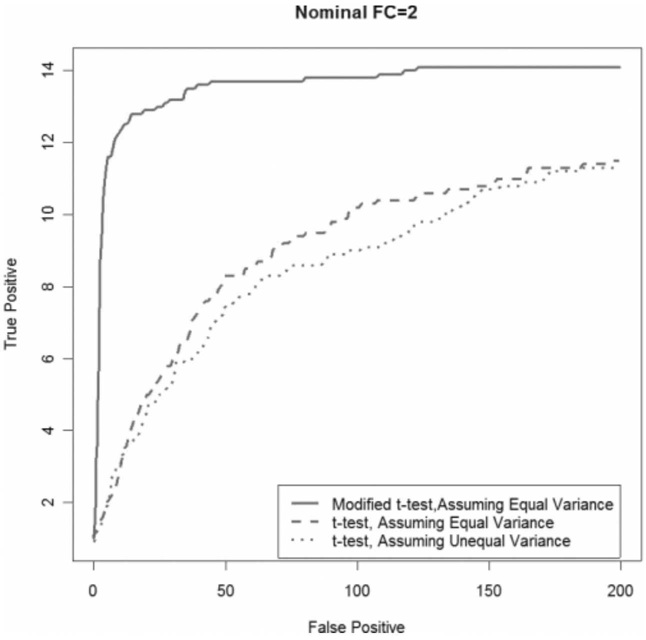

Fig. 6. ROC curves for the ordinary gene-by-gene *t*-tests assuming equal variance or Welch unequal variance test, and for the modified *t*-test that uses the posterior variance from a hierarchical variance model. Data are from the Affymetrix spike-in data set and comparisons are between all pairs of experiments in which the nominal fold change was 2.

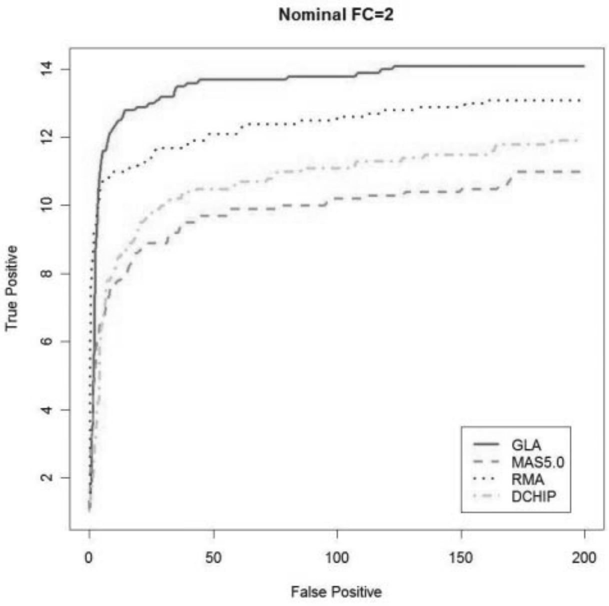

Fig. 7. ROC curves for the modified *t*-test that uses the posterior variance from a hierarchical variance model for four expresssion measures. Data are from the Affymetrix spike-in data set and comparisons are between all pairs of experiments in which the nominal fold change was 2.

improved denominator for *t*-tests or *F*-tests (Rocke, 2004; Smyth, 2004; Wright and Simon, 2003). For convenience, we work in the Gaussian framework and assume an inverse gamma prior for the true variances and a gamma distribution for observed variances. The posterior estimates of mean squared error and degree of freedom are expected to improve substantially the power of statistical tests in identifying important genes.

In the Affymetrix spike-in study, there are three experiments and each of 20 arrays are done in triplets except one that is only in duplicate. Using pairs of experiments with nominal fold change equal to 2, average ROC curves can be created, and are shown in Figures 6 and 7. In Figure 6, the modified unpaired *t*-test assuming equal variance detects a many more true positives for any level of FPs than either of the gene–bygene unpaired *t*-tests. This shows that the empirical Bayes correction increase the power of detection considerably, and also suggests that GLA expression index gains some degree of desirable variance homogeneity, since the equal variance test performs better than the Welch test. In Figure 7, we can see that when we use the *P*-value of the modified *t*-test as a statistic, the ROC curve of the GLA method dominates the other three methods. The GLA method identifies more true positives for any level of FPs compared with the other three summary methods. We also see that RMA does better in this criterioin than MAS 5.0 and MBEI. We can conclude that when using *P*-values of modified *t*-test as the identification statistic, the GLA method gives the highest sensitivity and specificity among the four expression measures.

5 CONCLUDING REMARKS

When working with Affymetrix GeneChip arrays, statistical analysis can be done on the probe level, probe set level or expression index level. The usual approach is the expression index level, in which a single expression index is calculated for each probe set. In this paper, we propose an expression index method which includes a probe level GLog variance stabilizing transformation, probe level loess normalization, and then taking array-wise average for each probe set. An advantage of this algorithm is that it is relatively easy to understand and simple to implement. Using the Affycomp evaluation suite, with some additional measures based on statistical tests, the GLA expression index dominates a number of other possible choices, including MAS5.0, MBEI and RMA on many important measures, especially in the identification of differentially expressed genes. The only possibly important disadvantage is that estimated fold change is smaller than true fold change in this method to a greater extent than the other methods. However, if reliable detection of differential expression is the goal, the GLA method may be superior to these alternatives.

There continue to be new methods proposed of producing expression indices for Affymetrix arrays (e.g. Lemon *et al*., 2003; Wu and Irizarry, 2005), and some of these may be more competitive with GLA than the alternatives examined, so conclusions that hold for all time are difficult to make. As future comparisons are made in the affycomp framework, we hope that the statistical measures we propose of determining differential expression will also be examined, since this is a major focus of the analysis of many array studies. In any case, we believe that this method is worth consideration for anyone analyzing Affymetrix GeneChip data.

We are continuing to make improvements in the method, including methods for using a sample of genes to compute the optimal transformation to speed up the process, and we are investigating the effects of different normalization methods and bandwidths. The method has been applied in several recent studies (Z. Goldberg, D.M. Rocke, C.W. Schwietert, A. Santana, J. Lehmann, R.L. Stern, and C. Hartmann-Siantar, manuscript in preparation); (M.A. Rea, L. Zhou, Q. Qin, Y. Barrandon, K. Easley, S. Gungner, W.S. Holland, P.H. Gumerlock, D.M. Rocke and R.H. Rice, submitted for publication), and as experience accumulates, perhaps reliable default values for the transformation can be provided that depend on the platform but not the experiment, thus removing one computational impediment to the use of the GLA method.

ACKNOWLEDGEMENTS

The authors are grateful to the developers and maintainers of the Affycomp benchmark, Rafael A. Irizarry and Harris A. Jaffee, and to the two referees whose comments greatly improved the presentation of the results in this paper. The research reported in this paper was supported by grants from National Institute of Environmental Health Sciences, National Institutes of Health (P43-ES04699), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (R01-AR044518-06) and the UC Davis Health System.

Conflicts of Interest: none declared.

REFERENCES

Affymetrix (2001), *Microarray Suite User Guide, Version 5.* Affymetrix.

James,A.C. *et al*. (2003) Sensitivity and specificity of five abundance estimators for high-density oligonucleotide microarrays. *Bioinformatics*, **20**, 1060–1065.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

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

Durbin,B.P. and Rocke,D.M. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **19**, 1360–1367.

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

Efron,B. *et al*. (2002) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*., **96**, 1151–1160.

Gentleman,R.C. *et al*. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Hawkins,D.M. (2002) Diagnostics for conformity of paired quantitative measurements. *Stat. Med*., **21**, 1913–1935.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, 96S–104S.

Huber,W. *et al*. (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol*., **2**, Article 3.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat*., **5**, 299–314.

Ideker,T. *et al*. (2001) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol*., **7**, 805–817.

Irizarry,R.A. *et al*. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, e15.

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

Kerr,K. and Churchill,G. (2001) Experimental design for gene expression microarrays. *Genet. Res*., **77**, 123–128.

Kerr,K. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol*., **7**, 819–837.

Lemon,W.J., Liyanarachchi,S. and You,M. (2003) A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biol*., **4**, R67.

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

Li,C. and Wong,W.H. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*., **2**, RESEARCH0032.

Munson,P. (2001) A ‘Consistency’ test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. In *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data,* Bethesda, MD.

R Development Core Team (2005), *R: a Language and Environment for Statistical Computing. Vienna, Austria.* R Foundation for Statistical Computing.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Rocke,D.M. (2004) Design and analysis of experiments with high throughput biological assay data. *Semin. Cell Dev. Biol*., **15**, 708–713.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol*., **8**, 557–569.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3,** Article 3.

Schadt,E. *et al*. (2001) Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem*., **80**, 192–202.

Schadt,E. *et al*. (2002) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem*., **84**, 120–125.

Wright,G.W. and Simon,R.W. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.

Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol*., **12**, 882–893.