Biostatistics (2005), **6**, 2, *pp.* 279–291
doi:10.1093/biostatistics/kxi009

Hierarchical Bayes models for cDNA microarray gene expression

INGRID LÖNNSTEDT*

Department of Mathematics, Uppsala University, Box 480, Uppsala SE-751 06, Sweden
ingrid@math.uu.se

TOM BRITTON

Department of Mathematics, Stockholm University, Stockholm SE-106 91, Sweden

SUMMARY

cDNA microarrays are used in many contexts to compare mRNA levels between samples of cells. Microarray experiments typically give us expression measurements on 1000–20 000 genes, but with few replicates for each gene. Traditional methods using means and standard deviations to detect differential expression are not satisfactory in this context. A handful of alternative statistics have been developed, including several empirical Bayes methods. In the present paper we present two full hierarchical Bayes models for detecting gene expression, of which one (D) describes our microarray data very well. We also compare the full Bayes and empirical Bayes approaches with respect to model assumptions, false discovery rates and computer running time. The proposed models are compared to existing empirical Bayes models in a simulation study and for a set of data (Yuen *et al.*, 2002), where 27 genes have been categorized by quantitative real-time PCR. It turns out that the existing empirical Bayes methods have at least as good performance as the full Bayes ones.

Keywords: cDNA microarray; Differential expression; Empirical Bayes; Hierarchical Bayes; MCMC simulations; ROC curve.

1. INTRODUCTION

1.1 *cDNA microarray data*

cDNA microarrays are used to compare gene expression in different samples of cells. They enable us to study the expression levels of thousands of genes simultaneously. The technique has a wide range of applications including learning how genes interact, which genes are used in different cell types and which genes change their expression in cells due to disease or drug stimuli.

For two cell samples to be compared, cDNA labeled with fluorescent dye (usually red and green) is hybridized to the cDNA spots on a glass slide (*microarray*). Each spot represents one out of several thousand genes to be examined. If a gene has a higher level of expression in the 'red' sample than in the 'green,' there will be more red than green dye on this spot. The dye intensities (R and G) are detected

*To whom correspondence should be addressed.

using a laser scanner. After a process of image analysis the information can be summarized as genewise (spotwise) log-ratios (M) of the dye intensities [discussed in Dudoit *et al.* (2002)].

Throughout this paper we assume a replicated microarray experiment so that our data consist of an $N \times n$ matrix $M = (M_{ij})$. The number of genes N represented on each microarray is large, say 10 000. The number of replicates n (microarrays) is usually small (2–10). For gene i on microarray j, $M_{ij} = \log_2 \frac{R_{ij}}{G_{ij}}$. We further assume that our log-ratios are properly normalized for systematic errors, e.g. according to Yang *et al.* (2002) or Kerr and Churchill (2001).

Our task is to determine which genes have different expression levels in the two samples, i.e. which genes have M-values genuinely different from zero.

1.2 *Statistics for differential expression*

Common statistics which might be used to select the differentially expressed genes are the mean and standardized mean expression levels, $(M_{i\cdot})$ and $(t_i) = (M_{i\cdot}/(s_i/\sqrt{n}))$, s_i being the sample standard deviation of the values M_{ij}, $j = 1, \ldots, n$. There are problems with both of these traditional statistics. A large mean might be driven by an outlier, something which occurs quite frequently in this context. A large t might arise because of a small denominator $s_i/\sqrt{n}$, even though the mean itself is small. Because of the small number of replicates, there will usually be genes with very small standard errors, and some of these will have small means as well.

Several alternative statistics have been proposed to overcome these problems. A sophisticated statistic SAM (Tusher *et al.*, 2001) slightly moderates the t-statistic, adding a suitable constant in the denominator. SAM can be thought of as a Bayesian t-statistic, where the denominator is a sum of a global prior standard error and the gene-specific $s_i/\sqrt{n}$.

In Lönnstedt and Speed (2002) an empirical Bayes log odds-ratio B is presented, which is summarized below including some new parameter estimates relevant to this paper. B is reparameterized into LOD in Smyth (2004), where a complete, stable parameter estimation procedure is suggested. Smyth (2004) also notes that LOD (or B) is monotonic increasing in a tuned t-statistic similar to SAM.

In a slightly different empirical Bayes framework, Baldi and Long (2001) model each channel ($\log_2 R$ or $\log_2 G$) separately, using Normal distributions with conjugate priors (normal and inverse gamma). This results in two posterior models for each gene and the posterior probablilty that the two location parameters are equal [$P(\log_2 R = \log_2 G|\text{Data})$].

Additional empirical Bayes models exist, e.g. Long *et al.* (2001) and Gottardo *et al.* (2003). Empirical Bayes approaches to inference seem natural, since microarrays hold information about thousands of genes simultaneously. This overall information is summarized into prior parameters, to be combined with means and standard deviations at the gene level. They do, however, make distributional assumptions in order to produce conjugate priors for the parameters. In Section 2 we derive a method of moments estimate of the proportion p of differentially expressed genes in a microarray experiment for the B model. The poor performance of this estimate motivates leaving the empirical Bayes model for full Bayesian treatment developed in Section 3, to choose parameters and distributions that better fit our data.

Broët *et al.* (2002) present a full Bayesian hierarchical model which allows the genes to be assigned to one of an unknown number of expression levels. The genes assigned to a given level are modeled to have the same mean and variance. Our models are mixtures of regulated and unregulated genes, where the former (as well as the latter in the second model) are allowed gene-specific means and variances. This simplifies the numerical evaluation and maintains a low number of hyperparameters. Model C is slightly more realistic than the empirical Bayes method B, whilst model D is constructed to represent our data very well. We investigate the gains of using more realistic models than empirical Bayes versus the cost of the more complicated numerical evaluations. A full Bayesian hierarchical model is also

used in Parmigiani *et al.* (2002) for a different purpose of classifying tumors from unreplicated gene expression data.

In Section 4 we compare our proposed full Bayesian methods with some of the empirical Bayes models mentioned above. We conclude that the empirical Bayes methods perform at least as well as our full Bayes methods on our microarray data. They are also computationally far cheaper.

1.3 *Data sets*

In this paper we use three microarray data sets, which we call SRBI, SRBI$^-$ and the Yuen data sets.

The SRBI experiment compares gene expression in liver cells from scavenger receptor transgenic (SR-BI) mice to those of the corresponding wild-type control mice, see Callow *et al.* (2000). Eight SR-BI mice are all compared to a reference sample of pooled cDNA from the control mice, resulting in eight microarray data sets. After image processing and normalization as described in Buckley (2000) and Yang *et al.* (2002), our data consisted of the eight sets of log-ratios (M-values). There were 6356 genes on each microarray.

The SRBI$^-$ data set comes from the same experiment as SRBI. Here, the eight control mice are compared to the reference sample of their pooled cDNA, so that averaging the eight sets of log-ratios, we expect no genes to have changed. This data set is used for the simulations in Section 4.

The Yuen data set comes from the microarray experiment in Yuen *et al.* (2002). It contains three replicated microarray slides, on each of which 956 genes are spotted in triplicates. For 47 of the genes, relative expression measurements have been established using quantitative real-time PCR assays. Of these 47 genes, 27 were categorized: 17 as differentially expressed and 10 as not differentially expressed.

2. The proportion of differentially expressed genes

2.1 *Empirical Bayes method B*

In Lönnstedt and Speed (2002) a statistic B was presented for the selection of differentially expressed genes in replicated microarray experiments. B is a log posterior odds of differential expression, based on the log-ratios M_{ij}, $i = 1, \ldots, N$ and $j = 1, \ldots, n$ for N genes each represented on n replicate microarray slides. The M_{ij} are regarded as Normally distributed random variables,

$$M_{ij}|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2). \tag{2.1}$$

Given the parameters all genes and replicates are assumed independent. This is of course not true but provides a pragmatic basis for modeling the data. Most genes have $\mu_i = 0$, but a small proportion p of genes have $\mu_i \neq 0$, indicated by $I_i = 1$ as opposed to $I_i = 0$. The log posterior odds for gene g to be differentially expressed is calculated as $B_g = \log(\Pr(I_g = 1|M)/\Pr(I_g = 0|M))$. By assuming conjugate prior distributions for the variances of the genes [inverse gamma, $n\beta/2\sigma_i^2 \sim \Gamma(\alpha, 1)$], as well as for their means where not zero [$\mu_i|\sigma_i^2, I_i \neq 0 \sim N(0, c\sigma_i^2)$], we obtain an explicit formula for B,

$$B_g = \log\left(\frac{p}{1-p} \frac{1}{\sqrt{1+nc}} \left[\frac{\beta + s_g^2 + M_{g.}^2}{\beta + s_g^2 + \frac{M_{g.}^2}{1+nc}} \right]^{\alpha+\frac{n}{2}} \right). \tag{2.2}$$

The use of B provides a ranking among the genes based on high relative expression level and replicability. It rules out the groups of genes that are easily falsely selected if only the average expression or a t-statistic was used (Section 1.2).

2.2 *Estimating the proportion p of differentially expressed genes*

The hyperparameter p (the proportion of differentially expressed genes) is usually not estimated (or not successfully estimated) in empirical Bayes models. Also, the constant c in the prior distribution for the non-zero means was not estimated in Lönnstedt and Speed (2002). A proper estimation procedure for these parameters would give a meaning to the numerical level of B. It would allow for resampling-based multiple testing in sufficiently large data sets, giving traditional p-values, which are commonly asked for by microarray users. B would hold the statistical evidence for each gene, rather than just being a ranking over all genes. Therefore, the estimation of p (and c) is a main motivation for the models derived in this paper. In this section we derive estimates $\hat{p}_B$ and $\hat{c}_B$ for the B model in Section 2.1 using the method of moments.

First, we observe from the distributional assumptions in Section 2.1 that for the genewise average M-values ($M_{i\cdot} = \sum_j M_{ij}/n$), we have $M_{i\cdot}|\sigma_i^2 \sim N(0, \sigma_i^2/n)$ if $I_i = 0$ and $N(0, \sigma_i^2(\frac{1+nc}{n}))$ if $I_i = 1$. Now, the variances σ_i^2 are unknown, so we use the sample variances s_i^2 and t-distributions. Letting $Y_i = M_{i\cdot}/s_i$ our key equations will be $Y_i\sqrt{n} \sim t_{n-1}$ if $I_i = 0$ and $Y_i\sqrt{n/(1+nc)} \sim t_{n-1}$ if $I_i = 1$. We will use the sample moments of Y_i to estimate $p = P(I_i = 1)$ and c. These formulae are now identical for all genes i, so we omit the index i for the rest of this section. Y can be written as a product of independent random variables $Y = ZT$, where $T \sim t_{n-1}$ and $Z = \sqrt{1/n}$ with probability $1-p$ and $\sqrt{(1+nc)/n}$ with probability p . By independence, $E(Y^r) = E(Z^r)E(T^r)$ for the rth moment of Y. Odd moments of T are zero by the symmetry of the t-distribution, so we use the two equations generated by the first absolute and second moments: $E|Y| = E|Z|E|T|$ and $E(Y^2) = E(Z^2)E(T^2)$. By substituting expressions for these moments into the Y_i equations, we get

$$\begin{cases} pc = k_1 \\ p(\sqrt{1+nc} - 1) = k_2 \end{cases}, \quad \text{where} \begin{cases} k_1 = \frac{n-3}{n-1}E(Y^2) - \frac{1}{n} \\ k_2 = \frac{(n-2)\sqrt{\pi}\Gamma(\frac{n-1}{2})\sqrt{n}}{2\sqrt{n-1}\Gamma(\frac{n}{2})}E|Y| - 1 \end{cases},$$

which has solutions $\hat{p}_B = \frac{k_2^2}{nk_1 - 2k_2}$ and $\hat{c}_B = \frac{k_1}{k_2}(n\frac{k_1}{k_2} - 2)$ provided $n \geqslant 4$. The parameters p and c are hence estimated using the method of moments by replacing $E|Y|$ and $E(Y^2)$ by the corresponding sample moments.

2.3 *Evaluation of the estimate of p*

When evaluating the estimates $\hat{p}_B$ and $\hat{c}_B$ above for data simulated from the model presented in Section 2.1, the estimates show good concordance with the true values used in the simulations, although the sampling distribution for c is very asymmetric. From 100 data sets of size 7000 genes and 8 replicates simulated with true $(p, c) = (0.01, 1.2)$, the average estimates were (0.0111, 2.76) and the medians were (0.00916, 1.39).

Since our estimates require four or more replicates, we cannot evaluate estimates for the Yuen data set. The estimates for the SRBI data are $\hat{p}_B = 0.942$ and $\hat{c}_B = 1.24$. The estimate $\hat{p}_B = 0.942$ seems strange, since we expect only a small proportion of the genes to be differentially expressed.

These unrealistic estimates suggest that our data do not fit the empirical Bayes model very well. This is not surprising, since our aim is to separate a mixture of zero-centred Normal distributions with different variances. By experience we know that our data are not Normally distributed, and the actual tail distribution of the data will slightly influence the estimates. We therefore investigate whether we can find a better model for microarray data using a full Bayesian setup.

3. HIERARCHICAL BAYES MODELS

In this section we present two specific hierarchical Bayes models for microarray data. We generally assume a model for a data set D, parameters u and hyperparameters θ, such that $D \sim \pi(D|u)$, $u \sim \pi(u|\theta)$ and $\theta \sim \pi(\theta)$. For details on hierarchical Bayes models and Markov Chain Monte Carlo (MCMC) we refer to Gilks *et al.* (1996, Chapter 1).

3.1 *Hierarchical Bayes model C*

From above, D are our M-values, the parameters u are our means μ_i and variances σ_i^2 and the hyperparameters θ are any parameters in the prior distributions of these parameters. In a first attempt to improve on the empirical Bayes model in Section 2.1, we only want to remove the dependency between the variances of the M-values σ_i^2 and the variances in the prior distribution of the non-zero means $c\sigma_i^2$. We assume (2.1) and that all genes have mean $\mu_i = 0$ ($I_i = 0$) except the proportion p, for which we apply a normal prior distribution $\mu_i|\tau^2 \sim N(0, \tau^2)$ ($I_i = 1$). We assume the variances of all genes to be a priori i.i.d. with $\sigma_i^2|\alpha, \beta \sim \Gamma^{-1}(\alpha, \beta)$ (inverse gamma).

The hyperparameters $(p, \tau^2, \alpha, \beta)$ are assumed to be independent. We use a beta prior $\beta(r, s)$ for p [e.g. $(r, s) = (1, 1)$], and the improper reference prior $1/\tau^2$ for τ^2, suggested by Yang and Berger (1996). For α and β we initially intended to use the reference prior suggested in Liseo (1993), but since microarrays yield data sets with thousands of variances, the resulting posterior distribution was too peaked for numerical evaluations. However, in such a peaked distribution, sampled hyperparameter values will be very close to any sensible estimates of these hyperparameters. Hence, we use the maximum likelihood estimates for α and β, rather than sampling from posterior distributions. The log-likelihood function for α and β given the variances is up to an additive constant,

$$\ell(\alpha, \beta \mid (\sigma_i^2)) = -N \log\Gamma(\alpha) - N\alpha \log\beta - \frac{\sum_i (1/\sigma_i^2)}{\beta} - (\alpha - 1)\sum_i \log(\sigma_i^2).$$

Setting the partial derivatives to zero we find the estimate $\hat{\beta} = \sum(1/\sigma_i^2)/(N\hat{\alpha})$. We estimate $\hat{\alpha}$ numerically. We call this approach model C. The prior distributions and resulting marginal posterior distributions for the Gibbs sampler MCMC method are summarized in Table 1.

In each step of the Markov chain we update each of the means μ_i and variances σ_i^2 followed by p, τ^2, α and β by sampling from the posterior distributions in Table 1. We discard the first 100 000 steps, then simulate further 1 000 000 steps out of which every 100th hyperparameter set is kept and averaged to give the estimates. The log odds-ratio C follows as a by-product, since

$$C_i = \log\left(\frac{\Pr(I_i = 1|D)}{\Pr(I_i = 0|D)}\right) = \log\frac{p_i}{1 - p_i},$$

where the p_i are the genewise posterior probabilities of differential expression from Table 1.

3.2 *Verification and discussion of model C*

To verify the estimates of the hyperparameters in the C model, we simulated small and large data sets with reasonable true parameter values $(p, \tau^2, \alpha, \beta) = (0.01, 1, 2.5, 3)$. As an example, Figure 1 shows this for the parameter p. We estimate the parameters by their posterior means.

Next we estimated the hyperparameters $(p, \tau^2, \alpha, \beta)$ for our two real data sets as (0.877, 0.078, 2.31, 7.97) for SRBI and (0.017, 4.07, 1.74, 6.55) for the Yuen data. The estimates $\hat{\alpha}_C$ and $\hat{\beta}_C$ are sensible, but $\hat{p}_C$ and $\hat{\tau}_C^2$ are very different between the data sets and far from correct for the SRBI data, as for the empirical Bayes model B. We conclude that avoiding the connection between the data variance and the

Table 1. *Prior and marginal posterior distributions for the full Bayes model C*

Prior distribution	Marginal posterior distribution
$M_{ij}\|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2)$	
For each gene i:	
$\mu_i \mid p, \tau^2 \begin{cases} \equiv 0, & \text{pr. } 1-p \\ \sim N(0,\tau^2), & \text{pr. } p \end{cases}$	$\mu_i \mid p, \tau^2, \sigma_i^2, (M_{ij}) \begin{cases} \equiv 0, \text{ pr. } 1-p_i^{(*)} \\ \sim N\left(\frac{nM_{i\cdot}}{n+\sigma_i^2/\tau^2}, \left(\frac{n}{\sigma_i^2}+\frac{1}{\tau^2}\right)^{-1}\right), & \text{pr. } p_i^{(*)} \end{cases}$
$\sigma_i^2 \mid \alpha, \beta \sim \Gamma^{-1}(\alpha, \beta)$	$\sigma_i^2 \mid \alpha, \beta, \mu_i, (M_{ij}) \sim \Gamma^{-1}\left(\frac{n}{2}+\alpha, \left(\frac{\sum_j (M_{ij}-\mu_i)^2}{2}+\frac{1}{\beta}\right)^{-1}\right)$
$p \sim \beta(r, s)$	$p \mid (\mu_i) \sim \beta(s_1 + r, N - s_1 + s)$, where $s_1 = \sum I_{(\mu_i \neq 0)}$
$\tau^2 \sim \pi(\tau^2) = \frac{1}{\tau^2}$ (improper prior)	$\tau^2 \mid (\mu_i) \sim \Gamma^{-1}\left(\frac{s_1}{2}, \frac{2}{\sum \mu_i^2}\right)$
α and β are ML-estimated, see text	

$^{(*)}\, p_i = \frac{p_{1i}}{1-p+p_{1i}}$, where $p_{1i} = p\sqrt{\frac{\sigma_i^2}{n\tau^2+\sigma_i^2}}\, \mathrm{e}^{\frac{n\tau^2 M_{i\cdot}^2}{2\sigma_i^2(\tau^2+\sigma_i^2/n)}}$.

variance of the non-zero means did not have much impact. The model is still not ideal for this microarray data set.

3.3 *Model discussion*

To understand how the model could be improved we changed model C in different ways and simulated data sets under each modification. The more similar the simulated data was to our real SRBI data set (Figure 2), the better was the model. The various modifications are summarized below.

Using a log-normal instead of an inverse gamma prior for σ_i^2 is a slight improvement on model C, but we believe the gain is too small to be worth the more complicated theory and longer computer running time it requires.

Microarray users often suggest that genes with a large degree of differential expression have larger variances. We tried to use (a) an inverse gamma prior for $\sigma_i^2|\mu_i$ with scale parameter of the form $1/(c_1|\mu_i| + c_2)$, so that genes with a large mean are assigned a large variance and (b) different inverse

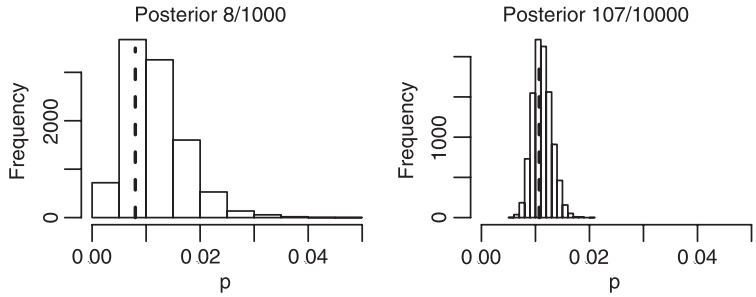

Fig. 1. Estimation of the proportion p of differentially expressed genes from data simulated from model C. Left: empirical posterior distribution of p for a small data set (1000 genes, 8 replicates) with 8 truly regulated genes. Right: empirical posterior distribution of p for a large data set (10 000 genes, 8 replicates) with 107 truly regulated genes. Dashed lines show true p. For details see text. The precision grows as the data set grows. A priori, $p \sim U(0, 1)$.

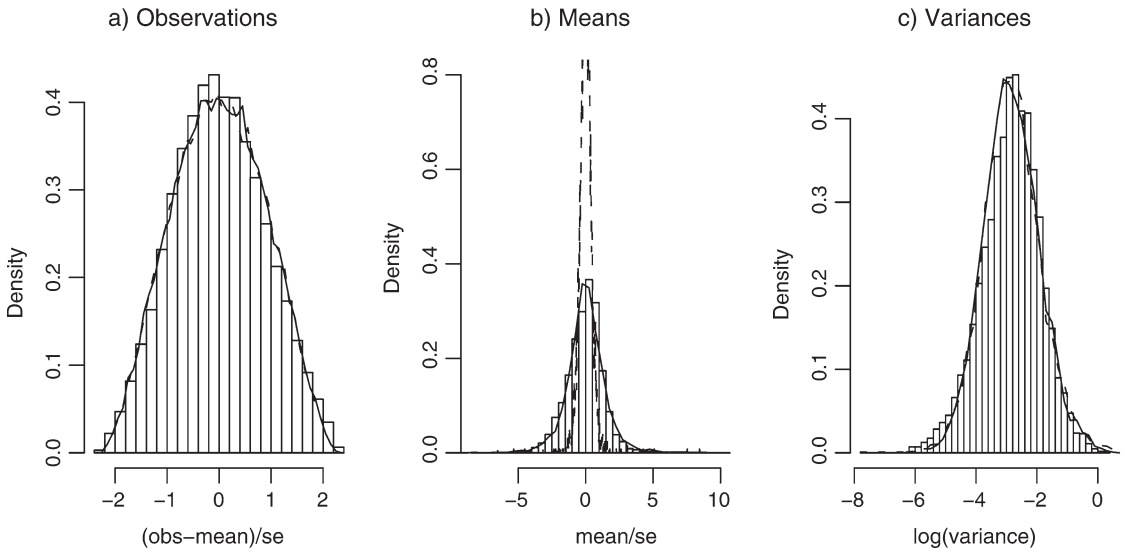

Fig. 2. SRBI compared to simulated data from models C and D. Bars represent SRBI, lines the model C-simulated data and dashed lines model D-simulated data. (a) $(M_{ij} - M_{i\cdot})/s_i$, (b) $M_{i\cdot}/s_i$ and (c) $\log(s_i^2)$, where $M_{i\cdot}$ and s_i are the genewise means and standard errors. The other models suggested in the text give results very similar to those of model C, i.e. the mean/se (b) has a more peaked density than real data.

gamma priors for $\sigma_i^2|\mu_i$ for zero and non-zero means, respectively, parameterized so that genes with $\mu_i \neq 0$ are assigned a larger variance than genes for which $\mu_i = 0$. None of these approaches resulted in a visible improvement.

In models B and C, the means of the regulated genes are modelled by zero-centered Normal distributions, although a gene with mean 0 is of course not regulated. Using (a) gamma priors of positive and negative non-zero means and (b) uniform priors of non-zero means, both parameterized so that zero means could not fall into the class of differentially expressed genes, we could not see any improvement.

Microarray data are thought to have thicker tails than the Normal distribution, but neither using t-distribution nor bilateral exponential distibutions for each μ_i and $M_{ij}|\mu_i, \sigma_i^2$ helped.

Only when allowing the non-regulated genes to have a small variance around zero could we make a large improvement towards fitting real data. In such a model the dispersion from zero can be decomposed into a true non-zero mean for regulated means, an *experimental* error and a residual standard error σ_i^2.

Figure 2 shows histograms of the means, variances and observations of the SRBI data, compared to the empirical densities of the same quantities for simulated data sets from models C and D. Model D includes the suggested experimental variance and is presented next.

3.4 *Hierarchical Bayes model D*

As a result of the last section we define a model D by changing the prior distribution for the means μ_i in model C. The normal prior distribution in (2.1) was kept as a sampling distribution, but we start from the distributions of the mean and variance, so that the formulae will hold for any linear model estimates (see Table 2). The prior distributions for the variances σ_i^2 and its hyperparameters were treated as in model C.

The means of the unregulated genes (proportion $1 - p$) have prior distribution $N(0, \tau_0^2)$ indicated by $I_i = 0$. Differentially expressed genes are modeled by a Normal distribution on either side of 0, in line with suggestions in the fourth paragraph of Section 3.3. There, we used gamma distributions,

Table 2. *Prior and marginal posterior distributions for the full Bayes model D. Here $M_{i\cdot}$ and s_i^2 can be any effect and variance estimates from a linear model. If $M_{i\cdot}$ is the average, $k = 1/n$ and $d = n - 1$*

Prior distribution	Marginal posterior distribution
$M_{i\cdot}\mid\mu_i, \sigma_i \sim N(\mu_i, k\sigma_i^2)$ $ds_i^2/\sigma_i^2\mid\sigma_i^2 \sim \chi^2(d)$	
For each gene i:	
$\mu_i \mid p, \tau_0^2, \lambda, \tau_1^2 \sim$	$\mu_i \mid p, \tau_0^2, \lambda, \tau_1^2\sigma_i^2, (M_{i\cdot}) \sim$
$\begin{cases} N(0, \tau_0^2), & \text{pr. } 1-p, \\ N(\lambda, \tau_1^2), & \text{pr. } p/2, \\ N(-\lambda, \tau_1^2), & \text{pr. } p/2, \end{cases}$	$\begin{cases} N\left(\frac{\tau_0^2 M_{i\cdot}}{\tau_0^2+k\sigma_i^2}, \left(\frac{1}{k\sigma_i^2}+\frac{1}{\tau_0^2}\right)^{-1}\right), & \text{pr. } 1-p_{0i}^{(*)} \quad (I_i = 0) \\ N\left(\frac{\tau_1^2 M_{i\cdot}+k\sigma_i^2\lambda}{\tau_1^2+k\sigma_i^2}, \left(\frac{1}{k\sigma_i^2}+\frac{1}{\tau_1^2}\right)^{-1}\right), & \text{pr. } p_{1i}^{(*)} \quad (I_i = 1) \\ N\left(\frac{\tau_1^2 M_{i\cdot}-k\sigma_i^2\lambda}{\tau_1^2+k\sigma_i^2}, \left(\frac{1}{k\sigma_i^2}+\frac{1}{\tau_1^2}\right)^{-1}\right), & \text{pr. } p_{-1i}^{(*)} \quad (I_i = -1) \end{cases}$
$\sigma_i^2 \mid \alpha, \beta \sim \Gamma^{-1}(\alpha, \beta)$	$\sigma_i^2 \mid \alpha, \beta, \mu_i, (M_{ij}) \sim \Gamma^{-1}\left(\frac{d+1}{2}+\alpha, \left(\frac{d}{2}s_i^2+\frac{1}{2k}(M_{i\cdot}-\mu_i)^2+\frac{1}{\beta}\right)^{-1}\right)$
$p \sim \beta(r, s)$	$p \mid (\mu_i) \sim \beta(s_1+s_{-1}+r, s_0+s)$, where $s_m = \sum I_{(I_i=m)}$
$\tau_0^2 \sim \pi(\tau_0^2) = \frac{1}{\tau_0^2}$ (improper prior)	$\tau_0^2 \mid (\mu_i) \sim \Gamma^{-1}\left(\frac{s_0}{2}, \frac{2}{\sum \mu_i^2 I_{(I_i=0)}}\right)$
$\tau_1^2 \sim \pi(\tau_1^2) = \frac{1}{\tau_1^2}$, $\tau_1^2 > \tau_0^2$ (improper prior)	$\tau_1^2 \mid (\mu_i) \sim \Gamma^{-1}\left(\frac{s_1+s_{-1}}{2}, \frac{2}{\sum(\mu_i I_i-\lambda)^2 I_{(I_i\neq 0)}}\right)$, $\tau_1^2 > \tau_0^2$ (truncated)
$\lambda \sim U(a, b)$	$\lambda \mid (\mu_i), \tau_1^2 \sim N\left(\frac{\mu_i I_i}{s_1+s_{-1}}, \frac{\tau_1^2}{s_1+s_{-1}}\right)$, $a < \lambda < b$ (truncated)
α and β are ML-estimated as for model C	

$$(*)\begin{cases} p_{0i} = \frac{f_{0i}}{f_{0i}+f_{1i}+f_{-1i}} \\ p_{1i} = \frac{f_{1i}}{f_{0i}+f_{1i}+f_{-1i}} \\ p_{-1i} = \frac{f_{-1i}}{f_{0i}+f_{1i}+f_{-1i}} \end{cases}, \quad \text{where } f_{0i} = \frac{1-p}{\sqrt{\tau_0^2+k\sigma_i^2}}\, e^{\frac{-M_{i\cdot}^2}{2(\tau_0^2+k\sigma_i^2)}},\ f_{li} = \frac{p}{\sqrt{\tau_1^2+k\sigma_i^2}}\, e^{-\frac{1}{2}\frac{(M_{i\cdot}-l\lambda)^2}{k\sigma_i^2+\tau_1^2}},\ l = -1, 1.$$

but conjugate Normal distributions ease the calculations and behave similarly. We included this in the model for interpretational reasons, although it did not turn out to be very important in Section 3.3. Hence, $\mu_i \sim N(\lambda, \tau_1^2)$ $(I_i = 1)$ or $\mu_i \sim N(-\lambda, \tau_1^2)$ $(I_i = -1)$ with equal probability $p/2$.

The model has three hyperparameters for μ_i, namely τ_0^2, τ_1^2 and λ. We applied the improper priors $\pi(\tau_0^2) = 1/\tau_0^2$ and $\pi(\tau_1^2) = 1/\tau_1^2$, $\tau_1^2 > \tau_0^2$. Generally, we have $\tau_0^2 \ll \tau_1^2$. We let $\lambda \sim U(a, b)$, where we suggest $a = \text{med}(|M_{i\cdot}|)$ and $b = 5 \times \max(|M_{i\cdot}|)$ in the implementation. Using improper priors can cause difficulties with Markov chain sampling. The limits of the τ_1^2 and λ prior distributions prevent the chain from jumping to incorrect interpretations such as $\tau_1^2 = 1000$ or $\lambda = 100$. Model D is summarized with the marginal posterior distributions in Table 2.

Markov chains were simulated as for model C, and the log odds-ratio was derived as

$$D_i = \log\left(\frac{\Pr(I_i \neq 0|D)}{\Pr(I_i = 0|D)}\right) = \log\frac{p_i}{1-p_i},$$

where p_i are the genewise posterior probabilities of differential expression from Table 2.

Model D was checked for the parameter estimates as for model C (see Section 3.2). The estimates of $(p, \tau_0^2, \lambda, \tau_1^2, \alpha, \beta)$ for our real data are (0.00349, 0.0676, 0.782, 1.16, 1.78, 10.9) for SRBI and

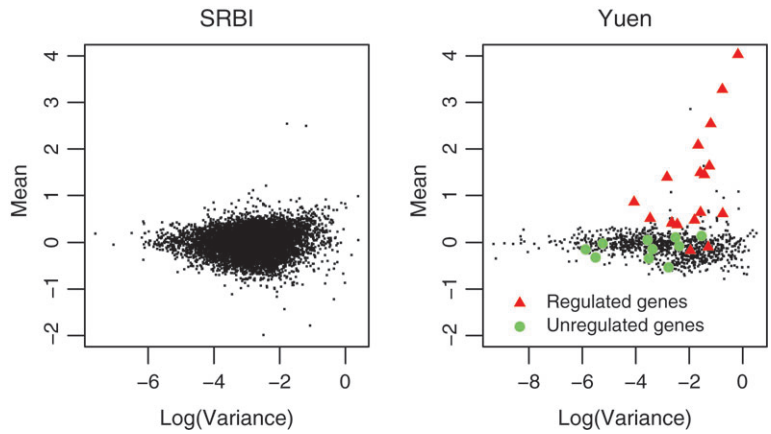

Fig. 3. Mean and variance scatter diagrams of the SRBI and Yuen data sets. The highlighted genes are known to be regulated or unregulated.

(0.0103, 0.00966, 1.73, 3.20, 1.57, 8.37) for the Yuen data set. The estimates of the proportion p of differentially expressed genes are more plausible than for model B, although the true values are unknown. The other estimates seem reasonable as judged from Figure 3. The Yuen data, with many fewer genes than SRBI (956 compared to 6356), seem to have a much larger spread of the extreme genes, as suggested by λ and τ_1^2, but a smaller variance of the other genes, as τ_0^2 suggests.

4. Comparison of statistics

The different empirical Bayes methods combine the gene-specific means and variances with the global distributions in different ways. To compare different statistics we have studied the numbers of false positive and false negative genes for different cutoff levels for simulated data and for the verified genes in the Yuen data set.

We included the following statistics in both studies:

average	The gene-specific mean statistic.
t	The t-statistic.
LOD	The log odds-ratio B reparameterized by Smyth (2004).
C	The log odds-ratio in Section 3.1.
D	The log odds-ratio in Section 3.4, which we believe has the most accurate model assumptions.
SAM	The magnitude of the t-statistic penalized with an extra factor a in the denominator (Tusher *et al.*, 2001). We use the 90th percentile of the standard errors for a, but there are other versions.
Baldi	$1 - p$, where p is the posterior probability of differential expression in Baldi and Long (2001).

4.1 *Simulated data*

By sampling with replacement of 2450 genes from the SRBI$^-$ data set we obtained the unregulated genes in one simulated set. Fifty regulated genes were added by sampling 25 means from $\Gamma(3.5, 0.25)$, 25 from

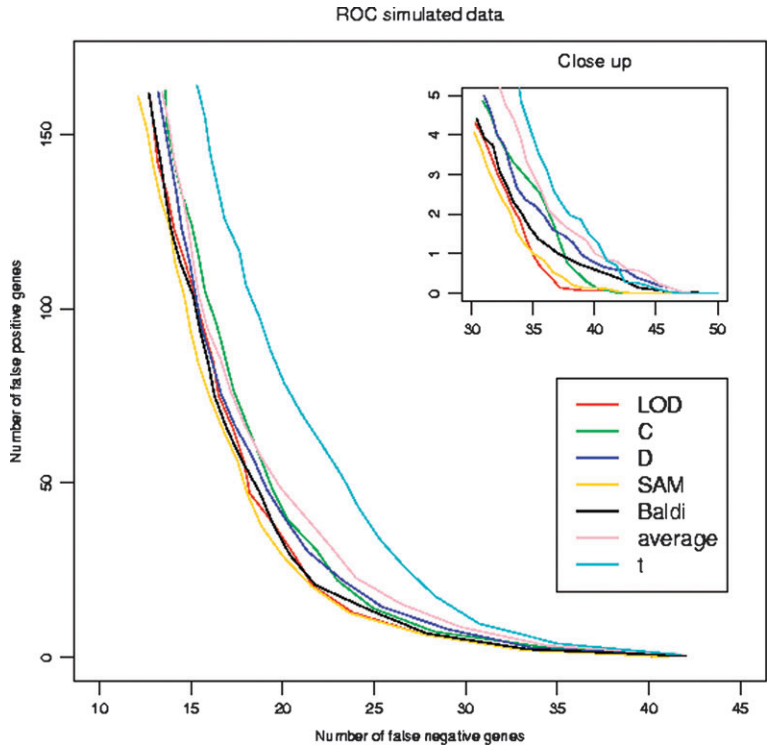

Fig. 4. ROC of the simulated data sets for seven different statistics, comparing the false discovery rates in terms of number of false negative and positive genes for different cutoffs. The lower the false numbers, the better the method.

$-\Gamma(3.5, 0.25)$ and 50 variances from all the gene sample variances in SRBI$^-$. Means and variances were matched randomly and M-values for the 50 regulated genes were sampled from the Normal distribution given the mean and variance of each gene. This procedure yields a data set with 2500 genes and 8 replicates, within which 1% of genes are upregulated and 1% are downregulated. We simulated 30 different such data sets.

The seven statistics were calculated for each data set. For C and D, we used 300 000 simulation steps, of which the first 10 000 were discarded and every 10th among the remainder were used. For each of the statistics to be compared, it is unclear where to draw the cutoff line connected to a specific level of significance when testing the hypothesis of no differential expression for a gene. We decided on a series of 20 cutoff values for each statistic, in a range from the 200th to the top observed level for that statistic.

For each data set and statistic, one cutoff level yields a number of false positive genes and a number of false negative genes. For the gene exactly at the cutoff level we accepted the null hypothesis of no differential expression. Each pair of counts results in one point in the false positives to false negatives graph, or receiver operating characteristic (ROC), of Figure 4. For a specific cutoff, each statistic was summarized by the average among the 30 simulated data sets. The line which is closest to the origin reflects the lowest number of false positive and negative genes and is taken to represent the best method.

We can see that the t-statistic is by far the worst method to use. Neither C nor D seem to be as good as the empirical Bayes methods LOD, SAM and Baldi, which are very close to each other.

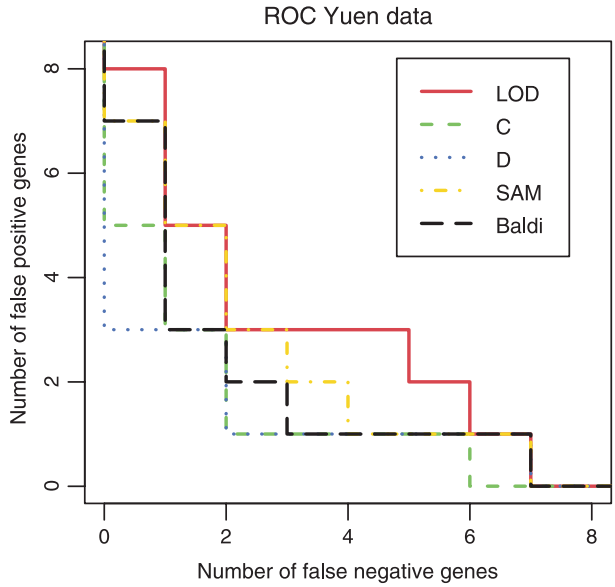

Fig. 5. ROCs of the Yuen data set for five different statistics, comparing the false discovery rates in terms of number of false negative and positive genes for different cutoffs. The lower the false numbers, the better the method. The regulated genes in this data set also have the largest absolute averages, so the *D* method shows the best performance (see text).

4.2 *Yuen data*

Here, 17 genes were found to be differentially expressed and 10 genes not differentially expressed (Yuen *et al.*, 2002). A sequence of cutoffs for each statistic was defined by the observed scores for each of the 27 known genes for that statistic. An ROC graph was constructed in Figure 5 as for the simulated data, but here the numbers of false negatives and false positives were only counted among the verified genes.

For clarity, the average and t-statistic are not displayed in Figure 5. The t-statistic showed the worst performance while the average was in the middle of the set of statistics. For the Yuen data set the LOD method is the worst of the Bayesian methods, and *D* is the best. By Figure 3 the regulated genes are also the genes with the largest averages, which we think are unusually large for microarray data. Comparing the different Bayesian statistics we conclude that *D* has the strongest dependence on the average compared to the variance.

5. Discussion and conclusions

We have presented two new hyperparameter estimates for the empirical Bayes model *B* in Lönnstedt and Speed (2002), including the estimate of the proportion of differentially expressed genes. The evaluation of the estimates for real data (e.g. $\hat{p}_B = 0.942$) suggests that the empirical Bayes model fits our microarray data poorly. This is not too surprising, since we know that the distribution of microarray data commonly is far from normal in the tails. We aimed to see whether the log odds-ratio could be improved by a model that better fits the data.

We present two full Bayesian models called *C* and *D*. Model *C* is just a minor change to *B*, avoiding the proportionality between gene variances and the prior variance of the means. We did not observe any major differences in favor of model *C*. Model *D* seems to fit our microarray data very well. By assuming that the genes are subject to an experimental variance component, in addition to the observation or residual

variance and, for regulated genes, the variance of a prior mean, we managed to simulate noise reasonably similar to that in our real data set.

Our paper also includes a comparison between existing empirical Bayes methods and the new models *C* and *D*. Our main conclusion is that for our data the full Bayesian models do not provide any improvement over the empirical Bayes methods. Neither *C* nor *D* seems worth the 5–20 hours long computer running time which they demand for a moderately large data set. We also note that model *C* is easier to evaluate numerically than the Broët *et al*. (2002) method. In our comparisons, however, the performance of the statistics is quite different between different data sets, which merely tells us that today's microarray data is difficult to analyze—they hold much noise and gain much from replication. Potentially interesting genes therefore should be investigated by follow-up experiments, e.g. PCR methods.

It is interesting that the empirical Bayes methods seem to outperform method *D*, which makes the more accurate model assumptions, as judged by the strange parameter estimates for *B* and by Figure 2. By comparison with model *D*, the empirical Bayes methods more strongly downweight the score for genes with a high variance over the microarray slides.

Since the model assumptions appear reasonable, the level of the log odds-ratio *D* should reflect the posterior probability of differential expression. A gene with log odds-ratio above zero has a larger probability of being regulated than of being unregulated. We checked the log odds-ratios *D* among the known regulated genes in the Yuen data set and found that 8 of the 17 genes had a positive score, suggesting that *D* is restrictive in rejecting the null hypothesis. All the unregulated genes in the Yuen data had negative scores. The empirical Bayes LOD had only 5 of the 17 scores above zero, and was hence even more restrictive, although the level of LOD is not really interpretable in that sense.

This study gives confidence in favour of using empirical Bayes methods for microarray data, and also widens the possibility to use full Bayes methods in situations where empirical Bayes cannot be applied.

ACKNOWLEDGMENTS

We thank Silvelyn Zwanzig and Terry Speed for valuable discussions and Terry Speed for letting us use the SRBI data set. We also thank Stuart Sealfon, Tony Yuen and Elisa Wurmbach for giving us access to the Yuen data, and Sanjib Basu for valuable comments. Our thanks also go to an anonymous referee for careful reading and constructive criticism of the manuscript. Financial support from Stockholm Bioinformatics Center is gratefully acknowledged.

REFERENCES

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.

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.

BUCKLEY, M. J. (2000). *The Spot User's Guide*. CSIRO Mathematical and Information Sciences. http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm.

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 Research* **10**, 2022–2029.

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

GILKS, W. R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

GOTTARDO, R., PANNUCCI, J. A., KUSKE, C. R. AND BRETTIN, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**, 597–620.

KERR, M. K. AND CHURCHILL, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

LISEO, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

LONG, A. D., MANGALAM, H. J., CHAN, B. Y. P., TROLLERI, L., HATFIELD, G. W. AND BALDI, P. (2001). Global gene expression profiling in *Escherichia coli* K 12: improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *Journal of Biological Chemistry* **276**, 19937–19944.

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

PARMIGIANI, G., GARRETT, E. S., ANBAZHAGAN, R. AND GABRIELSON, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B* **64**, 717–736.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, article 3.

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

YANG, R. AND BERGER, J. O. (1996). A catalog of noninformative priors, *Discussion Paper*, 97-42, ISDS. Durham, NC: Duke University.

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. AND SPEED, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

YUEN, T., WURMBACH, E., PFEFFER, R. L., EBERSOLE, B. J. AND SEALFON, S. C. (2002). Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Research* **30**, e48.

[*Received August 29, 2003; first revision January 23, 2004; second revision May 18, 2004; third revision August 12, 2004; fourth revision October 29, 2004; accepted for publication November 24, 2004*]