Biostatistics (2004), **5**, 2, *pp.* 155–176
Printed in Great Britain

Detecting differential gene expression with a semiparametric hierarchical mixture method

MICHAEL A. NEWTON*

Department of Statistics, University of Wisconsin–Madison, 1210 West Dayton St., Madison, WI 53706-1685, USA and Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, 600 Highland Ave., Madison, WI 53792, USA

newton@stat.wisc.edu

AMINE NOUEIRY

Institute for Molecular Virology, University of Wisconsin–Madison,1525 Linden Drive Madison, WI 53706, USA

DEEPAYAN SARKAR

Department of Statistics, University of Wisconsin–Madison, 1210 West Dayton St., Madison, WI 53706-1685, USA

PAUL AHLQUIST

Institute for Molecular Virology, University of Wisconsin–Madison,1525 Linden Drive Madison, WI 53706, USA and Howard Hughes Medical Institute, USA

SUMMARY

Mixture modeling provides an effective approach to the differential expression problem in microarray data analysis. Methods based on fully parametric mixture models are available, but lack of fit in some examples indicates that more flexible models may be beneficial. Existing, more flexible, mixture models work at the level of one-dimensional gene-specific summary statistics, and so when there are relatively few measurements per gene these methods may not provide sensitive detectors of differential expression. We propose a hierarchical mixture model to provide methodology that is both sensitive in detecting differential expression and sufficiently flexible to account for the complex variability of normalized microarray data. EM-based algorithms are used to fit both parametric and semiparametric versions of the model. We restrict attention to the two-sample comparison problem; an experiment involving Affymetrix microarrays and yeast translation provides the motivating case study. Gene-specific posterior probabilities of differential expression form the basis of statistical inference; they define short gene lists and false discovery rates. Compared to several competing methodologies, the proposed methodology exhibits good operating characteristics in a simulation study, on the analysis of spike-in data, and in a cross-validation calculation.

Keywords: Empirical Bayes; False discovery rate; Microarrays; Ordered alternatives; Posterior probability; Statistical analysis.

*To whom correspondence should be addressed.

1. INTRODUCTION

Microarray technologies assay a cellular state by measuring gene expression in parallel for many genes. All technologies assess the abundance of gene-specific messenger RNA molecules by taking advantage of the biochemical process of hybridization, but different technical approaches are used, including spotted cDNA microarrays and oligonucleotide microarrays (e.g. Nguyen *et al.*, 2002). Microarrays present some challenging statistical problems, in part because there are many sources of variation in the measurement process, and in part because the number of measurements per gene is usually quite small compared to the number of genes (e.g. Parmigiani *et al.*, 2003). A seemingly elementary statistical question is, 'Who is up/down?' (Speed, 2002, 13th problem). That is, in a comparative analysis between two cellular states, which genes are up-regulated, which ones are down-regulated and which ones remain unchanged in their expression levels? It may be necessary to report a short list of differentially expressed genes, to rank order the genes by evidence favoring differential expression, or to estimate the extent of differential expression in terms of an overall proportion of affected genes. In spite of considerable research on the differential expression problem, existing statistical methods are limited in the amount of reliable information they can extract from microarray data.

A small experiment to study genetic translation in the yeast *Saccharomyces cerevisiae* provides a motivating example. Like many microarray studies, this one involves cells grown in controlled conditions and there is a limited amount of replication. A wildtype yeast strain (WT) is being compared to a mutant strain (MUT) that is identical to WT except for a single mutation in DED1, a gene implicated in translational activation (Noueiry *et al.*, 2000). Knowing which genes are differentially expressed between MUT and WT provides insight into the function of DED1. Also, it is important to understand the direction and magnitude of effects and how these change according to a certain treatment on the RNA. Part of the experiment, which we take up in Section 6, involves three replicate Affymetrix S98 microarrays probing the WT transcripts and three replicate microarrays probing the MUT transcripts. After normalization and probe-set summarization, gene-specific inferences immediately must cope with the multiple three versus three comparisons. Nonparametric, permutation-based approaches are hampered by the very small sample size. There is also a concern that gene-specific tests such as the t-test are overly conservative and will not have the sensitivity to identify the genes of interest.

The formation of gene-specific summary statistics provides a basis for the rank ordering genes and the creation of short lists of genes inferred to be differentially expressed; obviously the precise form of the summary statistic affects this inference. One-dimensional, gene-specific summary statistics are usually isolated from each other in the sense that evaluation of the statistic for one gene does not use data from any other genes (e.g. gene-specific t-statistics, Dudoit *et al.*, 2002). By contrast, information sharing can be beneficial, because it can counteract the effects of low sample size (e.g. regularized t-statistics, Baldi and Long, 2001; Tusher *et al.*, 2001). To the extent that some methods enable information sharing, they do so in an elementary way that may belie the complex patterns of variation in microarray data. It is noteworthy that gene-specific statistics which use data from across the genome arise as a biproduct of hierarchical statistical modeling.

Whatever method is used to summarize gene-level data and to create a rank ordering of genes, there remains the problem of inferring which genes are differentially expressed: in other words, the problem of forming a short list by selecting the top ranking genes. This may be of secondary concern if the laboratory is interested in looking at, say, the top ten genes, regardless of presented variation. Typically, however, there is interest to calibrate the list so that errors are controlled in some way (see Dudoit *et al.*, 2003). Behind most approaches is the concept of a true list relative to which we may control some measure of type I and type II errors. In doing so we are obliged to introduce probability. Methods that rely on label permutation for this purpose are appealing because their conditional inference can provide error rate control which is exact and which does not depend on among gene dependence or any details of the distribution of expression values. However, when the number of replicate microarrays is limited, as for example in the yeast experiment, probabilistic statements based on label permutation become less effective because of the coarse distributions involved. There are only ten ways to divide six microarrays into two equal-sized groups, for example. Both gene-specific permutation-based hypothesis tests, and the nonparametric mixture approach of Efron *et al.* (2001) are affected adversely by the low complexity of the permutation distribution caused by limited replication.

An alternative strategy is to adopt some assumptions about the distribution of the expression measures, such as within-gene log-normality and independence of replicate measurements on a null hypothesis of equivalent expression. Another useful and fairly innocuous assumption is the discrete mixture model, which is becoming popular in microarray data analysis (Newton *et al.*, 2001; Efron *et al.*, 2001; Allison *et al.*, 2002; Broët *et al.*, 2002; Lee *et al.*, 2002; Lonnstedt and Speed, 2002; Pan, 2002; Kendziorski *et al.*, 2003; Storey and Tibshirani, 2003). Each gene is viewed as tossing a coin to decide *a priori* whether or not it will be differentially expressed. The success probability of the coin represents the unknown fraction of genes that are truly differentially expressed, and the outcome of each coin toss can be assessed only indirectly through measurements of expression: the differentially expressed genes present data according to a different distribution than the equivalently expressed genes. The most frequently considered case involves a mixture model on one-dimensional gene-specific p-values: given equivalent expression, these p-values are uniformly distributed on (0, 1) (usually after further modeling assumptions), but otherwise they ought to tend towards the origin. An advantage of discrete mixture modeling is that it provides a direct method to control the average rate of type I errors on the reported list—the false discovery rate (FDR). This can be done via p-value analysis or by more explicit modeling and the construction of gene-specific posterior probabilities of differential expression (see Section 3). By avoiding the reduction of gene-level data to p-values prior to mixture modeling, the latter approach may retain a degree of sensitivity.

To address the differential expression problem, we propose a methodology based on a hierarchical mixture model. The model is hierarchical in the sense that an observation component describes the conditional distribution of measurements given expected expression values, and a second component (the mean component) describes the distribution of these expected expression values. Such hierarchical modeling enables the sharing of information among genes; genes become linked by virtue of having expected expression values drawn from a common, albeit unknown, probability distribution. The rationale for using such random gene effects is that sensitivity may improve when we have very little information per gene. Inferences concerning the ranking of genes are directly affected by the form of this random effects distribution. We consider both parametric and nonparametric forms for the distribution of expected expression values. Although the parametric form produces analytically tractable inferences, the nonparametric form improves overall model fit and may give more robust inferences. We also adopt a discrete mixture model over patterns of differential expression. This enables the calculation of gene-specific posterior probabilities of differential expression and the reporting of gene lists with targeted FDR.

In what follows, Section 2 develops the model and summarizes estimation, and Section 3 reviews the method of gene-specific inference by posterior probability. Section 4 reports a small simulation study comparing the proposal to gene-specific t-testing and the mixture method of Efron *et al.* (2001). We make further comparisons on data from a spike-in experiment in Section 5. Section 6 reports the analysis of the yeast translation experiment, and a discussion follows in Section 7. Details of the parametric sub-model, the EM algorithm, and numerical evaluations are contained in an appendix. We also contribute **R** language code to implement the calculations (Ihaka and Gentleman, 1996).

2. Hierarchical mixture model

Let $\mathbf{x}_g = (x_{g,1}, \dots, x_{g,m})$ denote the replicate, normalized expression measurements on gene g in the first cellular state (i.e. condition) and $\mathbf{y}_g = (y_{g,1}, \dots, y_{g,n})$ the replicate, normalized measurements in the second condition. Expectations $\mu_{g,1} = E(x_{g,i})$ and $\mu_{g,2} = E(y_{g,i})$ are assumed to mediate differential expression in the sense that given these parameter values, the measurements in $\mathbf{x}_g$ form a random sample (independent and identically distributed) from an observation component $p(x_{g,i}|\mu_{g,1})$, and likewise $\mathbf{y}_g$ is an independent random sample from $p(y_{g,i}|\mu_{g,2})$. The calculations reported here use a Gamma observation component in which the shape parameter (hence, coefficient of variation CV) is constant across genes within a condition, but may vary between conditions (see Appendix, Section A.1) although other formulations such as log-normal may be beneficial (Kendziorski *et al.*, 2003). The Gamma distribution is analytically tractable, has some theoretical appeal for expression data, and fits well in our experience (Newton *et al.*, 2001). Further, many proposed models of expression variability encode approximately constant CV (Chen *et al.*, 1997; Ideker *et al.*, 2000; Baggerly *et al.*, 2001; Li and Wong, 2001; Rocke and Durbin, 2001; Theilhaber *et al.*, 2001; Tsodikov *et al.*, 2002). Our experience with various normalization procedures is that nearly constant CV is common, although there may be excess variation at the lowest expression levels and reduced variation at the highest levels.

Next, expectations $(\mu_{g,1}, \mu_{g,2})$ are themselves considered to be a random pair drawn from an unknown bivariate distribution f. (This is in contrast to a fixed-effects approach, e.g. Kerr *et al.*, 2000.) We organize f as a discrete mixture over three potentially interesting hypotheses: equivalent expression, $H_{g,0} : \mu_{g,1} = \mu_{g,2}$, under-expression in the first cellular state, $H_{g,1} : \mu_{g,1} < \mu_{g,2}$, and over-expression in the first cellular state $H_{g,2} : \mu_{g,1} > \mu_{g,2}$:

$$f(\mu_{g,1}, \mu_{g,2}) = p_0 f_0(\mu_{g,1}, \mu_{g,2}) + p_1 f_1(\mu_{g,1}, \mu_{g,2}) + p_2 f_2(\mu_{g,1}, \mu_{g,2}). \tag{2.1}$$

Scalers p_0, p_1, and p_2 give the marginal proportion of genes satisfying each of the three hypotheses. The densities f_0, f_1, and f_2 describe fluctuations of the means within each hypothesis. Though it is tempting to place no further structural constraints on f it is necessary to do so in order that all the components are estimable. We do so by relating the joint distribution f to a one-dimensional *base* distribution π which generates the gene and condition-specific expected values. Gene g draws its μ as follows: two independent draws U_g and V_g arise from π, and a three-sided die with probabilities (p_0, p_1, p_2) is cast to give the discrete outcome Z_g. If the die comes up for $H_{g,0}$, then $\mu_{g,1} = \mu_{g,2} = U_g$, and V_g is ignored. If the die comes up for $H_{g,1}$, then $\mu_{g,1} = \min(U_g, V_g)$ and $\mu_{g,2} = \max(U_g, V_g)$, and vice versa for if the die comes up for $H_{g,2}$. Thus, $f_0(\mu_{g,1}, \mu_{g,2}) = \pi(\mu_{g,1})\, 1[\mu_{g,1} = \mu_{g,2}]$, $f_1(\mu_{g,1}, \mu_{g,2}) = 2\pi(\mu_{g,1})\pi(\mu_{g,2})\, 1[\mu_{g,1} < \mu_{g,2}]$, and $f_2(\mu_{g,1}, \mu_{g,2}) = 2\pi(\mu_{g,1})\pi(\mu_{g,2})\, 1[\mu_{g,1} > \mu_{g,2}]$, with the factor of 2 coming from the order constraint and where $1[\,]$ is the indicator function. This device amounts to saying that, up to ordering, the assertion of differential expression is the assertion that expected expressions are unrelated in the sense that they are stochastically independent. The assumed independence of Z_g from the pair (U_g, V_g) encodes the idea that differential expression itself is unrelated to the overall level of expression; in the absence of special information this seems to be a reasonable starting position.

To fit the hierarchical mixture model is to obtain estimates of the mixing proportions (p_0, p_1, p_2), the base distribution π, and any unspecified parameters in the observation component. We have implemented two methods of estimation which differ by their treatment of the base distribution π:

1. Parametric π: The two-parameter inverse Gamma distribution for π is conjugate to the Gamma observation component (Section A.1). Implement maximum likelihood estimation for all parameters via an EM algorithm in which the missing data are the outcomes Z_g of the gene-specific three-sided die tosses.
2. Nonparametric π: For numerical stability, treat U_g and V_g as logarithms of expected expression values, rather than straight expected values. Fix a dense, equally spaced grid to represent the support of π (e.g. 500 points spanning the range of the log data). Thus, π is represented as a probability vector on this support. Implement (nonparametric) maximum likelihood via an EM algorithm for π and (p_0, p_1, p_2) in which the missing data are the pairs (U_g, V_g) and the outcomes Z_g. Fix the

shape parameters of the observation component at estimates obtained separately by the method of moments.

Further details of the EM algorithms are in Section A.2. Regarding the observation component parameters, note that the proposed model entails a Gamma observation component with shape parameter a_j in condition $j = 1, 2$. The coefficient of variation (standard deviation)/mean $= 1/\sqrt{a_j}$, so we see that the shape parameters can be estimated by a method of moments using the collection of within-condition, gene-specific sample coefficients of variation $\mathrm{cv}_{j,g}$ computed from the normalized data. The estimate $\hat{a}_j = 1/\mathrm{mean}(\mathrm{cv}^2_{j,g})$ works well in simulations. In the fully parametric case (1) above, these shape estimates are used as starting values for the EM algorithm which is targeting the full maximum likelihood estimates. In case (2), we maintain the method-of-moment estimates throughout and update only π and (p_0, p_1, p_2) during EM iterations.

The fully parametric case (1) is similar to the Gamma–Gamma (GG) methodology described in Kendziorski *et al.* (2003), except that here we allow ordered alternative hypotheses. It is useful in the yeast example and elsewhere to allow ordered alternatives, but doing so creates an interesting technical problem because the latent variables (U_g, V_g) are harder to integrate away (see equation A.4 in Section A.1). The fully parametric model has $1/U_g$ and $1/V_g$ distributed according to a Gamma distribution with shape parameter a_0 and rate parameter $a_0 x_0$. Conveniently in this parametrization, the parameter x_0 is a measure of location in the marginal distribution of the measurements. Case (2) involves a flexible nonparametric model for π coupled with the Gamma observation component, and so it is naturally referred to as a semiparametric model; the grid-based EM algorithm is a convenient approach to nonparametric likelihood estimation of π (Lindsay, 1995).

One way to assess goodness of fit in either case is to compute the marginal distribution induced by the hierarchical model and compare this to the empirical marginal distribution of measurements. For example, computing the marginal distribution of measurements $x_{g,i}$ in the first condition requires integrating the Gamma observation component against the fitted f from (2.1),

$$p(x_{g,i}) = \int p(x_{g,i}|\mu_{g,1})\, p(\mu_{g,1})\, \mathrm{d}\mu_{g,1},$$

where $p(\mu_{g,1}) = \int f(\mu_{g,1}, \mu_{g,2})\, \mathrm{d}\mu_{g,2}$. In case (1), this is available analytically as the univariate compoound Gamma distribution (equation A.2). It can be computed numerically in case (2).

3. Inference

Having fitted the hierarchical mixture model, gene-specific inference is based on posterior probabilities

$$P(H_{g,j}|\mathbf{x}_g, \mathbf{y}_g) = p_j\, p(\mathbf{x}_g, \mathbf{y}_g|H_{g,j})/p(\mathbf{x}_g, \mathbf{y}_g) \tag{3.1}$$

where

$$p(\mathbf{x}_g, \mathbf{y}_g|H_{g,j}) = \int\int p(\mathbf{x}_g|\mu_1)\, p(\mathbf{y}_g|\mu_2)\, f_j(\mu_1, \mu_2)\, \mathrm{d}\mu_1\, \mathrm{d}\mu_2 \tag{3.2}$$

and where $p(\mathbf{x}_g, \mathbf{y}_g) = \sum_j p_j\, p(\mathbf{x}_g, \mathbf{y}_g|H_{g,j})$ is the marginal (predictive) density of the data. As above, these integrals are available in closed form for the parametric version of the model (Section A.1, equation A.4), and they need to be computed numerically in the semiparametric version. A notable consequence of the EM fitting procedure is that the average gene-specific posterior probability of hypothesis j equals the estimated overall proportion of genes which satisfy this hypothesis, i.e. $p_j = \frac{1}{N}\sum_{g=1}^{N} P(H_{g,j}|\mathbf{x}_g, \mathbf{y}_g)$.

Insofar as the fitted mixture model is accurate and among-gene dependence can be ignored, the gene-specific posterior probabilities (3.1) form the basis of optimal statistical inference about differential expression. For example, Bayesian theory indicates that to minimize the probability of making a mistake, we ought to declare hypothesis $H_{g,j}$ true for gene g if that hypothesis has higher posterior probability than any of the other ones (e.g. Berger, 1985, p. 163). Similarly, optimal ranking and selection schemes are based on these marginal posterior probabilities. The inference problem might be to produce a list of genes for which evidence favors differential expression. For any fixed-size list, the optimal procedure is to form that list using those genes having the highest marginal posterior probability of differential expression. To do otherwise is to report a list with higher posterior expected loss, where the loss is the number of mistakes on the list. In this sense, the optimal ranking of genes is the ranking based on posterior probability of differential expression. A short list of differentially expressed genes is obtained by ranking from smallest to largest by $P(H_{g,0}|\mathbf{x}_g, \mathbf{y}_g)$ and cutting the list at some point chosen either for convenience (i.e. we only want to look at a fixed number of extreme genes) or in order to control the number of type I errors (false discoveries) in the list.

Suppose our goal is to identify a list J of genes g for which hypothesis $H_{g,j}$ is probably true (fix either $j = 1$ or 2), and we want the list to be as large as possible while bounding the rate of false discoveries by α. We use a *direct posterior probability approach* to achieve this goal. With data and fitted model in hand, we rank the genes according to increasing values of $\beta_g = 1 - P(H_{g,j}|\mathbf{x}_g, \mathbf{y}_g)$, and our reported list J contains genes g having values β_g less than some bound κ. Given the data, the expected number of false discoveries is

$$C(\kappa) = \sum_g \beta_g 1[\beta_g \leqslant \kappa]$$

since β_g is the conditional probability that placing gene g on the list creates a type I error. In the typical situation there will be some genes for which $\beta_g \leqslant \alpha$. As long as this is true, we can find a data-dependent $\kappa \leqslant 1$ as large as possible so that $C(\kappa)/|J| \leqslant \alpha$, where $|J| > 0$ is the size of the list. Notice that $C(\kappa)/|J|$ is the expected rate of false detections, given the data. By bounding the rate conditionally, we bound it on average over data sets. The bound is approximate, however, because it rests on the accuracy of the fitted model; one expects that careful modeling and diagnostic checking can reduce the effect of this approximation.

Alternative approaches to bounding the FDR use gene-specific p-values. In the calculations reported here, we compare the direct posterior probability approach to the algorithm of Storey and Tibshirani (2003) as it is applied to p-values from gene-specific t-tests. That method first transforms p-values to q-values; a list with target FDR level α is formed by including genes for which the q-value is bounded above by α.

Instead of testing for the presence of differential expression, the goal may be to estimate the magnitude of differential expression. The hierarchical mixture model may be adapted to this purpose. For example, the target fold change $\rho_g = \mu_{g,1}/\mu_{g,2}$ may be estimated by the empirical ratio $\bar{x}_g/\bar{y}_g$, or a model-based estimate may be computed. Posterior skewness suggests that we work on the scale of $\log(\rho_g)$; its posterior expectation is the Bayes estimate under squared-error loss and may be computed by numerical integration.

4. SIMULATION STUDY

We performed a simulation study to evaluate the proposed methodology. The three scenarios discussed here were designed to present variation similar to what we have observed in practice, although they are necessarily overly simplified representations. In each scenario, we have $N = 5000$ genes, $m = n = 3$ replicates per condition, and a Gamma observation component with shape parameters $a_1 = a_2 = 20$ that are common to all genes. The scenarios differ in the status of the underlying mixing components in f:

I. Conjugate, Inverse Gamma, shape $a_0 = 2$, location $x_0 = 1000$.
II. Uniform on $5 \leqslant A = \log(\sqrt{\mu_{g,1}\mu_{g,2}}) \leqslant 11$ and $-2 \leqslant M = \log(\mu_{g,1}/\mu_{g,2}) \leqslant 2$; and $M = 0$ if $\mu_{g,1} = \mu_{g,2}$.
III. Uniform on $5 \leqslant A = \log(\sqrt{\mu_{g,1}\mu_{g,2}}) \leqslant 11$ and $-1 \leqslant M = \log(\mu_{g,1}/\mu_{g,2}) \leqslant 1$; and $M = 0$ if $\mu_{g,1} = \mu_{g,2}$.

Scenarios II and III involve uniform distributions on the (M, A) scale and this very roughly approximates the apparent relationship between the means in some examples. The (M, A) notation is borrowed from Dudoit *et al.* (2002), though we use it to describe underlying means rather than statistics (and we use a natural log scale rather than base 2). Scenario I fully matches the proposed parametric model, and so we expect the parametric methodology to perform best. Scenarios II and III retain the parametric observation component but violate the parametric mean component; in fact, they do not encode conditional independence of the means given differential expression.

For each scenario we considered different levels of differential expression. We report in Table 1 results for mixing proportion vectors $(0.9, 0.05, 0.05)$; $(0.8, 0.1, 0.1)$; two data sets were simulated in every case. Both parametric and semiparametric models were fitted to each data set, and posterior probabilities were computed for each gene over the three hypotheses of interest. For comparison, we computed standard two-sided gene-specific t-tests on the log-transformed measurements. The p-values were transformed to q-values following the method in Storey and Tibshirani (2003). We specified the inference problem to be to obtain a list of differentially expressed genes in which the target FDR is $\alpha = 0.05$. Knowing the underlying hypotheses for each gene, we obtained empirical estimates of sensitivity, specificity, and realized rates of false discovery and false non-discovery.

The fully parametric model is correct in scenario I, and yet the parametric and semiparametric methods have indistinguishable operating characteristics. Evidently little is lost by adopting such a flexible model for π. In scenarios II and III, the parametric method begins to lose sensitivity to the semiparametric method, as expected, but the loss of sensitivity is only marginal. For all methods, sensitivity is relatively low but increases with an increase in the overall rate of differential expression, and the specificity is always high in the cases considered. Both parametric and semiparametric methods target a 5% FDR, though the bound is approximate because of estimation error; according to the simulation results the FDR is well-controlled in the cases considered. The gene-specific t-test procedure bounds FDR, but it does so at a great expense in sensitivity. The t-test is extremely conservative in the cases considered.

We also applied the nonparametric mixture method from Efron *et al.* (2001) to each data set, and we refer to that as the ETST method. It is meaningful to make this comparison because ETST is one of the few mixture-based approaches in which gene-specific posterior probabilities, as opposed to p-values, are the primary object of inference. The ETST mixture method has a great deal of flexibility and requires some care in implementation (see Section A.3 for specifics). We applied it to all simulated examples; Figure 1 summarizes results for one typical case. The ETST method falters compared to the semiparametric method in the sense that the FDR of a region is overestimated; there is no non-empty list of genes with estimated FDR of 5%, for example. By contrast, the realized FDR is close to its target value for the semiparametric method.

The simulated data were also used to assess the accuracy of estimation of true fold change $\log(\mu_{g,1}/\mu_{g,2})$. As expected by shrinkage, the Bayes estimates exhibited a mean squared error that is smaller than that of log empirical fold change estimate, on average over genes. Most of this improvement derives from genes of modest effect; the quality of the Bayes ranking is very similar to the quality of the empirical fold ranking if we focus on the most differentially expressed genes (data not shown).

Table 1. *Operating characteristics, simulation study: The methods are S (semiparametric), P (parametric), and T (t-statistic).* $\#\hat{DE}/n$ *is the proportion of genes that are detected to be differentially expressed. Out of the truly differentially expressed genes, Sens. is the proportion that are detected. Out of the truly equivalently expressed genes Spec. is the proportion that were not detected as DE. FDR is the realized rate of false detections in the detected list and FNDR is the realized rate of false non-detections in the non-detected list. Results shown are averages over the two data sets simulated and fitted in each condition; variation is very small*

Scenario	$P(DE)$	Method	$\#\hat{DE}/n$	Sens.	Spec.	FDR	FNDR
I	0.1	S	0.07	0.62	1.00	0.05	0.04
		P	0.06	0.61	1.00	0.04	0.04
		T	0.02	0.24	1.00	0.03	0.08
	0.2	S	0.14	0.67	0.99	0.06	0.08
		P	0.14	0.67	0.99	0.06	0.08
		T	0.09	0.41	0.99	0.05	0.13
II	0.1	S	0.07	0.72	1.00	0.03	0.03
		P	0.07	0.71	1.00	0.03	0.03
		T	0.04	0.40	1.00	0.05	0.06
	0.2	S	0.15	0.74	0.99	0.04	0.06
		P	0.15	0.73	0.99	0.03	0.06
		T	0.12	0.56	0.99	0.04	0.10
III	0.1	S	0.04	0.34	1.00	0.01	0.07
		P	0.03	0.31	1.00	0.01	0.07
		T	0.00	0.00	1.00	NaN	0.10
	0.2	S	0.08	0.42	1.00	0.01	0.10
		P	0.07	0.38	1.00	0.00	0.13
		T	0.00	0.01	1.00	0.05	0.20

5. Spike-in study

We applied the proposed methodology to data from the Gene-Logic spike-in experiment in which a small set of genes is known to be differentially expressed (Antonellis *et al.*, 2002). In the Gene-Logic study, known concentrations of 11 cloned bacterial and phagemid RNAs were added to RNA derived from an acute myeloid leukemia (AML) cell line and the complexes were probed with U95A Affymetrix microarrays. From the full study we selected six arrays, three from the 25pM concentration group and three from the 50pM concentration group, so as to enable a two-group comparison (see Section A.3 for array labels). Probe-level data were analysed using the robust multi-array average (RMA) method (Irizarry *et al.*, 2003) to produce gene-level measurements, and then we applied the hierarchical mixture methodology.

Table 2 summarizes results for the 10 spike-ins which differ in true concentration between the two conditions by two-fold. To enable comparisons the table considers the rank (out of 12 626 targets) of these spike-ins as determined by different procedures. The simple fold change (log base 2 of ratio of within-group arithmetic mean expressions) ranks the spike-ins very well, as do both the parametric and semiparametric hierarchical methods. Of note is that the gene-specific t-test does a poor job ranking the spike-ins. The smallest q-value in this case is 0.49 and so no spike-ins can be found if we try to

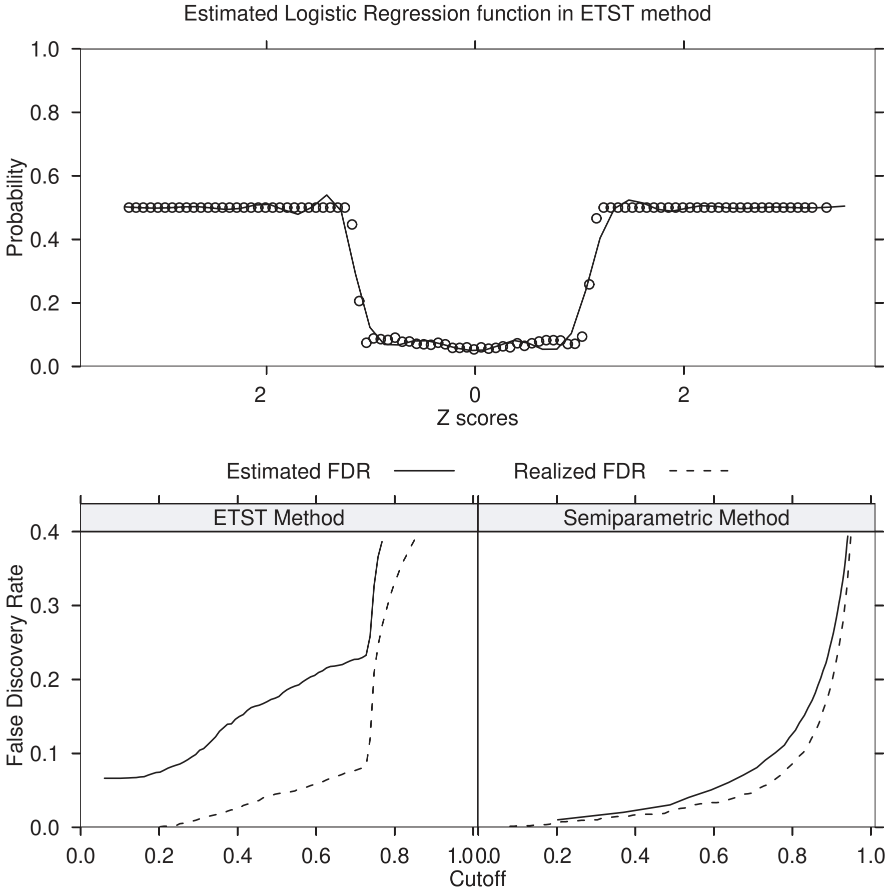

Fig. 1. ETST method applied to a simulated data set (scenario II, 40% DE): The plot on the top shows the logistic regression fit to the Z-scores. The range of the Z-scores is divided into 100 bins, and the points are the proportion in each bin of Z-scores from the mixture distribution. The fitted regression function is a natural spline with 20 d.f. The bottom plot summarizes the properties of the estimated FDR for both the ETST method and the semiparametric method. As can be seen, the estimated FDR in the ETST method starts quite high, and thus cannot be used to control the FDR at, say, 5%.

maintain a small FDR using p-values from the t-test. (This test uses log-normal measurements and equal variance between groups but different variance across genes.) In targeting a 10% FDR list by our proposed method, 8 of 10 spike-ins are on the semiparametric list and all 10 are on the parametric list. A 15% FDR-semiparametric list contains all 10 of the spike-ins. Further, ranking by Bayes estimated fold-change is very similar to the probability-based ranking.

6. Yeast translation experiment

To identify genes for which translation is affected by the DED1 gene, yeast mRNA was fractionated on a sucrose gradient and separated into a translating fraction associated with ribosomes and a non-translating, ribosome-free fraction following the approach of Johannes *et al.* (1999). Affymetrix Yeast

Table 2. *Spike-in study: column 1 has the spike-in label. Column pairs refer to empirical fold change, t-statistic p-value, semiparametric method and parametric method DE posterior probabilities. Shown are values and then rank out of 12 626 total targets*

Spike-in	Fold	F-rank	p-value	p-rank	SP-post	SP-rank	P-post	P-rank
AFFX-BioB-3-at	−0.79	7	0.026	154	0.999	9	0.999	11
AFFX-BioB-5-at	−0.86	4	0.009	46	1.000	5	1.000	5
AFFX-BioB-M-at	−1.17	1	0.010	52	1.000	1	1.000	1
AFFX-BioC-3-at	−0.88	2	0.010	51	1.000	2	1.000	2
AFFX-BioC-5-at	−0.82	6	0.016	101	1.000	6	1.000	7
AFFX-BioDn-3-at	−0.88	3	0.013	77	1.000	3	1.000	3
AFFX-DapX-3-at	−0.65	19	0.015	94	0.511	41	0.661	34
AFFX-DapX-5-at	−0.86	5	0.006	36	1.000	4	1.000	4
AFFX-DapX-M-at	−0.66	18	0.004	24	0.580	38	0.732	32
AFFX-CreX-5-at	−0.66	17	0.047	333	0.762	31	0.817	30

S98 microarrays representing all 6130 known yeast genes were used to probe the translating and non-translating fractions in both the mutant strain (MUT) and the wildtype strain (WT). The measurement process was repeated for three MUT replicates and three WT replicates in both translating and non-translating fractions to yield 12 microarrays. Separately for the two fractions, probe-level data were analysed using the RMA method to yield a single measure of expression for each gene on each microarray. A full characterization of genes affected by DED1 is underway. In the present methodologic study, we focus on the six non-translating microarrays (3 MUT and 3 WT); we use the data primarily for demonstration, noting that the level of replication and the patterns of variation may be common features of many expression studies.

Figure 2 shows that the sample coefficient of variation (CV) for the three non-translating WT microarrays does not have a strong systematic relationship with the sample mean expression. This is precisely what we expect if the measurement standard deviation increases linearly with the mean in the underlying data-generation process, as we have, for example, in the constant-shape Gamma model. Figure 3 reveals another property of the non-translating WT yeast measurements that also may be a common feature in microarray data. Each of the nine quantile–quantile plots compares data from a vertical band of Figure 2 to the quantiles of a fitted Gamma distribution. These diagnostic plots (and others not shown) show that the distribution of expression measurements is well approximated by the Gamma distribution when we focus locally on genes that have similar mean expression.

Figure 4 gives the parametric and semiparametric estimates of the base distribution π for the comparison of non-translating WT and MUT transcripts. The center and scale of the two estimates are similar, although, as expected by the allowed flexibility, the semiparametric estimate has much more detailed structure and fits well (Figure 5). Parameter estimates are reported in the caption of Figure 4. Shape parameters of the observation component are estimated to be smaller by the fully parametric method; this corresponds to bigger coefficient of variation and may be due to lack of fit by the parametric inverse-Gamma mean component. The model fit also entails estimated proportions of genes satisfying the three hypotheses: for the semiparametric fit we get $(p_0, p_1, p_2) = (0.554, 0.352, 0.093)$, which is slightly more differential expression than we estimate from the parametric model: (0.634, 0.291, 0.074). In either case the rate of differential expression may seem high, but one must remember that (a) DED1 is a critical translation factor and mutations could have substantial impact, and (b) the reported rates are overall proportions in the genome; they do not in themselves indicate the proportion of genes that are *significantly* differentially expressed.

One gene-specific question of interest is to ask, in the non-translating fraction of RNA, which genes are

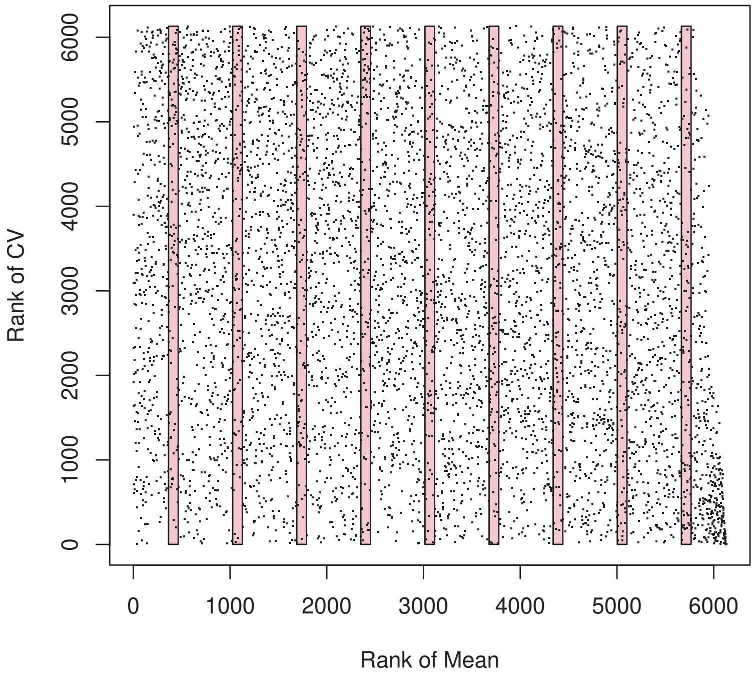

Fig. 2. Ranked CV vs. ranked Mean: $N = 6130$ points (genes) from WT data. From $m = 3$ replicate measures per gene (after normalization, raw scale) we compute the sample mean and CV which is the sample standard deviation over the sample mean. Ranking both axes spreads points. Non-uniformity indicates violation of constant CV assumption (e.g. most highly expressed genes exhibit slightly reduced variation). Strips contain 100 genes, for Figure. 3.

down-regulated in MUT compared to WT: i.e. $H_{2,g} : \mu_{1,g} > \mu_{2,g}$ in our notation where the **x** correspond to WT and **y** to MUT. The biological rationale, briefly, is that the mutant DED1 may be enhancing the efficiency of translation on these genes, depending on how they behave in the translating RNA fraction. As noted, the semiparametric estimate suggests that a proportion $p_2 = 0.093$ of genes are so down-regulated. Thus, a point estimate for the number of down-regulated genes is $572.4 = 6130 \times p_2$, though we may want to size a short list of down-regulated genes by considering FDR. For example, the largest list we can find with an estimated 5% FDR contains the 461 genes that have the highest $P(H_{2,g}|\mathbf{x}_g, \mathbf{y}_g)$. The cutoff happens to be $P(H_{2,g}|\mathbf{x}_g, \mathbf{y}_g) \geqslant 0.684$. It also happens that in the context of the fitted parametric model we obtain about the same cutoff for a 5% FDR list, though the list is smaller, having 351 genes; all of these genes are also on the list obtained by the semiparametric method, so there is good agreement and the semiparametric method is making more calls. For comparison we implemented one-sided gene-specific t-tests and formed a list of down-regulated genes according to the q-values that are bounded by 5%. This list contained only 33 genes, 31 of which are also on the semiparametric list.

The top panel of Figure 6 considers the model-based estimates of fold change as compared to the empirical fold change. As expected, most genes (94%) exhibit shrinkage towards $\rho = 1$ (upper left and lower right quadrants), although the attenuations are variable and genes with highest fold change show little shrinkage. In this example, genes with empirical fold more than about two-fold have a very high differential expression probability.

In the present statistical study we have no further molecular validation of the model predictions, but we

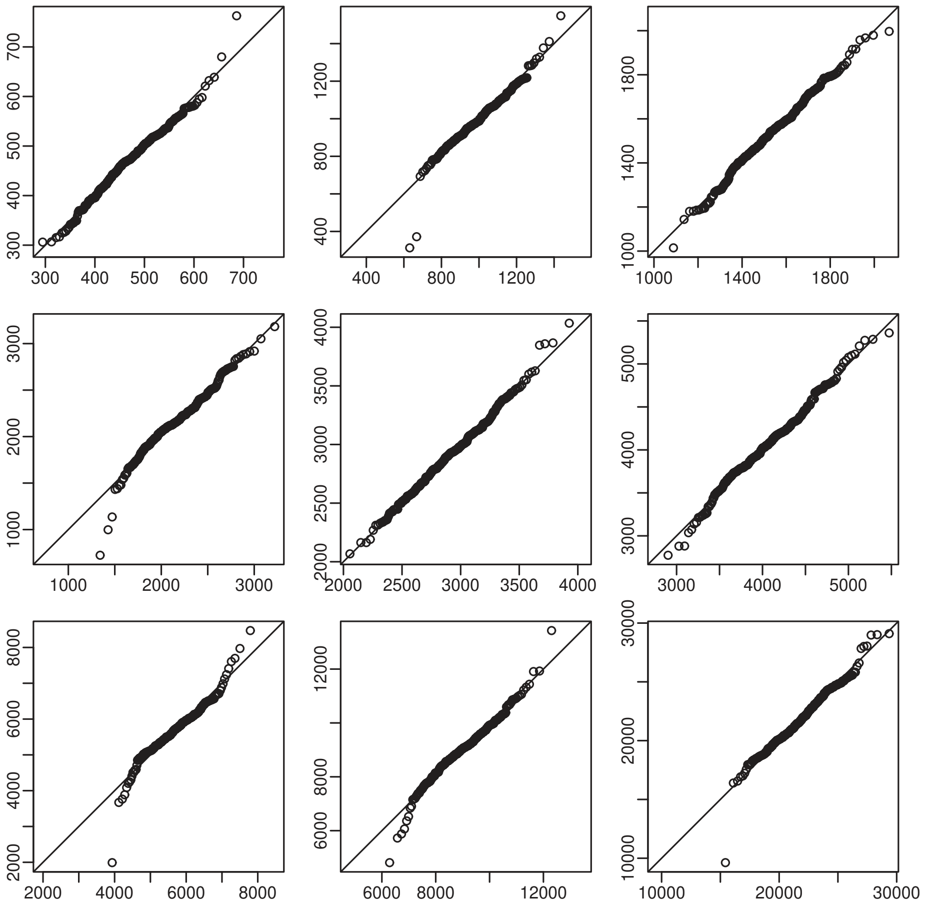

Fig. 3. Gamma QQ plot WT data. Each strip from Figure 2 contains 100 genes and thus 300 expression measures. Plotted here is a QQ plot for a fitted Gamma observation component. Theoretical quantiles are on the x-axis and observed quantiles are on the y-axis.

can make some statements about robustness from the following calculation. We considered six data sets that were derived from the original three-on-three comparison by omitting one microarray at a time. For each of these five-array data sets we applied the methodology to identify genes that are down-regulated in MUT at a target 5% FDR. On average over the six leave-outs, the semiparametric method identified 386 down-regulated genes, 96% of which are, on average, contained in the list obtained from the full data set. Further, the set of genes in both the leave-out and the full list is on average 80% of the full list. This is a high degree of robustness. The parametric methodology, by comparison, makes smaller lists that are mostly contained in the full list but that constitute on average only 61% of the full list. Also, five of six leave-outs produce empty t-based lists.

7. Discussion

Hierarchical mixture models can form the basis of an effective methodology to address the differential expression problem. They provide both for gene ranking and for the creation of short gene lists with target

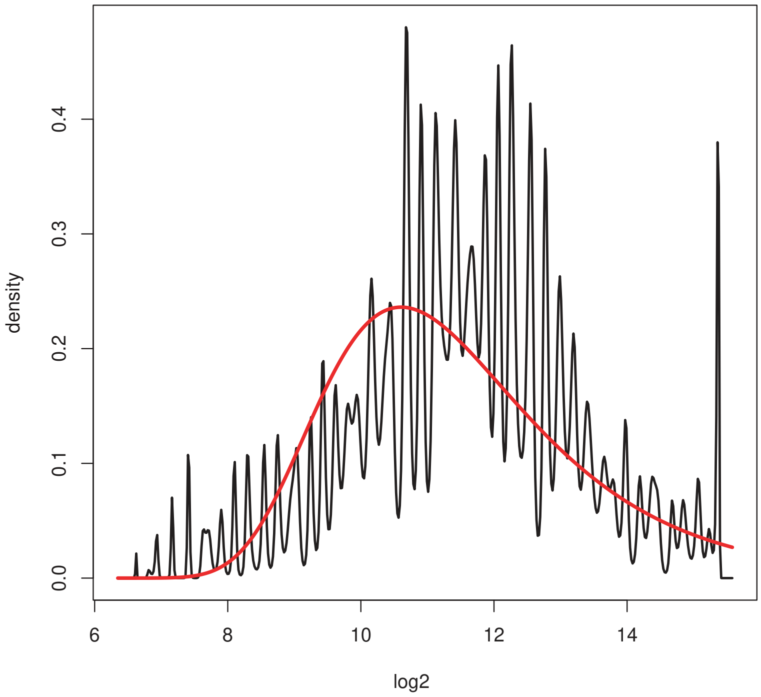

Fig. 4. Nonparametric (black) and parametric (red) estimates of the base distribution π for comparing non-translating MUT and WT RNAs. The semiparametric fit also has likelihood-estimated mixing proportions (0.554, 0.352, 0.093) and moment-estimated observation shape parameters $(a_1, a_2) = (48.1, 85.9)$. In the parametric fit, ML estimates of the mixing proportions are $(0.634, 0.291, 0.074)$ and remaining parameters are $(a_0, a_1, a_2, x_0) = (0.88, 34.2, 73.2, 1556.8)$.

error rates. The hierarchical structure permits a formal connection amongst genes. This is useful because isolated gene-specific calculations may be less efficient than methods which channel information from the whole genome into each gene-specific inference. Further, the mixture structure allows us to estimate global features such as the proportion of up- or down-regulated genes, and, at the same time, guides the formation of short lists with target FDRs. The proposed parametric model captures substantial sources of variation, though goodness of fit is improved when we consider a semiparametric model in which latent mean expression values fluctuate according to an unspecified distribution. Estimation via the EM algorithm is straightforward in either case.

We use a direct posterior probability approach to control FDR (Section 3), and we note that the idea is not new but is evident in the recent and fruitful literature on FDR (Efron *et al.*, 2001; Storey, 2002, 2003; Genovese and Wasserman, 2002a,b). Not accounting for the directionality of our inferences, what we denote by β_g is akin to the local FDR of Efron *et al.* (2001), except that we have fit a full probability model to the data rather than to a one-dimensional reduction; our ranking of genes by β_g and the formation of a gene list with level α FDR would give the same thing as if the Storey (2003) q-value method was applied to the β_g themselves (Storey, 2003, page 21) and if we formed the list of genes for which these q-values are bounded by α. We remark that in the context of a full probability model, as we have estimated, there is a certain simplicity in working directly with the gene-specific probabilities and assessing error rates of lists by looking collectively at these posterior probabilities. Gene-specific posterior probabilities

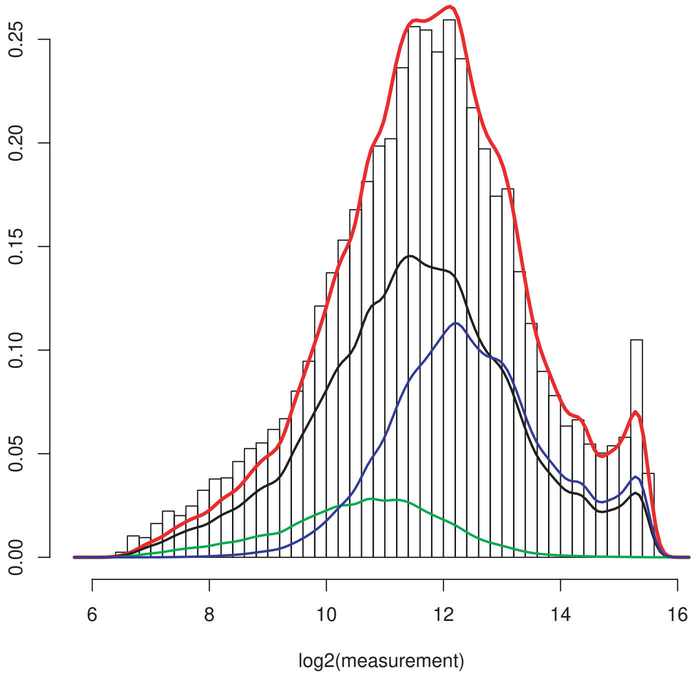

Fig. 5. Marginal diagnostic plot: histogram shows the complete set of MUT expression measurements compared to the marginal distributions (solid lines) which are induced by the fitted semiparametric model. The three interior curves correspond to the three mixing states $H_{g,0}$ $H_{g,1}$ and $H_{g,2}$ and the outer curve is the overall predictive distribution. Parametric fit (not shown) is inferior.

possess an interesting dual functionality: small β_g is the ticket with which gene g gets on the reported list; at the same time, β_g is the chance that the placement of gene g on the list is a false discovery. This fact is helpful in routine manipulations of microarray data. We note further that recent literature focuses primarily on inference using gene-specific p-values. Our modeling effort aims directly at gene-specific posterior probabilities rather than p-values; in the context of a statistical model that produces posterior probabilities, it is natural to use these objects directly, and so this is what we propose.

Though many methods take advantage of features that are typically present in microarray data, few methods rely on the explicit formulation of a probability model for the data. Our use of a Gamma observation component is quite compelling given diagnostic plots (Figures 2 and 3) and the recent characterization that the Gamma distribution is the only distribution for which the sample mean is independent of the sample coefficient of variation (Hwang and Hu, 1999). Flexibility in the mean component is also well justified so as to improve the overall model fit. In contrast to methods which invoke label-permutation to calibrate differential expression, our proposal considers differential expression to be the independence of latent expected expression values. This may be invalid if systematic effects are not properly eliminated by normalization, but it is suitable for carefully pre-processed data and allows for statistical inference even when there is limited replication.

There may be useful extensions of the present model. For example, it deals only with the simple two-group comparison and it ignores dependence among genes. Also, there may be useful improvements to be had on the computational side. Grid-based EM for estimating the base mixing distribution π is effective in the cases considered, though convergence may be slow compared to alternative optimization procedures.

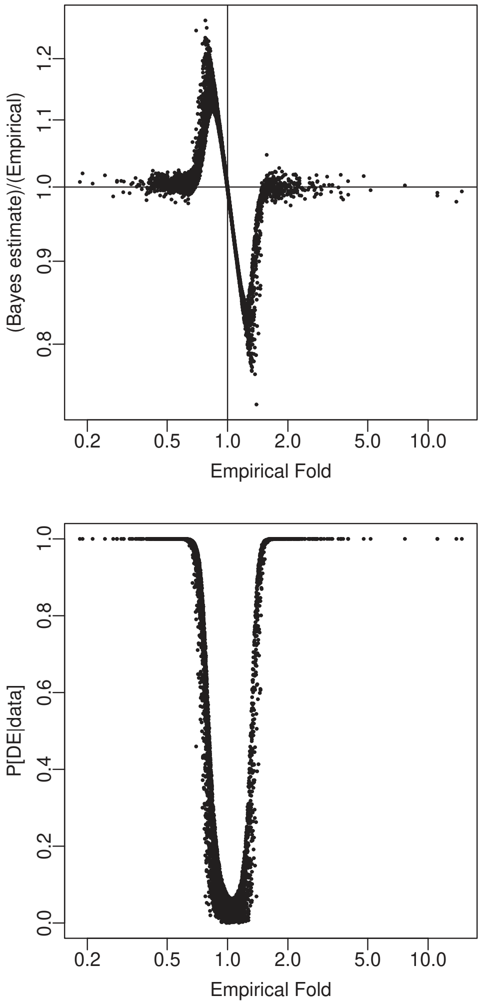

Fig. 6. Fold change estimates (top) and volcano plot (bottom): Ordinary fold change $\bar{x}_g/\bar{y}_g$ (horizontal) compared with semiparametric Bayes estimate $\hat{\rho}_g$ for WT/MUT (top). Upper left and lower right quadrants hold genes which shrink towards $\rho_g = 1$. Genes in the other quadrants (6%) do not shrink towards 1. The lower panel indicates how posterior probability tracks with empirical fold change.

Acknowledgements

The authors thank C. Kendziorski for useful comments and G. Rempala for pointing us to the Hwang and Hu paper. Remarks by an associate editor and anonymous referees were extremely helpful in revising an earlier draft. This research was supported by National Institutes of Health grants GM35072 and CA97944 to Paul Ahlquist and CA64364 to Michael Newton. Paul Ahlquist is an Investigator of the Howard Hughes Medical Institute. See `http://www.stat.wisc.edu/~newton/` for **R** code and further information.

APPENDIX

A.1 *Parametric submodel*

Here we describe the various marginal and conditional distributions induced by the parametric hierarchical mixture model. These objects are used both to drive the model fitting calculations (Section A.2) and to characterize the final gene-specific inferences. To simplify the development, we suppress in the notation the gene subscript g.

Gamma, inverse Gamma. The random variable V has a Gamma distribution with shape a and rate b, denoted $V \sim \text{Gamma}(a, b)$ if, for $v > 0$,

$$p(v) = b^a v^{a-1} \exp(-bv)/\Gamma(a). \tag{A.1}$$

Note that $E(V) = a/b$ in this parametrization and the coefficient of variation is $1/\sqrt{a}$. A variable U has an inverse Gamma distribution with shape a and rate b if $1/U \sim \text{Gamma}(a, b)$.

Compound Gamma. Let $\theta \sim \text{Gamma}(a_0, a_0 v_0)$ for hyperparameters a_0 and v_0. Given θ, let $V_1, V_2, \ldots, V_k$ be conditionally independent Gamma-distributed random variables in which V_i has shape a_i and rate $a_i\theta$. Thus, $1/\theta$ is the common conditional mean of the V_i. It is a classical result that upon integrating θ, the random variables $V_1, \ldots, V_k$ have a compound Gamma distribution with joint density

$$h(v_1, \ldots, v_k) = \frac{v_0\, \Gamma\left(\sum_{i=0}^k a_i\right)}{\left(\sum_{i=0}^k a_i v_i\right)^{\sum_{i=0}^k a_i}} \prod_{i=0}^k \left[\frac{a_i^{a_i} v_i^{a_i-1}}{\Gamma(a_i)}\right]. \tag{A.2}$$

This distribution arises as the predictive distribution of measurements on a gene when considered marginally with respect to the latent mean.

Mean component. Let π which defines f in (2.1) be the density function of an inverse Gamma distribution having shape a_0 and rate $a_0 x_0$. Thus

$$\begin{aligned} f_0(\mu_1, \mu_2) &= \pi(\mu_1)\, 1[\mu_1 = \mu_2] \\ f_1(\mu_1, \mu_2) &= 2\pi(\mu_1)\pi(\mu_2)\, 1[\mu_1 < \mu_2] \\ f_2(\mu_1, \mu_2) &= 2\pi(\mu_1)\pi(\mu_2)\, 1[\mu_1 > \mu_2]. \end{aligned} \tag{A.3}$$

Fitting the model amounts to estimating the discrete mixing proportions (p_0, p_1, p_2), the shape parameters (a_0, a_1, a_2) and the location parameter x_0. We use an EM algorithm and maximize marginal likelihood using data from all genes together under the independence assumption (Section A.2). Both this calculation and the calculation of posterior probabilities requires a formula for the marginal probability of gene-specific data $(\mathbf{x}, \mathbf{y})$ having integrated against $f(\mu_1, \mu_2)$, as in (3.1). We obtain

$$p(\mathbf{x}, \mathbf{y}) = p_0 h(\mathbf{x}, \mathbf{y}) + p_1 h(\mathbf{x}) h(\mathbf{y})\, 2P[B > b] + p_2 h(\mathbf{x}) h(\mathbf{y})\, 2P[B < b] \tag{A.4}$$

where h is the density of a compound Gamma (A.2), B is a Beta-distributed random variable with shapes $(a_0 + ma_1, a_0 + na_2)$, and

$$b = b(\mathbf{x}, \mathbf{y}) = \frac{a_0 x_0 + a_1 \sum_{i=1}^m x_i}{2a_0 x_0 + a_1 \sum_{i=1}^m x_i + a_2 \sum_{i=1}^n y_i}.$$

More specifically, the shape parameters entering h in (A.4) are a_1 for all m coordinates in $h(\mathbf{x})$, a_2 for all n coordinates of $h(\mathbf{y})$, and a concatenation of m a_1 and n a_2 in $h(\mathbf{x}, \mathbf{y})$.

We find it interesting how the compensating factor involving B enters the formula (A.4).

Gene-specific inference uses the posterior probability distribution over the three hypotheses about μ_1 and μ_2. For example, using elements in (A.4), the posterior probability of equivalent expression is

$$P(\mu_1 = \mu_2|\mathbf{x}, \mathbf{y}) = P(Z = 0|\mathbf{x}, \mathbf{y}) = p_0 h(\mathbf{x}, \mathbf{y})/p(\mathbf{x}, \mathbf{y}) \tag{A.5}$$

and the posterior probability of over-expression in the $\mathbf{x}$ condition is

$$P(\mu_1 > \mu_2|\mathbf{x}, \mathbf{y}) = P(Z = 2|\mathbf{x}, \mathbf{y}) = p_2 h(\mathbf{x}) h(\mathbf{y}) 2P[B < b]/p(\mathbf{x}, \mathbf{y}). \tag{A.6}$$

Derivation of (A.4). That (A.4) involves a discrete mixture of three terms follows from (2.1) and the structure of f in (A.3). The first term $p_0 h(\mathbf{x}, \mathbf{y})$ emerges from the definition of compound Gamma (A.2) because on the null all measurements have a common conditional mean. The next two terms involve the ordered alternatives. Consider the term $p_1(\mathbf{x}, \mathbf{y}) = h(\mathbf{x})h(\mathbf{y})2P(B > b)$ which corresponds to hypothesis H_1, under-expression in the first state. When $m = n = 1$ (i.e. no replication), the calculations are most simple to report. The argument is not much more difficult in the general case since $\sum x_i$, $\prod x_i$, $\sum y_i$ and $\prod y_i$ are sufficient statistics. Here we give the argument when $m = n = 1$, and we further suppose $a_1 = a_2 = a$, though this also is not necessary.

Let $\theta_1 = 1/\mu_1$ and $\theta_2 = 1/\mu_2$ denote the inverted expectations. By (A.3), their joint PDF is

$$2p(\theta_1)p(\theta_2)1[\theta_1 > \theta_2]$$

with common Gamma densities $p(\theta_1)$ and $p(\theta_2)$, as in (A.1) with shape a_0 and rate $a_0 x_0$. Thus, the marginal predictive density

$$\begin{aligned} p_1(x, y) &= \int_0^\infty \int_0^\infty p(x|\theta_1)p(y|\theta_2)\, 2p(\theta_1)p(\theta_2)1[\theta_1 > \theta_2]\, \mathrm{d}\theta_1 \mathrm{d}\theta_2 \\ &= \int_0^\infty 2p(y|\theta_2)p(\theta_2)I(\theta_2)\, \mathrm{d}\theta_2 \end{aligned} \tag{A.7}$$

where

$$\begin{aligned} I(\theta_2) &= \int_{\theta_2}^\infty p(x|\theta_1)p(\theta_1)\, \mathrm{d}\theta_1 \\ &= \int_{\theta_2}^\infty \frac{(a_0x_0)^{a_0}\theta_1^{a_0-1}\exp\{-\theta_1 a_0 x_0\}}{\Gamma(a_0)} \frac{(a\theta_1)^a x^{a-1}\exp\{-a\theta_1 x\}}{\Gamma(a)}\, \mathrm{d}\theta_1 \\ &= h(x) \int_{\theta_2}^\infty \frac{(a_0x_0 + ax)^{a_0+a}\theta_1^{a_0+a-1}\exp\{-\theta_1(a_0x_0 + ax)\}}{\Gamma(a_0 + a)}\, \mathrm{d}\theta_1 \\ &= h(x) \int_{\theta_2(a_0x_0+ax)}^\infty \frac{\psi_1^{a_0+a-1}\exp(-\psi_1)}{\Gamma(a_0 + a)}\, \mathrm{d}\psi_1. \end{aligned}$$

Now we plug this back into (A.7), and switch the order of integration to draw out $h(y)$ as we have just drawn out $h(x)$. Specifically,

$$\begin{aligned}
p_1(x, y) &= 2h(x) \int_0^\infty \int_{\theta_2(a_0x_0+ax)}^\infty p(y|\theta_2)p(\theta_2)\frac{\psi_1^{a_0+a-1}\exp(-\psi_1)}{\Gamma(a_0+a)}\,\mathrm{d}\theta_2\mathrm{d}\psi_1\\
&= 2h(x)h(y)\int_0^\infty \int_0^{\psi_1/(a_0x_0+ax)} p(\psi_1)\,\frac{p(y|\theta_2)p(\theta_2)}{h(y)}\,\mathrm{d}\theta_2\mathrm{d}\psi_1\\
&= 2h(x)h(y)\int_0^\infty p(\psi_1)\left(\int_0^{\psi_1/(a_0x_0+ax)} p(\theta_2|y)\,\mathrm{d}\theta_2\right)\mathrm{d}\psi_1
\end{aligned}$$

where $p(\psi_1)$ is a Gamma PDF with rate 1 and shape $a_0 + a$, and $p(\theta_2|y)$ is the posterior calculated under the Gamma prior for θ_2, which also turns out to be a Gamma with shape $a + a_0$ by conjugacy. Adjusting rates by a change of variables, we have

$$\begin{aligned}
p_1(x, y) &= 2h(x)h(y)\int_0^\infty \int_0^r p(\psi_1)p(\psi_2)\,\mathrm{d}\psi_2\mathrm{d}\psi_1\\
&= 2h(x)h(y)\,P\,(\psi_1 > \psi_2/r)\\
&= 2h(x)h(y)\,P\,[B > (1-B)/r]\\
&= 2h(x)h(y)\,P\,[B > 1/(1+r)]
\end{aligned}$$

where now ψ_1 and ψ_2 are i.i.d. Gamma variables with shape $a_0 + a$, and $r = (ay + a_0x_0)/(ax + a_0x_0)$, and $B = \psi_1/(\psi_1 + \psi_2)$ is a Beta distributed random variable. Thus we obtain the second term in (A.4) with $b = 1/(1 + r)$. The third term is similarly derived.

A.2 *Model fitting details*

Semiparametric model. Having replicate measurements within each gene/condition setting allows us to get a method-of-moments estimate of the observation component shape parameters a_1 and a_2 as indicated in Section 2. These are treated as fixed in the following.

The model asserts that i.i.d. draws U_g and V_g arise from an unknown base distribution π for all genes g, and also that there are i.i.d. discrete random variables Z_g on $\{0, 1, 2\}$ with unknown mixing probabilities (p_0, p_1, p_2) which indicate patterns of differential expression. The base distribution π is considered to be a probability vector on a finite grid $\mathcal{G}$ in the space of log mean expressions. In our calculations we used a grid of 500 values equally spaced in the range of the log expression measurements, though the software allows a user-supplied grid. The EM algorithm starts at a uniform π and equal mixing probabilities $(1/3, 1/3, 1/3)$. To generate the algorithm, note that the complete data likelihood L_c is

$$\begin{aligned}
L_c &= \prod_g p(\mathbf{x}_g, \mathbf{y}_g, Z_g, U_g, V_g)\\
&= \prod_g p(\mathbf{x}_g|Z_g, U_g, V_g)\; p(\mathbf{y}_g|Z_g, U_g, V_g)\; p(Z_g)\,\pi(U_g)\,\pi(V_g)\\
&= A \times B \times C
\end{aligned}$$

where, in the last line, the product is being reorganized as a product over the three values of Z_g. For example,

$$A = \prod_{g:Z_g=0} p(\mathbf{x}_g|U_g)\; p(\mathbf{y}_g|U_g)\; P(Z_g = 0)\,\pi(U_g)\,\pi(V_g),$$

and

$$B = \prod_{g:Z_g=1} p(\mathbf{x}_g|U_g \wedge V_g)\, p(\mathbf{y}_g|U_g \vee V_g)\, P(Z_g = 1)\, \pi(U_g)\, \pi(V_g).$$

Next, expand $\pi(U_g) = \prod_{u\in\mathcal{G}} [\pi(u)]^{1(U_g=u)}$ and similarly expand $\pi(V_g)$ and use this to simplify the complete-data log-likelihood as

$$\log L_c = \sum_{j=0}^{2} \log(p_j) \left(\sum_g 1[Z_g = j] \right) + \sum_{u\in\mathcal{G}} \log[\pi(u)] \left\{ \left(\sum_g 1[U_g = u] \right) \right. \\ \left. + \left(\sum_g 1[V_g = u] \right) \right\} + K$$

where K involves the observation components but neither π nor (p_0, p_1, p_2). The E-step is complete upon taking conditional expectations; the M-step is

$$\hat{p}_j = \frac{\sum_g P(Z_g = j|\mathbf{x}_g, \mathbf{y}_g)}{N} \tag{A.8}$$
$$\hat{\pi}(u) = \frac{\sum_g \left[P(U_g = u|\mathbf{x}_g, \mathbf{y}_g) + P(V_g = u|\mathbf{x}_g, \mathbf{y}_g) \right]}{2N}$$

where N is the number of genes. It remains to derive a useful formula for these conditional probabilities. Of course by Bayes' rule (3.1),

$$P(Z_g = j|\mathbf{x}_g, \mathbf{y}_g) \propto p_j\, p(\mathbf{x}_g, \mathbf{y}_j|Z_g = j)$$

and this can be renormalized after evaluating it for all three hypotheses j. Each j involves integrating over the mean space. For example,

$$p(\mathbf{x}_g, \mathbf{y}_g|Z_g = 1) = \sum_{u\in\mathcal{G}} \sum_{v\in\mathcal{G}} p(\mathbf{x}_g|u \wedge v)\, p(\mathbf{y}_g|u \vee v)\, \pi(u)\, \pi(v). \tag{A.9}$$

Conveniently, the double summation in (A.9) reduces and may be computed using simple vector inner products. To see this, consider univariate distributions associated with sampling from π, such as

$$\pi(\mathbf{x}_g) = \sum_{u\in\mathcal{G}} p(\mathbf{x}_g|u)\, \pi(u), \qquad \pi(u|\mathbf{x_g}) = p(\mathbf{x}_g|u)\, \pi(u)/\pi(\mathbf{x}_g)$$

and the tail posterior $\pi(U_g > v|\mathbf{x}_g) = \sum_{u>v} \pi(u|\mathbf{x}_g)$. Likelihoods $p(\mathbf{x}_g|u)$ and $p(\mathbf{y}_g|u)$ and the base distribution π are stored on the grid $\mathcal{G}$ so these objects are readily computed. With this, the predictive probability (A.9) can be written as

$$\begin{aligned} p(\mathbf{x}_g, \mathbf{y}_g|Z_g = 1) &= \epsilon + \sum_{u\neq v} p(\mathbf{x}_g|u \wedge v)\, p(\mathbf{y}_g|u \vee v)\, \pi(u)\, \pi(v) \\ &= \epsilon + 2 \sum_{u<v} p(\mathbf{x}_g|u)\, p(\mathbf{y}_g|v)\, \pi(u)\, \pi(v) \\ &= \epsilon + 2\pi(\mathbf{x}_g)\, \pi(\mathbf{y}_g) \sum_{u<v} \pi(u|\mathbf{x}_g)\, \pi(v|\mathbf{y}_g) \\ &= \epsilon + 2\pi(\mathbf{x}_g)\, \pi(\mathbf{y}_g) \sum_{u\in\mathcal{G}} \pi(u|\mathbf{x}_g)\, \pi(V_g > u|\mathbf{y}_g). \end{aligned}$$

Here ϵ, which is on the order of $\sum_{u \in \mathcal{G}} \pi^2(u)$, may be ignored as it emerges from the possibility on $H_{g,1}$ that independent draws from π can be equal; ϵ is reduced as the grid $\mathcal{G}$ becomes more dense.

Continuing with the E-step, we require a simplified formula in (A.8) for

$$\psi_g(u) = P(U_g = u|\mathbf{x}_g, \mathbf{y}_g) + P(V_g = u|\mathbf{x}_g, \mathbf{y}_g)$$

for each $u \in \mathcal{G}$. We find by similar manipulations that $\psi_g(u)$ is proportional to

$$\begin{aligned} p_0 \pi(\mathbf{x}_g, \mathbf{y}_g) \left[\pi(u|\mathbf{x}_g, \mathbf{y}_g) + \pi(u)\right] & \\ + 2p_1\, \pi(\mathbf{x}_g)\, \pi(\mathbf{y}_g) & \left[\pi(u|\mathbf{y}_g)\, \pi(U_g \leqslant u|\mathbf{x}_g) + \pi(u|\mathbf{x}_g)\, \pi(V_g > u|\mathbf{y}_g)\right] \\ + 2p_2\, \pi(\mathbf{x}_g)\, \pi(\mathbf{y}_g) & \left[\pi(u|\mathbf{x}_g)\, \pi(V_g \leqslant u|\mathbf{y}_g) + \pi(u|\mathbf{y}_g)\, \pi(U_g > u|\mathbf{x}_g)\right]. \end{aligned}$$

The software provided starts with uniform estimate of π and (p_0, p_1, p_2) and iterates EM steps by computing the necessary one-dimensional probability vectors. In the reported calculations we used a grid $\mathcal{G}$ with 500 support points in the range of the log data, and we used 300 EM iterations. Convergence was monitored informally by graphical analysis of the model fits and posterior probabilities.

Using the formulae in Section A.1 it is straightforward to generate an EM algorithm for the fully parametric model in which only the discrete $\{Z_g\}$ are missing values. Moment-based estimates are used to start the EM iterations.

A.3 *Miscellaneous*

Spike-in. Arrays used in the spike-in comparison from Table 1 of Antonellis *et al.* (2002) are 92456hgu95a11, 92457hgu95a11, 92499hgu95a11, 92458hgu95a11, 92459hgu95a11, and 92500hgu95a11.

ETST method. The ETST method uses a regularized t-like statistic on the log-transformed measurements: in our notation

$$T_g = \frac{\frac{1}{m}\sum_i \log(x_{g,i}) - \frac{1}{n}\sum_i \log(y_{g,i})}{S_g + a_0} \tag{A.10}$$

where S_g is a gene-specific pooled sample standard deviation and a_0 is a regularizing constant. ETST note that having a_0 converge to ∞ amounts to assuming constant variance among genes on the log scale; to develop a fair comparison we adopt this limiting case. We implement this by replacing the denominator of (A.10) by the median of S_g values across genes. (Recall that the statistics T_g are transformed to Z-score values via a normal-scores transformation so that constant-ness across genes rather than large-ness is the key property of the denominator.) The ETST method involves a two-component mixture model on the transformed Z-scores. For the unpaired comparison problem, ETST propose that null scores be generated by permutation of sample labels, and then a nonparametric regression method is used to estimate the posterior probability of differential expression. With $m = n = 3$, there are only ten distinct arrangements of the six microarrays into two equal-sized groups, and so we evaluated null Z-scores exhaustively rather than by random shuffling. We used ETST equation (5.9) which provides a way to estimate the FDR of any given rejection region.

REFERENCES

ALLISON, D. B., GADBURY, G. L., HEO, M., FERNANDEZ, J. R., LEE, C. -K., PROLLA, T. A. AND WEINDRUCH, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **35**, 1–20.

ANTONELLIS, K. J., BEAZER-BARCLAY, Y. D., ELASHOFF, M., JELINSKY, S. A., WHITLEY, M. Z., BROWN, E.L. AND SCHERF, U. (2002). Optimization of an external standard for the normalization of Affymetrix GeneChip arrays. *Gene Logic Technical Note*.

BAGGERLY, K. A., COOMBES, K. R., HESS, K. R., STIVERS, D. N., ABRUZZO, L. V. AND ZHANG, W. (2001). Identifying differentially expressed genes in cDNA microarray experiments. *Journal of Computational Biology* **8**, 639–659.

BALDI, P. AND LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer.

BROËT, P., RICHARDSON, S. AND RADVANYI, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* **9**, 671–683.

CHEN, Y., DOUGHERTY, E. R. AND BITTNER, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

DUDOIT, S., YANG, Y. H., SPEED, T. P. AND CALLOW, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

DUDOIT, S., SHAFFER, J. P. AND BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. AND TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

GENOVESE, C. AND WASSERMAN, L. (2002a). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* **64**, 499–518.

GENOVESE, C. AND WASSERMAN, L. (2002b). A large sample approach to false discovery control. *Technical Report*. Carnegie Mellon: Department of Statistics.

HWANG, T -Y. AND HU, C -Y. (1999). On a characterization of the gamma distribution: the independence of the sample mean and the sample coefficient of variation. *Annals of the Institute of Statistical Mathematics* **51**, 749–753.

IDEKER, T., THORSSON, V., SIEGEL, A. F. AND HOOD, L. E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* **7**, 805–817.

IHAKA, R. AND GENTLEMAN, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314. (See `www.r-project.org`)

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. AND SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

JOHANNES, G., CARTER, M. S., EISEN, M. B., BROWN, P. O. AND SARNOW, P. (1999). Identification of eukaryotic mRNAs taht are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proceedings of the National Academy of Sciences* **96**, 13118–13123.

KENDZIORSKI, C. M., NEWTON, M. A., LAN, H. AND GOULD, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

KERR, M. K., MARTIN, M. AND CHURCHILL, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

LEE, M. L. T., LU, W., WHITMORE, G. A. AND BEIER, D. (2002). Models for microarray gene expression data. *Journal of Biopharmaceutical Statistics* **12**, 1–19.

LI, C. AND WONG, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* **98**, 31–36.

LINDSAY, B. (1995). *Mixture models: theory, geometry, and applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5. Hayward, CA: Institute for Mathematical Statistics.

LONNSTEDT, I. AND SPEED, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

NGUYEN, D. V., ARPAT, A. B., WANG, N. AND CARROLL, R. J. (2002). DNA microarray experiments: biological and technical aspects. *Biometrics* **58**, 701–717.

NOUEIRY, A. O., CHEN, J. AND AHLQUIST, P. (2000). A mutant allele of essential, general translation initiation factor DED1 selectively inhibits translation of a viral mRNA. *Proceedings of the National Acadamy of Sciences* **97**, 12985–90.

PAN, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546–554.

PARMIGIANI, G., GARRETT, E. S., IRIZARRAY, R. A. AND ZEGER, S. L. (2003). *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer.

ROCKE, D. M AND DURBIN, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology* **8**, 557–570.

SPEED, T. P. (2002). `http://www.stat.Berkeley.EDU/users/terry/zarray/Html/list.html`.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

STOREY, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.

STOREY, J. D. AND TIBSHIRANI, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445.

THEILHABER, J., BUSHNELL, S., JACKSON, A. AND FUCHS, R. (2001). Bayesian estimation of fold-changes in the analysis of gene expression: the PFOLD algorithm. *Journal of Computational Biology* **8**, 585–614.

TSODIKOV, A., SZABO, A. AND JONES, D. (2002). Adjustments and tests for differential expression with microarray data. *Bioinformatics* **18**, 251–260.

TUSHER, V., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Acadamy of Sciences* **98**, 5116–5121.

[*Received January 30, 2003; revised December 8, 2003; accepted for publication December 17, 2003*]