STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:3899–3914 (DOI: 10.1002/sim.1548)

On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles[‡]

C. M. Kendziorski[1,*,†], M. A. Newton[1,2], H. Lan[3] and M. N. Gould[3]

[1] *Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, U.S.A.*
[2] *Department of Statistics, University of Wisconsin, Madison, WI, U.S.A.*
[3] *McArdle Laboratory for Cancer Research, University of Wisconsin, Madison, WI, U.S.A.*

SUMMARY

DNA microarrays provide for unprecedented large-scale views of gene expression and, as a result, have emerged as a fundamental measurement tool in the study of diverse biological systems. Statistical questions abound, but many traditional data analytic approaches do not apply, in large part because thousands of individual genes are measured with relatively little replication. Empirical Bayes methods provide a natural approach to microarray data analysis because they can significantly reduce the dimensionality of an inference problem while compensating for relatively few replicates by using information across the array. We propose a general empirical Bayes modelling approach which allows for replicate expression profiles in multiple conditions. The hierarchical mixture model accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations. Two distinct parameterizations are considered: a model based on Gamma distributed measurements and one based on log-normally distributed measurements. False discovery rate and related operating characteristics of the methodology are assessed in a simulation study. We also show how the posterior odds of differential expression in one version of the model is related to the ratio of the arithmetic mean to the geometric mean of the two sample means. The methodology is used in a study of mammary cancer in the rat, where four distinct patterns of expression are possible. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: hierarchical model; mixture model; microarray; differential expression; breast cancer

1. INTRODUCTION

Enabled by resources created from genome sequencing projects, DNA microarray technology has emerged as a fundamental measurement tool in the study of diverse biological systems.

[*] Correspondence to: Christina Kendziorski, Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 6729 Medical Sciences Center, 1300 University Avenue, Madison, WI 53703, U.S.A.
[†] E-mail: kendzior@biostat.wisc.edu
[‡] Presented at the Twenty-second Annual Conference of the International Society for Clinical Biostatistics, 19–23 August 2001, Stockholm, Sweden.

Contract/grant sponsor: NIH; contract/grant numbers: TA-CA 09565, CA 64364, CA 28954, CA 77494

Received January 2002
Accepted February 2003

Microarrays offer an unprecedented ability to perform large-scale studies of gene expression. As a result, the focus of many research efforts has shifted from individual genes to multiple genes and the complicated and orchestrated ways in which they interact to maintain life.

With the shift from individual to integrated analysis in molecular biology comes a shift in the related statistical questions posed and methods required. The number of measurements of distinct genes across an array greatly exceeds that for any individual gene. Thus, we as statisticians are faced with the 'large p, small n' paradigm [1, 2]. Empirical Bayes methods provide a natural approach to microarray data analysis because they can significantly reduce the dimensionality of an inference problem involving many unknown parameters (for example, Efron and Morris [3, 4]). Our earlier work described a version of parametric empirical Bayes analysis for spotted microarrays and was restricted to so-called 'single-slide' data in which each gene produces two measurements, one from each cell condition [5]. The proposed empirical Bayes methodology provides improved estimation of expression fold-change and allows for the assessment of differential expression by the calculation of a posterior odds. In spite of there being very little data per gene, the methodology works because inference about a given gene uses information on the fluctuations of expression measurements from all genes. One goal of the present paper is to extend the parametric empirical Bayes calculations beyond the case of single-slides in two conditions, and thus to allow replicate expression profiles in multiple cell conditions.

The methodological work presented here is motivated in part by an experiment to study gene expression in a rat model of breast cancer (see Section 2). Microarray data were obtained from four distinct inbred lines (two parentals and two offspring congenic lines). The parental strains differ in their susceptibility to breast cancer and identifying differentially expressed genes could provide insight into the genetic basis of this difference. An interesting feature of the present study is the presence of the four interrelated groupings (the four inbred lines). For each gene, we are interested in making inference about the pattern of differential expression among the four groups. We are not simply asking if there is differential expression between two conditions.

The development of statistical methods to address the two condition problem has recently received much attention. A general approach is to conduct a hypothesis test at each gene and then correct for multiple tests. Most of the test statistics currently used are t (or t-like) and differ primarily in the estimation of the variance. Dudoit *et at.* [6] use a t-statistic with variance estimated by the within gene sample variance and go on to address the multiple testing problem extensively using permutation analysis. Tusher *et al.* [7] also use the within-gene sample variance, but adjust the denominator of their test statistic by adding a constant to account for the dependence between the relative difference in expression and absolute intensity; they address the multiple testing problem using the method of false discovery rates. Baldi and Long [8] use the posterior variance derived from a Bayesian analysis and do not consider the multiple testing problem. Methods such as these which treat the genes as separate fixed effects may have reduced efficiency when compared to empirical Bayes methods which treat the genes as arising from some population, and thus which allow a level of information sharing amongst genes.

The two-group empirical Bayes method originally proposed by Newton *et al.* [5] amounts to a simple mixture-model calculation. Stochastically, each gene is either differentially expressed or not. Those genes which are equivalently expressed present data according to some background distribution, and those which are differentially expressed present data according to a different distribution. The specific forms of these distributions arise by another layer of mixing over the latent mean expression level for each gene. The latent mean values are treated not as fixed effects (as they would be in the standard analyses outlined above) but follow some specified distribution. With these components in place, inference about differential expression amounts to computing the posterior probability of that event, conditional on the measurements. The analysis is *empirical* Bayes because the small number of unknown parameters which index the component distributions are estimated from the data. In Section 3 we describe the extension of this approach to replicate profiles in multiple conditions. We consider two distinct parametric families: a model based on Gamma distributed measurements and one based on log-normally distributed measurements. As it is often observed (for example, Chen *et al.* [9]), a constant coefficient of variation is built in to both models. The models also account for differential variation in apparent fold change (for example, Dudoit *et al.* [6], Newton *et al.* [5] and Tusher *et al.* [7]). We use the methodology to analyse rat mammary epithelium expression profiles in Section 6.

There are other mixture-modelling approaches to expression data analysis. Working with a specific experimental design, Efron *et al.* [10, 11] describe empirical Bayesian calculations which relax the parametric assumptions. After a long series of preprocessing steps, each gene yields a one-dimensional test statistic whose marginal distribution turns out to be known and whose null distribution (that is, on equivalent expression) can be non-parametrically estimated. Lee *et al.* [24] also use the idea of a two-group mixture model for expression analysis; their calculations were in a slightly different context and were applied to parameter estimates from a first-stage analysis. Here we do not endeavour to extend either of these approaches to the case of multiple conditions, but in Section 5 we do offer some numerical comparisons of false discovery rate between our proposal and the non-parametric method in the context of two conditions.

2. ANIMAL MODELS OF BREAST CANCER

The risk of developing breast cancer is affected by both environmental and genetic factors. Known genetic factors include inherited mutant alleles of genes such as p53, BRCA1 and BRCA2. Studies have indicated that individuals may carry genes that diminish the consequences of BRCA mutations [13, 14], but such modifier loci are particularly difficult to identify in human populations as potential genetic effects are confounded by environmental effects that are not easily controlled. An alternative approach is to study animal models of breast cancer. Ideally, human homologues of identified genes could then be used to directly evaluate their effects on breast cancer risk in human populations.

In an effort to identify potential resistance or modifier loci. Shepel *et al.* [15] considered crosses between an inbred Copenhagen (COP) rat strain that is almost completely resistant to mammary carcinogenesis and an inbred Wistar-Furth (WF) rat strain which is highly susceptible to mammary carcinoma following the carcinogen DMBA. Four regions likely to contain genes affecting tumour dynamics were identified. To further narrow down these regions, intermediate inbred lines are being produced which carry the homozygous WF/WF genotype throughout the genome except on a relatively small region of interest where the animals are homozygous COP/COP. Such animal populations are referred to as congenic lines (Figure 1). We are interested in identifying genes differentially regulated among the parental strains (COP

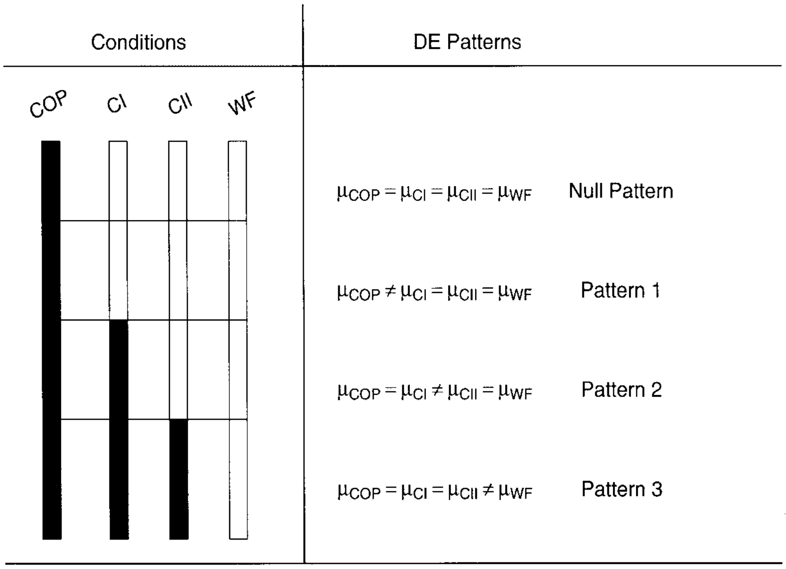
Conditions
DE Patterns
COP
CI
CII
WF
$\mu_{COP} = \mu_{CI} = \mu_{CII} = \mu_{WF}$
Null Pattern
$\mu_{COP} \neq \mu_{CI} = \mu_{CII} = \mu_{WF}$
Pattern 1
$\mu_{COP} = \mu_{CI} \neq \mu_{CII} = \mu_{WF}$
Pattern 2
$\mu_{COP} = \mu_{CI} = \mu_{CII} \neq \mu_{WF}$
Pattern 3

Figure 1. Schematic diagram showing animal lines (conditions) from which mRNAs were obtained (left) along with differential expression patterns (right). Genotypes shown in black (COP/COP) and white (WF/WF) are not drawn to scale (the homozygous COP/COP region is approximately 30 cM in congenic line CI and 1.5 cM in CII). True expression intensities for each group are denoted by μ. Note that differences in genotype do not imply differences in expression.

and WF) and the derived congenic lines as such genes may be consequential in the tumour susceptibility differences present among each of these strains. Our earlier work on empirical Bayes approaches for microarrays has been extended to address this problem. The extensions are discussed in detail in Section 3.

3. HIERARCHICAL MODELLING FRAMEWORK

Our models attempt to describe the probability distribution of a set of expression measurements taken on a gene g. We assume that some preprocessing technique has been used to adequately normalize the data so that the measurements can be viewed as *bona fide* approximations of relative gene expression in the sampled cells. The expression measurements may arise from cells under different conditions, and there may be replicate measurements in each condition. The number of expression patterns possible depends on the number of conditions from which the expression measurements were obtained. For example, if measurements are taken from two conditions, two patterns of expression – equivalent expression and differential expression between the two conditions – are possible. Given three conditions, $\binom{3}{0}+\binom{3}{2}+\binom{3}{3}=5$ expression patterns are possible. These include equivalent expression across the three conditions, altered expression in just one condition, and distinct expression in each condition. With microarrays from four cell conditions there are 15 different patterns. (The total number of patterns is equal to the Bell exponential number of possible set partitions, in fact.) As we see in the rat mammary study which has four cell conditions (Section 2), we can sometimes reduce the total number of patterns to a more manageable level, and in that case we reduce to four interesting patterns (Section 6).

Suppose in the general case that $m+1$ distinct patterns of expression are possible for a data vector $\mathbf{d}_g=(d_{g,1},\ldots,d_{g,N})$ measuring a gene g in N conditions. Then for any pattern k, the set of experimental conditions $S=\{1,2,\ldots,N\}$ is partitioned into $r(k)$ mutually exclusive and exhaustive subsets $\{S_{i,k};\ i=1,2,\ldots,r(k)\}$, where any measurements contained in a subset $S_{i,k}$ share a common latent mean level of expression.

On a null hypothesis there is equivalent expression among the conditions ($r(k)=1$ since all data values share the same mean expression level), and the data for a given gene arise from a joint probability density function (pdf) $f_0(\mathbf{d}_g)$. Alternatively the data are in expression pattern $k\neq 0$, and the joint pdf is $f_k(\mathbf{d}_g)$. *A priori* we do not know which situation manifests itself for gene g, and so we introduce discrete mixing parameters p_k to denote the unknown probabilities of expression pattern k. Thus, the marginal distribution of the data is given by a mixture of the form

$$\sum_{k=0}^{m} p_k f_k(\mathbf{d}_g) \tag{1}$$

The posterior probability of expression pattern k is then

$$P(k|\mathbf{d}_g)\propto p_k f_k(\mathbf{d}_g) \tag{2}$$

and the posterior odds in favour of pattern k is

$$\text{odds}_{g,k}=\frac{p_k}{1-p_k}\,\frac{f_k(\mathbf{d}_g)}{1-f_k(\mathbf{d}_g)} \tag{3}$$

Naturally, the pattern specific predictive density for pattern k is given by

$$f_k(\mathbf{d}_g)=\prod_{i=1}^{r(k)} f(\mathbf{d}_{g,S_{i,k}}) \tag{4}$$

where $f(\mathbf{d}_{g,S_{i,k}})$ is the pdf for the data indexed by subset $S_{i,k}$.

We assume that measurements which share a common mean expression level μ_g arise independently and identically from an observation component $f_{\text{obs}}(\cdot|\mu_g)$. Two parametric forms are considered in Section 4. Our approach is to consider μ_g as arising from some genome-wide distribution $\pi(\mu_g)$, which represents fluctuations in mean expression levels among genes. Were we to treat the μ_g as a fixed effect, we would not take advantage of information sharing. With these components in place, $f(\mathbf{d}_{g,S_{i,k}})$ is the predictive density of $\mathbf{d}_g$ having integrated away the mean value common to all measurements in subset $S_{i,k}$

$$f(\mathbf{d}_{g,S_{i,k}})=\int\left(\prod_{s\in S_{i,k}} f_{\text{obs}}(\mathbf{d}_{g,s}\mid\mu_g)\right)\pi(\mu_g)\,\mathrm{d}\mu_g \tag{5}$$

The posterior probabilities given in equation (2) summarize our inference about expression patterns at each gene. They can be used to identify genes with altered expression in at least one group, to classify genes into distinct expression groups, or to order genes within groups. Before posterior summaries can be evaluated, however, we must first specify distributional forms for the components of the hierarchical mixture model.

4. THE GAMMA-GAMMA AND LOGNORMAL-NORMAL MODELS

The general mixture model in Section 3 is specified by an observation component $f_{\text{obs}}(\cdot|\mu_g)$ which characterizes fluctuations in repeated measurements from a gene having latent mean expression level μ_g, and a second component $\pi(\mu_g)$ which describes fluctuations in these means among genes. Since properties of individual experiments affect each source of variation, the distributions governing these sources are to some extent experiment dependent. However, there are characteristics inherent to microarray data that are repeatedly observed across experiments. These include constant coefficients of variation [5, 9] as well as dependencies between intensity ratio variation and magnitude [5, 6]. These characteristics provide insight into appropriate distributional forms. Here we describe two particular versions of the general mixture formulation that maintain these properties.

In the gamma-gamma (GG) model, the observation component is a gamma distribution having shape parameter $\alpha > 0$ and a mean value μ_g; thus, with *scale* parameter $\lambda_g = \alpha/\mu_g$

$$f_{\text{obs}}(z|\mu_g) = \frac{\lambda_g^{\alpha} z^{\alpha-1} \exp\{-\lambda_g z\}}{\Gamma(\alpha)}$$

for measurements $z > 0$. Note that the coefficient of variation in this distribution is $1/\sqrt{\alpha}$, taken to be constant across genes g. Matched to this observation component is a marginal distribution $\pi(\mu_g)$ which we take to be an inverse gamma. More specifically, fixing α, the quantity $\lambda_g = \alpha/\mu_g$ has a gamma distribution with shape parameter α_0 and scale parameter v. Thus three parameters are involved, $\theta = (\alpha, \alpha_0, v)$, and, upon integration, the joint predictive density corresponding to (5) has the form

$$f(z_1, z_2, \ldots, z_n) = K \frac{(\prod_{i=1}^{n} z_i)^{\alpha-1}}{(v + \sum_{i=1}^{n} z_i)^{n\alpha+\alpha_0}} \tag{6}$$

where

$$K = \frac{v^{\alpha_0}\Gamma(n\alpha + \alpha_0)}{\Gamma^n(\alpha)\Gamma(\alpha_0)}$$

From this result one can calculate the posterior probability of any given expression pattern following the prescription in Section 3. In the special case of two conditions, the posterior odds for differential expression (3) simplify to

$$\text{odds}_g = \frac{p}{1-p} K' \frac{\left(\sum_{i=1}^{n_1} x_{g,i} + \sum_{i=1}^{n_2} y_{g,i} + v\right)^{N\alpha+\alpha_0}}{\left(\sum_{i=1}^{n_1} x_{g,i} + v\right)^{n_1\alpha+\alpha_0} \left(\sum_{i=1}^{n_2} y_{g,i} + v\right)^{n2\alpha+\alpha_0}} \tag{7}$$

where

$$K' = \frac{v_0^{\alpha}\Gamma(n_1\alpha + \alpha_0)\Gamma(n_2\alpha + \alpha_0)}{\Gamma(\alpha_0)\Gamma(N\alpha + \alpha_0)}$$

and recall that $N = n_1 + n_2$ is the total number of observations on gene g. The odds may be computed as soon as we have estimates in hand for $\theta = (\alpha, \alpha_0, v)$. In Section 7 we point out an interesting connection between these posterior odds and the arithmetic-geometric mean inequality.

The GG calculations derived above extend those presented in Newton *et al.* [5] to replicates and multiple conditions. Many investigators would consider as reasonable a different model for the array measurements – one in which the log-transformed measurements have a Gaussian observation component. We may use this in our hierarchical mixture model as follows. Let us say the natural logarithms of the measurements are denoted $\tilde{\mathbf{x}}_g$ and $\tilde{\mathbf{y}}_g$. The latent gene-specific mean μ_g is now a mean for the log-transformed measurements, and these measurements have a sampling variance σ^2 which we treat as common to all genes. Note that the coefficient of variation for the original measurements becomes $\sqrt{\{\exp(\sigma^2) - 1\}}$ in this model. A conjugate prior for the μ_g is normal with some underlying mean μ_0 and variance τ_0^2. Integrating as in (5), the joint predictive density f for an n-dimensional input becomes Gaussian with mean vector $\boldsymbol{\mu}_0 = (\mu_0, \mu_0, \ldots, \mu_0)^{\mathrm{T}}$ and exchangeable covariance matrix

$$\mathbf{\Sigma}_n = (\sigma^2)\mathbf{I}_n + (\tau_0^2)\mathbf{M}_n$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix and M_n is an $n \times n$ matrix of ones. This basic formulation has been well studied (for example, Carlin and Louis [16]). In our context there is an additional layer of discrete mixing, and we may derive the posterior probability of different expression patterns following (2). For the special case of two conditions, the odds of differential expression (3) may be written in terms of quadratic forms. Let $\delta_g = (\tilde{\mathbf{x}}_g, \tilde{\mathbf{y}}_g)^{\mathrm{T}} - \boldsymbol{\mu}_0$ denote the centred transformed full data vector for gene g

$$\text{odds}_g = \frac{p}{1-p}\sqrt{\left(\frac{|\Sigma_N|}{|\Sigma_*|}\right)}\exp\left\{-\frac{1}{2}\,\delta_g^{\mathrm{T}}(\Sigma_*^{-1} - \Sigma_N^{-1})\delta_g\right\}$$

where Σ_* is the $N \times N$ block-diagonal matrix with Σ_{n_1} in the upper left block and Σ_{n_2} in the lower right block.

For either the log-normal-normal (LNN) model or GG model, we can use the method of maximum (marginal) likelihood to obtain estimates of the small set of unknown parameters. (In the GG model, $\theta = (\alpha, \alpha_0, v)$ and in LNN, $\theta = (\mu_0, \sigma^2, \tau_0)$.) The mixing proportions are additional parameters. The marginal log-likelihood is a sum over genes g of terms (1) and this may be optimized by various methods. We use the S-plus program *nlminb* [17]. For two conditions, the mixing proportions are estimated directly using *nlminb*. In the case of three or more conditions, we use the EM algorithm to handle the vector of mixing proportions [18] (see Appendix).

5. SIMULATIONS

The proposed methodology provides a way to infer patterns of differential expression among two or more conditions, but it relies on parametric model assumptions and the implementation of numerical optimization methods. To assess the methodology we performed a small set of simulation studies. These provide some insight into whether or not the parameters are well estimated, how much inference is affected by fitting a model different from the one which generated the data, and perhaps most importantly, they provide information on error rates in the inference of differential expression.

First, we simulated the GG model with 10,000 genes in two conditions, having three replicates in each condition. We took model parameters similar to those obtained in Newton *et al.*

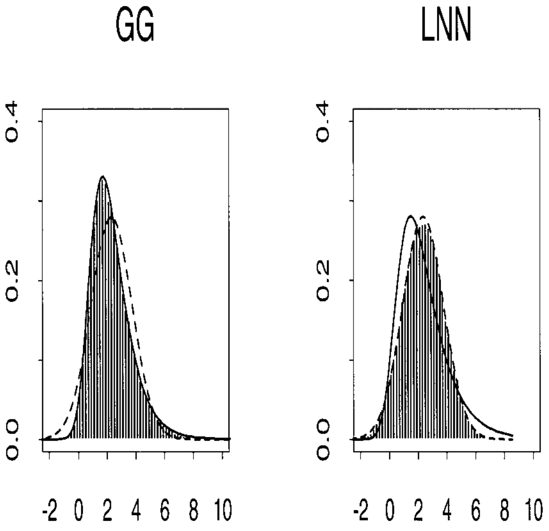

Figure 2. Histograms are of intensities (log scale) simulated under the GG (left) or LNN (right) model. Solid line is fitted marginal density from the GG model and dashed line is fitted marginal density from the LNN model.

[5] ($\alpha = 10, \alpha_0 = 0.9$ and $v = 0.5$). The prior probability that a gene is differentially expressed was set to $p = 0.2$. The GG and LNN mixture models described in Section 4 were each fit to these simulated data. Histograms of the simulated data along with the fitted marginal densities are shown in the left panel of Figure 2. As expected, the fitted GG marginal density more closely describes the simulated data.

Next, we simulated a similar data set under the LNN model ($\mu_0 = 2.3, \sigma = 0.3$, and $\tau = 1.39$); each mixture model was again fit to the simulated data. As shown in the right panel of Figure 2, the simulated data is better described by the LNN density. Although expected, this result illustrates that comparing the marginal densities to the empirical distribution can give insight into which model assumptions are more appropriate.

We did a more formal comparison of GG and LNN by calculating a log Bayes factor to measure the relative fit of these models [19]. The log Bayes factor is the difference of the log predictive densities calculated under GG or LNN assumptions (the general form of the predictive density is given by equation (1)). For each simulated data set, the Bayes factor correctly identified the model generating the simulated data. Considering the success of either approach in identifying the underlying model, one might think that the parametric assumptions have a substantial effect on which genes are identified as differentially expressed. We find that this is not the case.

The differences in the simulated data which allow for model identification do not seem to impact the mixture model's ability to identify differentially expressed genes. For this simulation, 1968 (1952) genes in the GG (LNN) data happened to be differentially expressed. Each method applied to each data set identified about 1470 genes as differentially expressed (that is, odds > 1). Out of those identified, approximately 95 per cent were correct.

Table I. Summary of parameter estimates for GG model applied to GG simulated data. Parameter estimates are averaged over 100 simulations; standard error is shown in parentheses. For each simulation, $(\alpha, \alpha_0, \nu) = (10, 0.9, 0.5)$.

	p				
	0.1	0.2	0.3	0.4	0.5
$\hat{\alpha}$	10.001 (0.0098)	9.997 (0.0087)	9.995 (0.0104)	9.993 (0.0099)	10.009 (0.01)
$\hat{\alpha}_0$	0.900 (0.0016)	0.900 (0.0015)	0.897 (0.0015)	0.900 (0.0014)	0.901 (0.0013)
$\hat{\nu}$	0.499 (0.0012)	0.500 (0.0011)	0.500 (0.0011)	0.500 (0.0012)	0.500 (0.0011)
$\hat{p}$	0.101 (0.0005)	0.201 (0.0007)	0.298 (0.0008)	0.401 (0.0008)	0.501 (0.0009)

Table II. Summary of parameter estimates for LNN model applied to LNN simulated data. Parameter estimates are averaged over 100 simulations; standard error is shown in parentheses. For each simulation, $(\mu_{10}, \sigma, \tau) = (2.33, 0.33.1.39)$.

	p				
	0.1	0.2	0.3	0.4	0.5
$\hat{\mu}_{10}$	2.328 (0.0018)	2.333 (0.0021)	2.329 (0.0015)	2.332 (0.0017)	2.33 (0.0015)
$\hat{\sigma}$	0.332 (0.0001)	0.331 (0.0001)	0.332 (0.0002)	0.331 (0.0002)	0.332 (0.0002)
$\hat{\tau}$	1.390 (0.0013)	1.386 (0.0014)	1.390 (0.0011)	1.392 (0.0012)	1.391 (0.0012)
$\hat{p}$	0.1 (0.0005)	0.2 (0.0007)	0.3 (0.0009)	0.399 (0.0009)	0.501 (0.0008)

Simulations were repeated to assess the sensitivity, specificity, positive and negative predictive values, and false discovery rates of the methodology. We varied the proportion p of differentially expressed genes from 0.1 to 0.5 (in increments of 0.1). For each fixed proportion, 100 sets having six arrays each (three replicates in two conditions, as above) were simulated. Parameter values were defined as above. Odds were calculated using both GG and LNN models. Sensitivity is calculated as the average (over the 100 simulations) of the fraction of differentially expressed genes correctly identified by the method (odds >1); specificity is the average of the fraction of equivalently expressed genes correctly identified (odds $\leqslant 1$). The positive predictive value (PPV) is defined as the average of the fraction of genes with odds >1 that are truly differentially expressed; the negative predictive value (NPV) is the average of the fraction of genes with odds $\leqslant 1$ that are equivalently expressed. The false discovery rate (FDR) is the average of the ratio of the number of false positives to the number of genes identified as differentially expressed.

Parameter estimates averaged over the 100 simulations are given in Tables I and II. As shown, the parameter estimates are close to the true values, with little standard error. The operating characteristics of each approach are similar under different simulation assumptions. For each method, the sensitivity ranges from 65 per cent to 80 per cent and is increasing with increasing p. The specificity is at or above 95 per cent for each method and each value of p considered. The positive predictive value ranges between 94 per cent and 95 per cent, while the negative predictive value decreases from near 97 per cent when $p = 0.1$ to near 80 per cent when $p = 0.5$. The average false detection rate (FDR), near 0.05 for all values of p, increased slightly with increasing p. A graphical representation is shown in Figure 3.

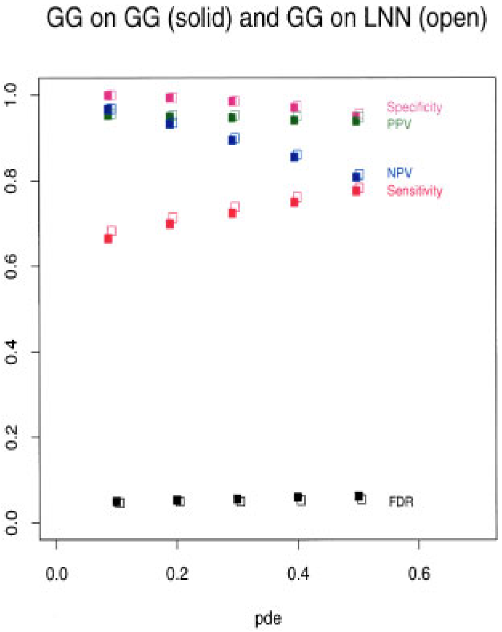

Figure 3. Operating characteristics: results are shown for the GG model applied to simulated data. To minimize overlap, we jittered the horizontal component. Closed characters imply identical model and simulation assumptions (GG model applied to GG data); open characters imply the opposite (GG model applied to LNN data).

The FDR estimates suggest that using an odds value greater than one as a rejection rule results in a type I error rate near 0.05. Interestingly, the estimates of the FDR are similar to those reported by Efron *et al.* [10] in assessment of their empirical Bayes approach. A lower bound on p for the data set considered there is estimated to be 0.189. The authors consider the FDR rates using the posterior probability of differential expression at values greater than and equal to 0.9. This corresponds to an odds$>$9. They report an FDR of 0.0048 at this level. Our results are similar. For $p=0.2$ and odds$>$9.1, the FDR averaged over 100 simulations was 0.0054 (GG on GG), 0.0052 (GG on LNN), 0.0057 (LNN on GG) and 0.0049 (LNN on LNN); standard errors were all less than 0.0003.

A possibility not so far considered is that the underlying distribution of μ_g is not well approximated by either of the parametric forms described above. For instance, if some genes are not truly expressed in one condition, then the distribution of μ_g may be better approximated by a bimodal distribution with one mode near zero. A full study of the robustness of our proposed methodology to the form of the mixing distribution is beyond the scope of the present paper, but we report here the results of a modest investigation to address the possibility of unexpressed genes and their effect on the inferences. Using the same parameters as above, we simulated 5000 genes according to the GG model with three replicates in each of two conditions and with $p=0.2$. A proportion ω of the genes are expressed and have mean intensity from an inverse gamma distribution with shape α_0 and scale v; the remaining $1-\omega$ fraction of the genes have mean intensity from an inverse gamma distribution with scale hv

where $h \ll 1$. In spite of some bias in estimated parameters, operating characteristics of the proposed method were not adversely affected by this model misspecification. For example, with $\omega = 3/4$ and $h = 1/10$, the average FDR and sensitivity were 0.036 and 0.777, respectively. These values are similar to those from the larger simulation study which assumed a unimodal distribution for μ_g. Other operating characteristics were also similar. By direct calculation, one may further derive a relationship between the odds for differential expression computed in the correct bimodal model to the same odds computed in the standard unimodal set up, using a fixed set of parameters. The ideally-computed bimodal odds, in the limit as $h \to 0$, are smaller only by a factor ω compared to the odds computed in the standard unimodal model (derivation not shown). Other investigations have shown that odds for differential expression computed using a non-parametric mixing distribution $\pi(\mu_g)$ can lead to similar inferences as the GG model presented here [20], thus suggesting a level of robustness of the proposed method.

6. CASE APPLICATION

As discussed in Section 2, we are interested in the identification of genes that are differentially regulated among parental rat strains (COP and WF) and two derived congenic lines (CI and CII) in mammary epithelial cells. The size of the homozygous COP/COP region is approximately 30 centimorgans (cM) in congenic line CI and 1.5 cM in congenic line CII (see Figure 1). A cM is a unit of measure to quantify distances between genome regions. In particular, 1 cM is equal to a 1 per cent chance that one location on the genome will be separated from a second location due to crossing over in a single generation. In humans, 1 cM is equivalent, on average, to 1 million base pairs.

By a standard protocol, mammary epithelial cells were harvested from untreated 12-week-old females. Messenger RNAs were extracted, prepared and then probed using a set of three Affymetrix Rat Genome U34 chips. In most cases, these mRNAs were pooled from samples of four genetically identical animals to reduce animal to animal variation. Intensity measurements were obtained for 26,379 genes recorded on 10 chip sets: 1 COP, 2 WF, 5 CI and 2 CII lines.

All data were processed through DNA-chip analyzer [21]. DNA-chip analyzer (dChip) uses a statistical model for probe level data to account for artifacts such as probe-specific biases. Corrected and normalized model-based estimates of gene expression were obtained for 25,248 genes (1131 were identified as outliers). A small fraction of the measurements are negative and these cannot be used by the model fitting procedures, so they are omitted for that purpose (796 out of 25,248). These observations can be included in the posterior probability calculations as long as they are set to a boundary value.

Both the GG and LNN models were used to categorize patterns of gene expression across the parental strains and derived congenic lines. For these four conditions, there are 15 possible expression patterns; however, if latent expression in each congenic matches one of the parentals, only four expression patterns are possible (see Figure 1). A null pattern consists of equivalent expression across the four conditions. The other three patterns allow for differential expression between the parental strains, with the congenic lines exhibiting the same mean expression as one of the parentals. Specifically, differential expression of the COP parent only is specified in pattern 1, between the congenics in pattern 2, and of the WF parent

Table III. Parameter estimates for GG ($\hat{\alpha}, \hat{\alpha}_0, \hat{\nu}$) and LNN ($\hat{\mu}_0, \hat{\sigma}, \hat{\tau}$) models used in two-group comparisons between parentals and in four-group comparisons among the parentals and derived inbred lines.

GG	$\hat{\alpha}$	$\hat{\alpha}_0$	$\hat{\nu}$	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
Two groups	12.490	0.919	35.842	0.998	0.002	NA	NA
Four groups	16.738	0.883	24.398	0.985	0.012	0.002	0.001
LNN	$\hat{\mu}_0$	$\hat{\sigma}$	$\hat{\tau}$	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
Two groups	6.775	0.292	1.193	0.993	0.007	NA	NA
Four groups	6.741	0.257	1.221	0.975	0.017	0.004	0.004

Table IV. Expression averages (left) and posterior pattern probabilities (right) for several genes classified as having expression pattern 3 by the GG model (see Figure 4). For each gene, the probability vector from the GG model is in the first row and the one from the LNN model is in the second row.

Gene ID	Group				Expression pattern			
	Cop	CI	CII	WF	null	P1	P2	P3
J00801	3066.3	4777.0	995.3	9082.9	0.05	0	0	0.95
					0.04	0	0	0.96
L08100	4367.5	4002.6	1278.3	14162.3	0	0	0	1
					0	0	0	1
J00772	392.0	325.8	121.7	678.9	0.04	0	0	0.96
					0.97	0.01	0.00	0.02

only in pattern 3. Note that differences in genotype need not imply differences in expression. Genes classified into the null pattern show equivalent expression across groups, but differ in genotype. Patterns 1 and 2 also allow for distinct genotype and expression patterns. Parameter estimates for each model are given in Table III.

Under the GG model, 24,795 genes had posterior probability greater than 0.5 of being in the null pattern; 250, 86 and 111 genes were classified into patterns 1, 2 and 3, respectively. We did not classify six genes because for them no pattern had posterior probability greater than 0.5. The LNN model identified slightly more genes as differentially expressed. Specifically, 24,164 were classified into the null pattern; 447, 346 and 280 were classified into patterns 1, 2 and 3 and 11 were not classified. Identified under both methods were 24,119 (null), 217 (pattern 1), 51 (pattern 2) and 78 (pattern 3) genes.

Three genes identified as pattern 3 by the GG model are shown in Table IV. Two of these genes (J00801 and L08100) are known markers of mammary gland differentiation, and a common belief is that differentiation protects against tumour development. For each of these genes, the average intensity in the WF condition is higher than that observed in the COP or congenic lines. This indicates increased expression (and increased differentiation) in the WF, which is unexpected since the WF strain is tumour susceptible. It may be the case that not all forms of differentiation are associated with resistance. Preliminary data in other rat strains and other experiments are supporting this hypothesis (Gould, unpublished data). The third gene

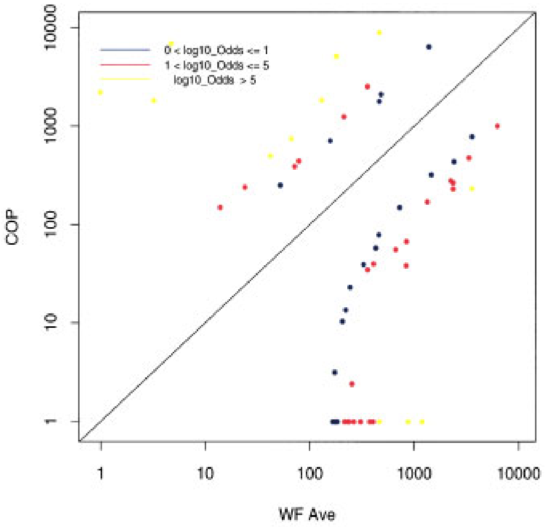

Figure 4. Average intensities across replicates for the WF and COP data. Only spots which exhibit significant differential expression (as determined by the GG model) are shown.

(J00772) is rat prostatein. Recent work suggests that this gene, normally associated with the ventral prostate, is strongly expressed in the stromal cells of the rat mammary gland (Watson and Gould, unpublished data). The GG calculations classify this gene as having elevated expression in WF, but the LNN calculations are equivocal, and consider it to be unchanged. Further study of this gene is warranted.

As a separate calculation, we analysed the data from the WF and COP parentals only, omitting the congenics. Table III contains parameter estimates. The odds calculation under GG assumptions estimates 58 genes to be differentially expressed. Of these, 57 are also identified by the LNN model. Figure 4 gives a graph of the average intensities (across replicates) for the spots identified as changed in the GG model. These results are consistent with the multiple group analysis. Each of the 58 genes identified as differentially expressed in the two-group analysis is also identified when comparing multiple groups; 48 of the 58 genes have posterior probability larger than 0.5 of being in pattern 1, four of the genes are in pattern 2, and five of the genes are in pattern 3. One gene was not classified. For both the four- and two-group analysis, Bayes factors indicated that the GG model fits better than the LLN model.

7. DISCUSSION

We have extended empirical Bayes methodology for gene expression data to account for replicate arrays, multiple conditions and a range of modelling assumptions. The general hierarchical mixture model proposed accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations. Since properties of individual experiments affect each of these features, the distributions governing them are to some extent experiment dependent. However, there are characteristics inherent to microarray data that are observed in many experiments, such as increasing variation with increasing mean and within-gene correlation. The present approach accounts for these properties and is also flexible in the sense that various choices can be made on the distributional forms at each level of the model.

Using a specified model and Bayes rule, posterior probabilities are obtained from which inferences regarding differential expression patterns can be made. The classification of genes as differentially expressed (DE) or equivalently expressed (EE) according to the state favoured by the posterior probabilities is an optimal procedure in the context of the mixture model. It minimizes the expected number of mistakes. Interestingly, this goal is different from the goal in classical testing which is to bound the type I error rate and then aim to maximize the power. To reduce type I errors we could make a more stringent decision rule and assume EE unless the odds favouring DE are much larger than 1:1. Preliminary results from a small simulation study indicated that this might not be necessary as the type I error rate is controlled. Additional operating characteristics evaluated in the simulation study were also found to be well controlled.

Under gamma-gamma model assumptions, the estimated positive predictive value was at least 94 per cent, regardless of the proportion p of differentially expressed genes. The negative predictive value decreased from 97 per cent to 80 per cent with increasing p. This indicates that although the method may be missing genes, a positive identification is most likely an accurate one. Estimates of the sensitivity and specificity reflect this as well. The sensitivity increased from 65 per cent to 80 per cent with increasing p, while the specificity was at or above 95 per cent for each value of p considered. The average false discovery rate (FDR) was near 0.05 for all values of p, indicating that control of the type I error rate is inherent to this empirical Bayes approach. Virtually identical numerical results were obtained under LNN assumptions. Furthermore, the results were similar when a bimodal distribution for the true underlying intensities was considered.

These results suggest that error rates are reasonably low and that particular modelling assumptions might have only a minimal impact on the accurate identification of differentially expressed genes. We note that such results are preliminary, and further work is required before any such conclusions can be made in general. Only two groups having a fixed number of replicates in each group were considered in our simulation study. The study could be extended to evaluate error rates in the case of multiple conditions for a varying number of replicates. Additional model forms should also be considered, both for data simulation and odds calculations. We are currently investigating the effects of non-parametric assumptions on the latent mean distribution $\pi(\mu_g)$ (see Newton [20]).

The proposed method assumes that intensity measurements approximate some true underlying expression level. Thus, expression profiles must be normalized in such a way so that any systematic sources of variation have been removed. DNA chip analyser [21] was used here, but many other methods are available. We also note that mRNA samples were pooled across subjects. Of course, under some conditions this can decrease measurement variability, thereby reducing the number of replicates required. However, owing to array specific effects, pooling does not eliminate the need for replication [22]. Both Kerr *et al.* [23] and Lee *et al.* [12] stress the importance of replication in microarray studies. In addition to array effects, if outliers (for example, contaminated samples) are present, pooling can lead to biased estimates of underlying expression. Optimal experimental designs which provide for maximum measurement accuracy using a minimum number of arrays have yet to be developed. This is an area that requires further investigation.

Finally, we note an interesting statistic which emerges from the odds of differential expression in the GG model (7) when comparing two conditions when the number of replicates in each group (n_1 and n_2, respectively) are large compared to α_0 and v. Let $\mathbf{x}_g = (x_{g,1}, x_{g,2}, \ldots, x_{g,n_1})$ denote the n_1 replicate measurements in the first condition and $\mathbf{y}_g = (y_{g,1}, y_{g,2}, \ldots, y_{g,n_2})$ denote the n_2 replicates in the second condition. Up to a power and a proportionality constant, the odds favouring DE are

$$\frac{(\bar{x}_g + \bar{y}_g)/2}{\sqrt{(x_g y_g)}}$$

where $\bar{x}_g$ and $\bar{y}_g$ are respective sample means (on the raw scale) of expression measurements in the two groups. The odds are related to the ratio of the arithmetic to the geometric mean of the sample means. Considering the arithmetic-geometric mean inequality, this seems to be an interesting measure of the difference between the two samples. A similar analysis of the LNN odds shows that one is related to the more familiar difference $\bar{\bar{x}}_g - \bar{\bar{y}}_g$, that is, the difference between the arithmetic means of the log-transformed responses (a t-like statistic). We think these facts give some credence to the model-based formulation and also suggest directions that the models could be extended.

APPENDIX: ESTIMATION IN THE MULTIPLE GROUP CASE

With data $\mathbf{d}_g$ governed by a mixture of the form (1), we introduce missing pattern indicators $z_{g,l}$ defined as one if the expression pattern of gene g is pattern l and zero otherwise. The *complete* data log-likelihood is

$$l_c(\theta) = \sum_g \left\{ \sum_{k=0}^{m} z_{g,k} [\log f_k(\mathbf{d}_g) + \log(p_k)] \right\}$$

For θ fixed at θ_0, calculation of the expectation conditional on the observed data and θ_0 (E-step) gives

$$\hat{l}_c(\theta) = \sum_g \left\{ \sum_{k=0}^{m} \hat{z}_{g,k} [\log f_k(\mathbf{d}_g) + \log(p_k)] \right\}$$

$\hat{z}_{g,l}$ is the posterior probability of expression pattern l for gene g

$$P(l|\mathbf{d}_g) = \frac{p_l f_l(\mathbf{d}_g)}{\sum_{k=0}^{m} p_k f_k(\mathbf{d}_g)}$$

where θ_0 parameterizes the densities f_k. We use the arithmetic mean of $\hat{z}_{g,k}$ to estimate p_k; *nlminb* in S-plus provides estimates of θ (M-step). This process is repeated until there is convergence in the estimates. Results are checked from various starting configurations.

ACKNOWLEDGEMENTS

The authors thank two referees for critical reviews of an earlier draft. Code and data used in this article are available at the first author's web site www.biostat.wisc.edu/˜kendzior/. This research was funded in part by training grant TA-CA 09565 for CMK, the NIH R01 grant CA64364 to MAN, and CA28954 and CA77494 to MNG.

REFERENCES

1. West M, Nevins JR, Marks JR, Spang R, Blanchette C, Zuzan H. DNA microarray data analysis and regression modeling for genetic expression profiling. Institute of Statistics and Decision Sciences, Working Paper #15, 2000.
2. West M. Bayesian regression analysis in the 'large p. small n' paradigm. Institute of Statistics and Decision Sciences, Working Paper #22, 2000.
3. Efron B, Morris C. Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1973; **35**:379–421.
4. Efron B, Morris C. Stein's paradox in statistics. *Scientific American* 1977; **236**:119–127.
5. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 2001; **8**:37–52.
6. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002; **12**:111–139.
7. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; **98**:5116–5121.
8. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* 2001; **17**(6):509–519.
9. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 1997; **2**(4):364–374.
10. Efron B, Tibshirani R, Goss V, Chu G. Microarrays and their use in a comparative experiment. Technical Report 37B/213, Stanford University Department of Statistics, 2000.
11. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**(456):1151–1160.
12. Lee MLT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences* 2000; **97**(18):9834–9839.
13. Langston AA, Malone KE, Thompson JD, Daling JR, Ostrander EA. BRCA1 mutations in a population-based sample of young women with breast cancer. *New England Journal of Medicine* 1996; **334**:137–142.
14. Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, *et al.* The risk of cancer associated with specific mutations of BRCAl and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine* 1997; **336**:1404–1408.
15. Shepel LA, Lan H, Haag J, Brasic GM, Gheen ME, Simon JS, Hoff P, Newton MA, Gould MN. Genetic identification of multiple loci that control breast cancer susceptibility in the rat. *Genetics* 1998; **149**:289–299.
16. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall: New York, 1996.
17. STATISTICAL SCIENCES. *S-PLUS Guide to Statistical and Mathematical Analysis*, *Version 3.2* StatSci, a division of MathSoft, Inc.: Seattle, 1993.
18. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.
19. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**(430):773–795.
20. Newton MA. On a nonparametric recursive estimator of the mixing distribution. *Sankhya* A 2002; **64**:1–17.
21. Li C, Wong W. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* 2001; **98**(1):31–36.
22. Kendziorski CM, Zhang Y, Lan H, Attie A. The efficiency of MRNA pooling in microarray experiments. *Biostatistics* 2003; **4**:465–477.
23. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 2000; **7**:819–837.
24. Lee MLT, Weining L, Whitmore GA, Beier D. Models for microarray gene expression data. *Proceedings of the ASA Joint Meetings*, *Atlanta*, *GA*, 2001.