BIOINFORMATICS

Vol. 18 no. 12 2002
Pages 1633–1640

Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model

R. Šášik [1,], E. Calvo [2] and J. Corbeil [1]*

[1]School of Medicine, University of California San Diego, La Jolla, CA 92093-0679, USA and [2]Centre Hospitalier de l'Université Laval Research Centre and Quebec Genome Centre, Québec, Canada G1V 4G2

Received on November 8, 2001; revised on April 3, 2002; May 25, 2002; accepted on June 6, 2002

ABSTRACT

Motivation: High-density oligonucleotide arrays (GeneChip, Affymetrix, Santa Clara, CA) have become a standard research tool in many areas of biomedical research. They quantitatively monitor the expression of thousands of genes simultaneously by measuring fluorescence from gene-specific targets or probes. The relationship between signal intensities and transcript abundance as well as normalization issues have been the focus of much recent attention (Hill *et al.*, 2001; Chudin *et al.*, 2002; Naef *et al.*, 2002a). It is desirable that a researcher has the best possible analytical tools to make the most of the information that this powerful technology has to offer. At present there are three analytical methods available: the newly released Affymetrix Microarray Suite 5.0 (AMS) software that accompanies the GeneChip product, the method of Li and Wong (LW; Li and Wong, 2001), and the method of Naef *et al.* (FN; Naef *et al.*, 2001). The AMS method is tailored for analysis of a single microarray, and can therefore be used with any experimental design. The LW method on the other hand depends on a large number of microarrays in an experiment and cannot be used for an isolated microarray, and the FN method is particular to *paired* microarrays, such as resulting from an experiment in which each 'treatment' sample has a corresponding 'control' sample. Our focus is on analysis of experiments in which there is a series of samples. In this case only the AMS, LW, and the method described in this paper can be used. The present method is model-based, like the LW method, but assumes multiplicative not additive noise, and employs elimination of statistically significant outliers for improved results. Unlike LW and AMS, we do not assume probe-specific background (measured by the so-called mismatch probes). Rather, we assume uniform background, whose level is estimated using both the mismatch and perfect match probe intensities.

Results: We present a new method for GeneChip analysis, based on a statistical model with multiplicative noise. We demonstrated that this method yields results superior to those obtained by the Affymetrix Microarray Suite 5.0 software and to those obtained by the model-based method of Li and Wong (Li and Wong, 2001). The present method eliminates the hard-to-interpret negative expression indices, and the binary 'presence' calls (present or absent) are replaced by the statistical significance (p-value) of gene expression. We have found that thresholding the p-values at the $(0.1)^{16}$–level produces about the same number of 'present' calls as the AMS software. By testing our method on a pair of replicate GeneChips (hybridized with the same cRNA), we found that 95.6% of data points lie within the 1.25–fold interval. In other words, our method had a 4.4% type I error rate at the 1.25–fold level. The error rate of the LW method was 15%, and that of the AMS method was 29%. There were *no* points outside the 2–fold interval with the present method. Analysis of variance (ANOVA) of another experiment with multiple replicates shows that this reduction of variance is *not* accompanied by a corresponding reduction of signal. On the contrary, the signal-to-noise ratio (as measured by the distribution of F-statistics) of the present method is on average 3.4-times better than that of AMS, and 1.4-times better than that of Li and Wong.

Contact: sasik@corgon.ucsd.edu

Availability: A Fortran 90 source code of this method is available from the corresponding author upon request.

INTRODUCTION

Gene expression array technology (Lockhart *et al.*, 1996) is rapidly becoming standard in many areas of biomedical research. High-density oligonucleotide GeneChips (Affymetrix, Santa Clara, CA) in particular provide a convenient medium on near-genomic scale for human research, and are a *de facto* industry's standard. Each gene (EST, or a chromosome segment) is represented on

*To whom correspondence should be addressed.

the GeneChip by an array of probe pairs (typically 16 or 20). Each probe pair consists of a perfect match (PM) probe and the corresponding mismatch (MM) probe. Each PM probe contains thousands of short (ideally identical, 25-mer) sequences that are taken from the transcribed sequence of the gene represented by that probe set. The MM probe on the other hand contains sequences identical to the PM probe except for a single nucleotide at the center of the sequence which is different. Upon hybridization with a fluorescent-dye-labeled RNA, each probe in the probe set captures a certain amount of RNA, whose fluorescent intensity is subsequently measured. The measured intensities of all probes in the probe set reflect in a unique way the expression level of that particular gene. Since the hybridizing efficiency of the sequences contained within a probe set is unknown and varied, it is impossible to determine the concentration of the transcript in the solution in an absolute way. Rather, we must content ourselves with a *relative* determination (i.e. the relative amount of change between two measurements). Although the AMS software that accompanies this technology is a sophisticated analytical tool for *individual* arrays, there are statistical methods which, when applied to a *series* of arrays, yield superior results. In the following we describe a method based on a statistical model with multiplicative noise and outlier detection and elimination. We apply this method to analysis of seven arrays, two of which were hybridized with the same RNA solution, and compare the results with two other methods, the AMS method and the LW method. We further perform analysis of variance on a larger set of murine microarrays with multiple replicates.

METHODS

T-cell RNA used in this study was extracted and prepared for hybridization using standard protocols available at http://genomics.ucsd.edu/protocols.html. Mouse mRNA was provided by Dr Fernand Labrie and processed by Dr Thomas Hudson's Microarray facility at the Montréal Genome Centre.

ALGORITHM AND ANALYSIS

Li and Wong (2001) introduce a statistical model based on the differences $PM - MM$:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \qquad (1)$$

where θ_i is the fitted expression index of sample i, $i = 1 \ldots n_s$, ϕ_j is the sensitivity of probe j, $j = 1 \ldots n_p$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. Probe sensitivities are subject to a constraint $\sum_j \phi_j = n_p$ in order that the solution be unique.

Their model is an improvement over the AMS software in that it does not average over the probes in a probe set, which results in the reduction of variance (cf. Figure 3a, b). However, it shares one troubling feature of AMS 5.0, in that the expression index θ_i can be negative (in analogy to the 'average difference' in the AMS 5.0). Furthermore, genes with negative θ can still be classified as present (as can genes with negative average difference in AMS 5.0). Both of these features are highly undesirable. We think that the root cause of these problems is the subtraction of the MM probe intensity in Equation (1), a practice adopted by both algorithms. This subtraction would be justified if the MM probes measured hybridization that is non-specific to the sequence of the PM probes. There is evidence however, that the MM probes are rather specific and are not an independent reference, as discussed by Naef *et al.* (2002b). Perhaps this is not surprising since there is only a single-nucleotide difference between the PM and MM sequences. A related factor is the hybridizing temperature, which should ideally be set so that the PM probes bind their corresponding RNA whereas the MM probes do not. Because of the varying G–C content of individual probes, the optimum hybridization temperature varies among the probes. When the actual hybridizing temperature, set at 42 °C by the experimental protocol, is lower than ideal, the MM probe will bind the transcript with non-zero affinity. In view of these facts we find the practice of subtraction of the MM probe intensities unjustifiable.

Here we propose a statistical model that does not suffer from the above problems. It is based on a similar assumption as the LW model, that fluorescent intensity of a PM probe (properly adjusted for background b) is directly proportional to the concentration c_i of the transcript, $PM_{ij} - b \sim \phi_j c_i$. We choose to write this relationship in the form $\log_2(PM_{ij} - b) \sim \log_2 \phi_j + \log_2 c_i$. Defining $\iota_{ij} \equiv \log_2(PM_{ij} - b)$, $\varphi_j \equiv \log_2 \phi_j$, and $\gamma_i \equiv \log_2 c_i$, we write our model as

$$\iota_{ij} = \varphi_j + \gamma_i + \varepsilon_{ij}, \qquad (2)$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$. There is one such equation for each transcript, and we allow for the possibility that the variance σ of the error term is different for each transcript. Since the errors ε_{ij} are additive at this level, they are multiplicative at the level of the original variables. We observe that in this model, regardless of the fitted values of φ_j or γ_i, the probe sensitivities ϕ_j and concentrations c_i *can never be negative*, because $\phi_j \equiv 2^{\varphi_j} > 0$; likewise for c_i.

It is an empirical fact that ratios of corresponding probe intensities between two samples, and other quantities derived from probe intensities, which are typically non-Gaussian distributed, become nearly normally distributed in the logarithmic space. We therefore proceed with the hope that Equation (2) will be better-behaved than Equation (1), and that the residuals ε_{ij} will be nearly normally distributed.

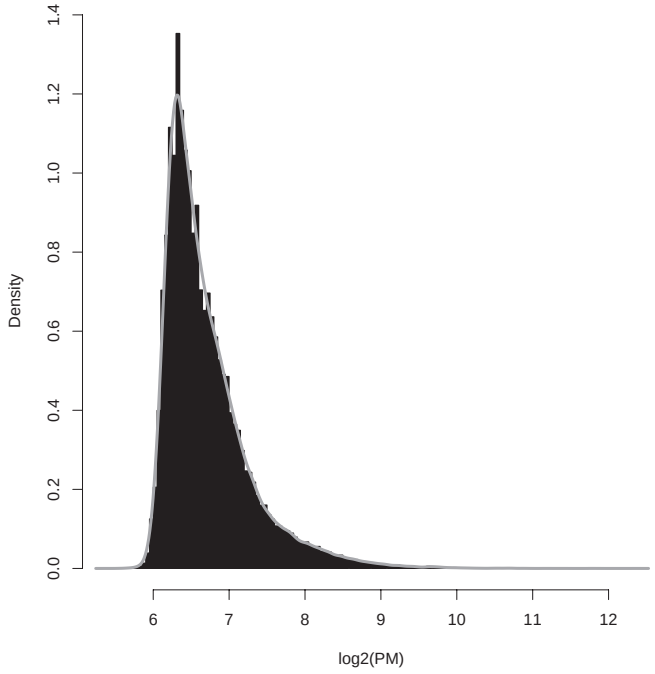

Fig. 1. Distribution of ($\log_2$) intensities of background PM cells.

Our program aims to find the best fit to Equation (2) by minimization of the sum of squares $\sum_{ij} \varepsilon_{ij}^2$. Then, if the distribution of residuals ε_{ij} is found to be significantly non-Gaussian, we eliminate the probe which contributes most to the sum of squares, and fit again Equation (2) using the reduced set of probes. We repeat this process until the distribution of residuals is not significantly different from Gaussian or the number of retained probes drops below a certain threshold. For mathematical convenience we choose the constraint on φ_j's to be $\sum_{ij} \mu_{ij}\varphi_j = 0$, where μ_{ij} is a function which is unity if the probe (i, j) is retained in the probe set and zero if it has been eliminated from the probe set.

When analyzing a batch of n_s samples, we first normalize all samples to each other using quantile normalization (Bolstad, 2002). In order to assess the background fluorescence in each sample, we adopt the procedure first proposed by (Naef *et al.*, 2001), in which a PM probe is considered to be a background probe when its intensity is within a δ-neighborhood of the corresponding MM probe intensity. The distribution of background probes (Figure 1) is not very sensitive to the value of δ—we use $\delta = 50$. This step is the only one in which MM probes are used. Next we subtract the mode b of the background probe intensity from every PM probe in the array. The probes whose intensity becomes negative as a result of background subtraction are eliminated (their μ_{ij} is set to zero) even before the first iteration is done.

Note that we do not simply dispose of the mismatch probes—they are used to estimate the level of background b, which is subsequently subtracted from all PM probes. Complete disposal of mismatch probes would lead to a loss of sensitivity for small-abundance transcripts, as demonstrated by a calibration experiment (Affymetrix, 2002). It is conceivable however that our method of subtracting flat background b is less than ideal in the very low-intensity region. This issue will be investigated in the future.

There are other authors (Efron *et al.*, 2002) who propose another way of dealing with the mismatch probes, in which the mean of $\log(PM_{ij}) - c \cdot \log(MM_{ij})$ (taken over the probe set) is suggested as the expression index. Here c is some constant (the authors suggest $c = 1/2$), which is found empirically. Since averaging is done over the probe set regardless of the probes' individual properties, this approach is not fundamentally different from that of AMS. Furthermore, it is not clear what the physical interpretation of the mismatch probes is in their prescription.

Our assessment of the normality of the residuals is based on the fact that a sum of squares of n independent normal variables with zero mean and unit variance has the χ^2-distribution with n degrees of freedom (Steel and Torrie, 1960). To this end we construct a variable $s_\sigma^2 = \sum_{ij} \varepsilon_{ij}^2/\sigma^2$, where σ is the standard deviation of the Gaussian 'core' of the residual distribution. We estimate σ as the interquartile range of the residual distribution divided by 1.3489... (this relationship holds true for a Gaussian distribution). Of course the residual distribution is generally non-Gaussian, but here we merely use this estimation of σ because the interquartile range is insensitive to outliers. When s_σ^2 is significantly large, i.e. when its p-value $p \equiv P(\chi^2 \geq s_\sigma^2)$ as derived from the χ^2-distribution is small, the residual distribution is deemed significantly non-Gaussian (has outliers) and the process of elimination of the worst offending probe continues. Iterations stop when the p-value of s_σ^2 is sufficiently large (we use $p > 0.2$), or the number of retained probes drops below a threshold (we use n_s). Since both σ and s_σ^2 are derived from the same data, this is a self-consistency test for normality of the residuals.

How this procedure works is depicted graphically in Figure 2. We show quantile–quantile plots of the residuals against the normal distribution. Figure 2a is the plot for a strongly expressed gene (probe ID 31330_at) after the initial fit to the model. All $16 \times 7 = 112$ PM probes were used in this fit. Deviations from normality are obvious and $p = 1.5 \times 10^{-21}$. Figure 2b displays the final distribution of residuals when 17 probes had been eliminated by the algorithm outlined above. The residual distribution is much closer to normal, and $p = 0.43$. In the like manner, Figure 2c displays the Q–Q plot of the residuals against the normal distribution for a gene which

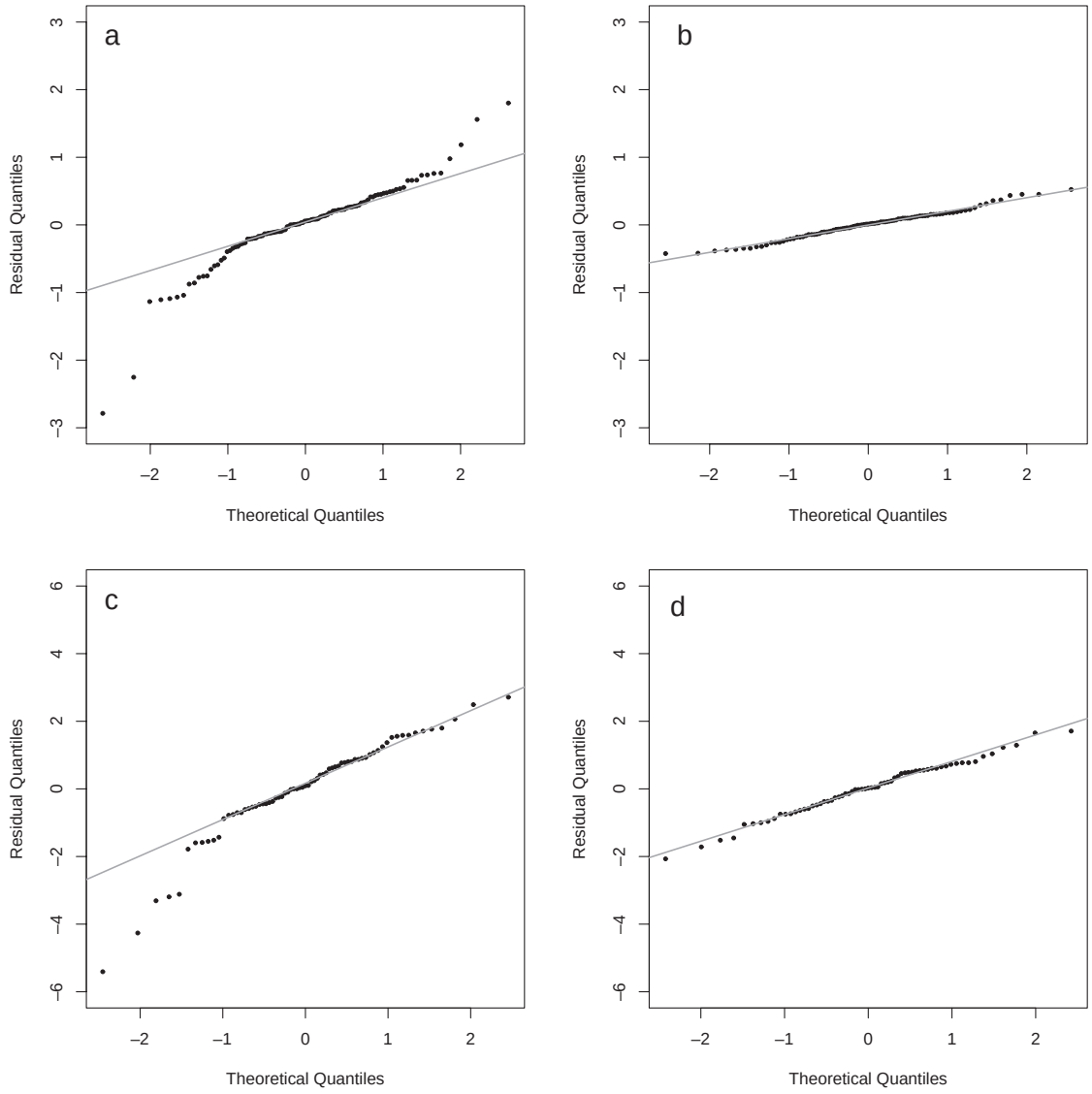

Fig. 2. Normal quantile-quantile plots of residuals ε_{ij} for probes 31330_at (a,b) and 160029_at (c,d), before (a,c) and after (b,d) iterations. The black lines pass through the first and third quartiles of the residual distribution.

is not significantly expressed (probe ID 160029_at). The initial number of probes used was 71 (the remaining 41 had intensities below background and were rejected), and $p = 4.3 \times 10^{-5}$. After elimination of 6 outlier probes (Figure 2d), $p = 0.76$.

After a few iterations, when the residual distribution has become approximately Gaussian, we calculate n_s p-values of a gene as follows: First, we calculate the p-value for every one of the $n_p \times n_s$ probes. Probes rejected from the fit are assigned a p-value of 1, and the remaining probes have p-values equal to the probability that a randomly chosen background cell has intensity greater or equal to theirs. This probability is found from the background probe distribution specific to that sample, such as the one in Figure 1. The heavy tail of this distribution at high intensity is probably caused by GC-rich probe pairs whose PM and MM probes both bind RNA with high affinity at the hybridizing temperature which is lower than optimal.

If the probe intensities were independent of each other, the p-value of a probe set would be given by the product of p-values of the probes in that set. We take the p-value of a probe set to be the product of the p-values of its constituent probes. This way the p-value of a probe set is the probability that a transcript is absent from the mixture, as measured simultaneously by all probes.

We now proceed with comparison of the present method

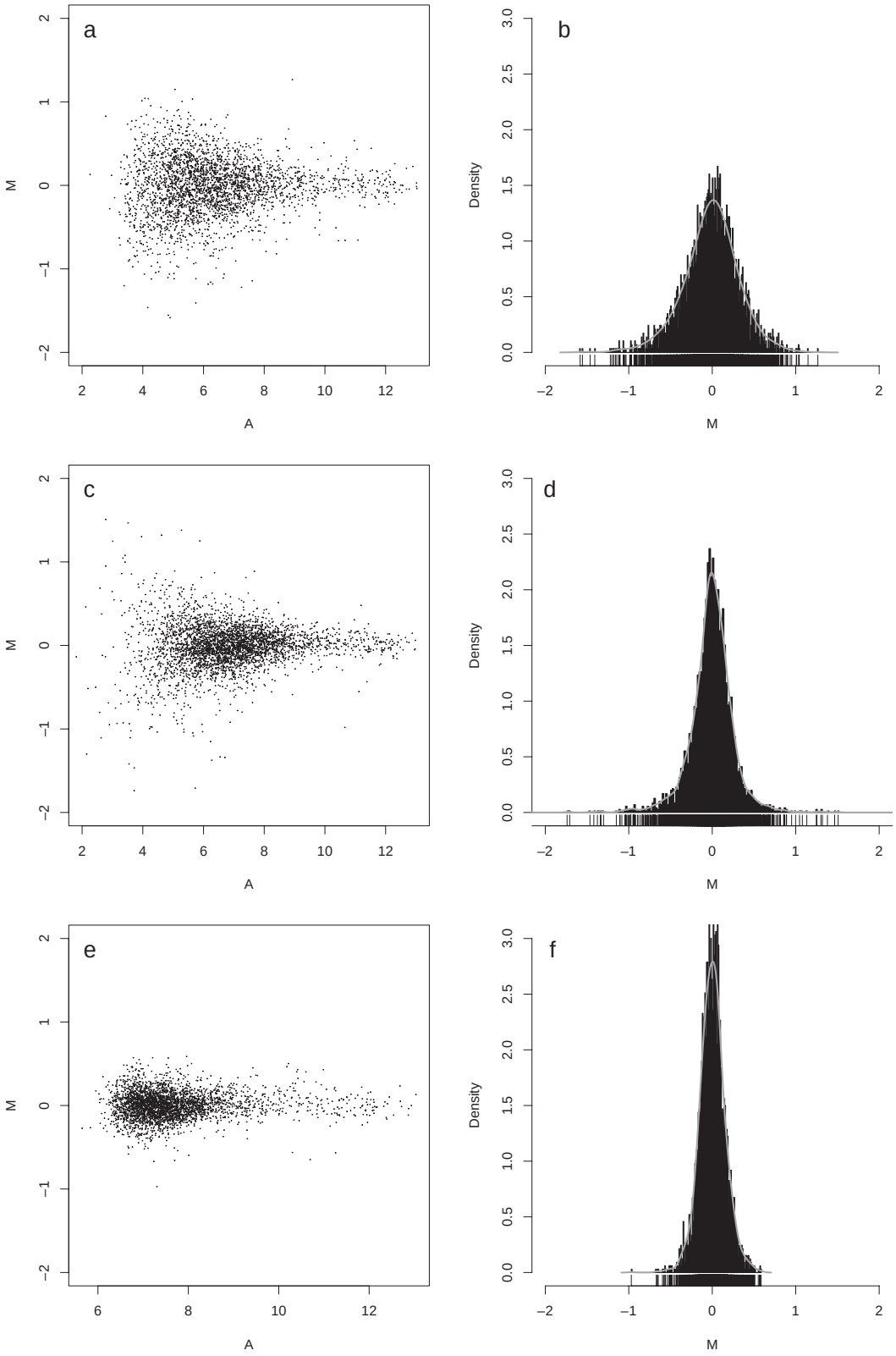

Fig. 3. The $A - M$ plots and distributions of M for the AMS 5.0 method (a,b), the LW method (c,d), and the present method (e,f). The AMS method produces 29% of points outside the 1.25-fold range, the LW method 15%, and the present method 4.4%.

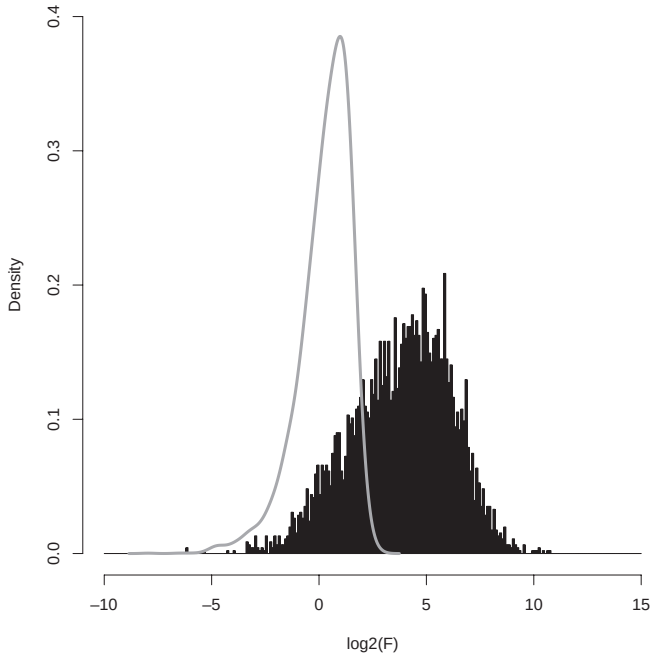

Fig. 4. Distribution of F-statistics for 4553 genes involved in the mouse study (black histogram). The black line traces the distribution of F-statistics obtained after a random permutation of sample labels.

of analysis with two established methods, the AMS and LW. To this end, we use an experimental standard to which all three methods will be compared. The standard in this case will be two samples, 1 and 2, hybridized with the same RNA solution. These two samples along with another five samples hybridized with RNA obtained from the same tissue type, harvested under a variety of conditions, form a set of seven to which we applied all three algorithms. Samples 1 and 2 are expected to report equal amounts of RNA for all genes. We will therefore compare measurement taken on sample 1 against sample 2, for all expressed genes.

Figure 3a shows the $A - M$ plot for the AMS method. Here $A \equiv \log_2(A_1 * A_2)/2$ and $M \equiv \log_2(A_2/A_1)$, where A_1 and A_2 are the average differences reported by the AMS software for a particular transcript in samples 1 and 2. Here we used only genes identified as present by the AMS software in both samples. A total of 2869 data points were plotted. Ideally, the points should line up at the $M = 0$ horizontal. The data have the characteristic spread as the agreement with $M = 0$ becomes progressively worse at low average differences. Figure 3b shows the histogram of M.

Figure 3c shows the $A - M$ plot for the LW method. Here A_1 and A_2 are the expression indices found by this method. Only genes identified as present in both samples by the LW software were used. It turns out that we had to further screen out genes with negative expression indices. A total of 3522 data points were plotted. The characteristic splay at low expression indices is there, but there is a marked improvement over the AMS method, as shown in the histogram of M (Figure 3d).

Figure 3e shows the $A - M$ plot obtained by the present method. Here A_1 and A_2 are the fitted concentrations c of transcripts in samples 1 and 2. We plotted data points for genes whose p-value was no greater than $(0.1)^{16}$ in both samples. This criterion selected a total of 3517 genes.

Inspection of histograms (Figure 3b, d, and f) and the corresponding scatterplots reveals that the LW method performs better than the AMS method in reducing the variance, and that the present method performs better than either the AMS or LW methods in this respect. Quantitatively, the number of genes with $|M| > \log_2(1.25)$ is 833 (29%) for the AMS 5.0 method, 545 (15%) for the LW method, and 156 (4.4%) for the present method.

Reduction of variance is certainly a desirable feature. However, one also has to show that the signal has not been reduced in the same proportion, in essence defeating our purpose. To this end we analyzed a series of experiments with murine RNA, in which there are six replicates for each of four different experimental conditions (classes). The four classes are characterized by different periods of time elapsed between injection of GDX and necroscopy, and by different tissue types. Each six replicates come from scanning three different microarrays—hybridized with the same cRNA—twice; once 'as is', and once after adding fluorescently-labeled antibody, which increases the fluorescent signal about two-fold. In order to determine whether there are any genes that are differentially expressed between any of the four classes, we perform analysis of variance (ANOVA), and compute the F-statistics for all genes (we use the logarithm of expression indexes in the calculation of F). We restrict ourselves to a subset of 4553 genes for which the product of p-values in any of the four classes was less than $(0.03^{16})^6$. The F-statistic is always positive, and in general large F-statistics indicate differential expression between at least two of the four experimental conditions. The F-statistic is a quantitative measure of bias versus variance—low within-class variance and large between-class bias makes for a large F and vice versa—and therefore very suitable for our purpose to objectively compare the three methods.

First we need to convince ourselves that there are differentially expressed genes in this experiment. In Figure 4 we plot the distribution of F-statistics as obtained in this experiment (black histogram), and the distribution of the same when the class labels have been randomly permuted. A random permutation of labels should create large within-class variance and, as a consequence, the F-statistics should drop. This is indeed the case. It is

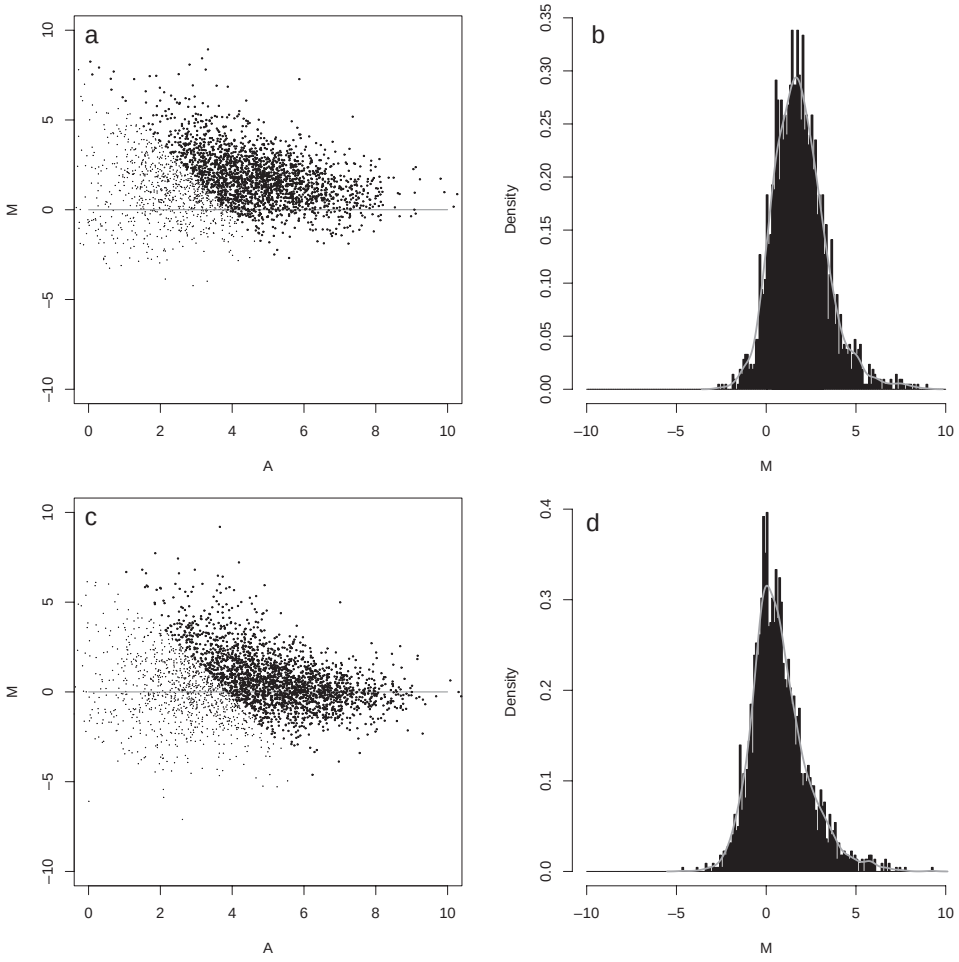

Fig. 5. The $A - M$ plots of F-statistics and distributions of M for significantly differentially regulated genes (bold points, adjusted p-value less than 0.05). Here $A \equiv \log_2(F * F')/2$ and $M \equiv \log_2(F/F')$, where F is the F-statistic obtained in the present method, and F' is the same for the AMS method (panels a, b) or the LW method (panels c, d).

obvious that a large fraction of genes in this study are differentially expressed. In order to find the subset of genes that are differentially expressed with statistical significance, we first calculate the unadjusted p-values for all F-tests using a permutation algorithm. In this algorithm the class labels are randomly permuted 10 000 times, the F-statistics are calculated, and the (unadjusted) p-values are given by the fraction of instances in which the permutation F-statistics were greater or equal to the F-statistics before permutation. Because of the large number of statistical tests performed (equal to the number of genes), adjustment of p-values for multiple testing is in order. We use the step-down permutation algorithm of (Westfall and Young, 1993), as described by Callow *et al.* (2000). Genes whose *adjusted* p-values were smaller than 0.05 (a total of 2412) were considered significantly differentially expressed. These genes are denoted by bold points in Figure 5a, c. The x-axes on these figures are the ($\log_2$) geometrical means of the F-statistic calculated from the present method, and F' calculated from AMS data (Figure 5a) and LW data (Figure 5c). The y-axes are the ($\log_2$) ratios of F and F'. The right-hand panels (Figure 5b, d) are histograms of ($\log_2$) ratios of F and F' for the 2412 significantly differentially expressed genes. In both instances (AMS and LW), both the mean and median F/F' ratio are positive, which means that *on average* the present method performs better in detecting differentially expressed genes than any of the other two methods considered.

DISCUSSION

The median F/F' ratio for the AMS 5.0 method is $2^{1.75} = 3.4$, and that of LW is $2^{0.52} = 1.4$. Such significant improvement, especially with respect to the AMS method,

is apparently due to our algorithm's ability to reduce variance while maintaining the signal. As another aspect of the same issue, 92% of the total number of significantly differentially expressed genes have a higher F-statistic in the present method compared to AMS 5.0, and 66% have a higher F-statistic in the present method compared to the LW method. The LW method compares fairly well to our method in the F test, but has more variance (Figure 3c, d). We think that the LW method could be improved by considering a uniform background b instead of the probe-dependent background it currently uses.

We conclude that overall the present method of analysis is a substantial improvement over the AMS method, and a significant improvement over the LW method.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Allergy and Infectious Diseases AI46237 and AI47703, the Center for AIDS Research Genomics Core Laboratory (AI36214), the Universitywide AIDS Research Program IS99-SD213 and the San Diego Veterans Medical Research Foundation. We thank Dr Fernand Labrie and Jean Morrissette for comments and suggestions.

REFERENCES

Affymetrix (2002), Technical note, http://www.affymetrix.com/products/25mer_content.html.

Bolstad,B. (2002) Probe level quantile normalization of high density oligonucleotide array data, (unpublished manuscript).

Callow,M.J., Dudoit,S., Gong,E.L., Speed,T.P. and Rubin,E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, **10**, 2022–2029.

Chudin,E., Walker,R., Kosaka,A., Wu,S.X., Rabert,D., Chang,T.K. and Kreder,D.E. (2002) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.*, **3**, research0005.1–0005.10.

Efron,B., Tibshirani,R., Goss,V. and Chu,G. (2002) Microarrays and their use in a comparative experiment, (unpublished manuscript).

Hill,A.A., Brown,E.L., Whitley,M.Z., Tucker-Kellogg,G., Hunter,C.P. and Slonim,D.K. (2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.*, **2**, research0055.1–0055.13.

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.

Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Naef,F., Lim,D.A., Patil,N. and Magnasco,M.O. (2001) From features to expression: High-density oligonucleotide array analysis revisited http://xxx.lanl.gov/ps/physics/0102010.

Naef,F., Hacker,C.R., Patil,N. and Magnasco,M. (2002a) Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.*, **3**, preprint0001.1–0001.24.

Naef,F., Lim,D.A., Patil,N. and Magnasco,M. (2002b) DNA hybridization to mismatched templates: a chip study. *Phys. Rev. E*, **65**, 040902(R).

Steel,R.G.D. and Torrie,J.H. (1960) *Principles and Procedures of Statistics*. McGraw-Hill, New York.

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Resting: Examples and Methods for p-value Adjustment*. Wiley, New York.