CSIRO PUBLISHING
www.publish.csiro.au/journals/ajea *Australian Journal of Experimental Agriculture*, 2005, **45**, 859–866

Using mixture models to detect differentially expressed genes

G. J. McLachlan[A,B,C,D], *R. W. Bean*[B], *L. Ben-Tovim Jones*[B] *and J. X. Zhu*[B]

[A]Department of Mathematics, University of Queensland, Qld 4072, Australia.
[B]ARC Centre in Bioinformatics, Institute for Molecular Bioscience,
University of Queensland, Qld 4072, Australia.
[C]ARC Special Research Centre for Functional and Applied Genomics,
University of Queensland, Qld 4072, Australia.
[D]Corresponding author. Email: gjm@maths.uq.edu.au

Abstract. An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. In this paper, we focus on the use of mixture models to handle the multiplicity issue. With this approach, a measure of the local false discovery rate is provided for each gene, and it can be implemented so that the implied global false discovery rate is bounded as with the Benjamini-Hochberg methodology based on tail areas. The latter procedure is too conservative, unless it is modified according to the prior probability that a gene is not differentially expressed. An attractive feature of the mixture model approach is that it provides a framework for the estimation of this probability and its subsequent use in forming a decision rule. The rule can also be formed to take the false negative rate into account.

Additional keywords: multiple hypothesis testing, false discovery rate, Bayes formula, Bayes rule.

Introduction

In recent times, there has been an explosion in the development of comprehensive, high-throughput methods for molecular biology experimentation. Deoxyribonucleic acid microarray technologies, such as the Affymetrix GeneChip or spotted (cDNA or oligonucleotide) arrays, provide a means for measuring tens of thousands of genes simultaneously. For a history of microarray development refer to McLachlan *et al*. (2004). An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes $C_1,\ldots,C_g$. The classes may correspond to tissues (cells) that are at different stages in some process, in distinct pathological states, or under different experimental conditions. By comparing gene expression profiles across these classes, researchers gain insight into the roles and reactions of various genes. A comparison can be drawn, for example, between healthy cells and cancerous cells within subjects in order to learn which genes tend to be over or underexpressed in the diseased cells. Regulation of such genes could produce effective cancer treatment and/or prophylaxis. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. In this paper, we focus on the use of mixture models to handle the multiplicity issue.

There is now an extensive range of literature on the problem of detecting differentially expressed genes in microarray data. Dudoit and Fridlyand (2002) conducted one of the first studies to recognise the importance of the multiplicity problem as one of the key statistical issues in microarray data analysis. An excellent review of this problem has been given recently by Dudoit *et al*. (2003).

The simplest method for identifying differentially expressed genes is to evaluate the log ratio between 2 classes (or averages of the log ratios when there are replicates) and consider all genes that differ by more than an arbitrary cutoff value to be differentially expressed (Schena *et al*. 1996; DeRisi *et al*. 1997). This test, sometimes called a fold change, is not a statistical test, and there is no associated level of confidence in the designation of a gene as being differentially expressed or not differentially expressed (Cui and Churchill 2003). Also, this method ignores the variance of the replicates in each class.

Statistical significance of the differential expressions can be tested by performing a test for each gene. When many hypotheses are tested, the probability that a type I error (a false positive error) is committed increases sharply with the number of hypotheses. This multiplicity problem is not unique to microarray analysis, but its magnitude where each experiment may involve many thousands of genes

 10.1071/EA05051 0816-1089/05/080859

dramatically intensifies the problem. We shall consider this further, but first we need to introduce some notation.

Notation

Although biological experiments vary considerably in their design, the data generated by microarray experiments can be viewed as a matrix of expression levels. For M microarray experiments (corresponding to M tissue samples), where we measure the expression levels of N genes in each experiment, the results can be represented by an $N \times M$ matrix. Typically, M is no more than 100 (usually much less in the present context), while the number of genes N is of the order of 10^4. The expression signature is the expression levels of the N genes for each tissue. Conversely, the expression profile is the expression levels across the different tissue samples for each gene. The M tissue samples might correspond to each of M different patients or, say, to samples from a single patient taken at M different time points. The expression levels are taken to be the measured (absolute) intensities for the Affymetrix platform, and the ratios of the intensities for the Cy5-channel (red) images and Cy3-channel (green) images for spotted (cDNA or oligonucleotide) arrays (e.g. Dudoit and Fridlyand 2002). The M tissue samples on the N available genes are classified with respect to g classes or conditions. We let y_{ijk} denote the (logged) gene expression level for the kth replicate of the jth gene in the ith class ($k = 1,\ldots,n_j$; $j = 1,\ldots,N$; $i = 1,\ldots,g$), where $M = n_1 + \ldots + n_g$ denotes the total number of microarrays. It is assumed that these expression levels have been preprocessed with adjustment for array effects. For simplicity, we shall take $g = 2$, but the methodology to be described applies in the case of multiple classes.

For gene j, we let $H_j = 0$ denote that the null hypothesis of no association between its expression level and its class membership holds, and we let $H_j = 1$ if it does not hold ($j = 1,\ldots,N$).

Test of a single hypothesis

A commonly used statistic for testing for a difference in the means of 2 classes is the well-known Student's t-statistic defined by:

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{s_j\sqrt{1/n_1 + 1/n_2}} \qquad (1)$$

where $\bar{y}_{ij}$ and s_{ij}^2 denote the sample mean and variance of the n_i replicates y_{ijk} ($k = 1,\ldots,n_i$)for the jth gene in the ith class C_i and $s_j^2 = \{(n_1 - 1)s_{1j}^2 + (n_2 - 1)s_{2j}^2\}/(M - 2)$ is the pooled within-class sample variance.

Because of the large number of genes in the microarray experiments, there will always be some genes with a very small sum of squares across replicates, so that their (absolute) t-values will be very large whether or not their averages are large. Tusher *et al*. (2001) have proposed a refinement that avoids this difficulty.

They used the modified t-statistic by adding a constant to the denominator of equation 1 to give:

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{s_j\sqrt{1/n_1 + 1/n_2} + a_0} \qquad (2)$$

The constant a_0 was chosen to make the coefficient of variation of t_j about constant as a function of s_j. This has the added effect of dampening values of t_j that arise from genes whose expression is near to zero.

In testing a single hypothesis on the jth gene, 2 types of errors can be committed: reject the null hypothesis when it holds (a type I error), or retain the null hypothesis when it does not hold (a type II error). With the traditional approach to the test of a single hypothesis, the aim is to maximise the power (1 minus the probability of making a type II error), while keeping the probability of a type I error at or below a specified level.

We let W_j denote a test statistic, such as the square of the t-statistic (equation 1), for testing the null hypothesis that the jth gene is not differentially expressed, where the null is to be rejected for sufficiently large (positive) values. In the case of an arbitrary number g of classes, we might take W_j to be the usual one-way analysis of variance F-statistic, which reduces to the square of the t-statistic in the case of $g = 2$.

An advantage of working with this F-statistic is that it can be easily transformed so that its null distribution is well approximated by the standard normal distribution.

A gene-specific summary is given by the observed value W_j of the test statistic W_j or the associated P-value, p_j, which can be expressed as $p_j = \text{pr}\{W_j \geq w_j \mid H_j = 0\}$.

Multiple hypothesis testing

The aim is to detect whether the expression levels of some of the thousands of genes are different in class C_1 than in class C_2. In the context of statistical inference, we can formulate the problem as a multiple hypothesis testing problem. When many hypotheses are tested, the probability that a type I error (a false positive error) is committed increases sharply with the number of hypotheses. In practice, the number of genes N can be very large. Thus if we were to carry out separate tests in the case of $N = 6000$ genes, the number of false positives could be quite large. For instance, if all $N = 6000$ genes were not differentially expressed, then the expected number of false positives would be 300. Thus there is a need to control the false positive rate.

The Bonferroni method is perhaps the best known method for dealing with multiple testing. It controls the family-wise error rate (FWER), which is the probability that at least 1 false positive error will be committed. The test of each null hypothesis is controlled so that the probability of a Type I error is less than or equal to a/N for some α. This ensures that the overall FWER is less than or equal to α. But control of the FWER is only appropriate in situations where the intent is to identify only a small number of genes that are truly

different. Otherwise, the severe loss of power in controlling the FWER is not justified (Reiner *et al*. 2003). Instead, it is more appropriate to emphasise the proportion of false positives among the identified differentially expressed genes. The expectation of this proportion is essentially the false discovery rate (FDR) of Benjamini and Hochberg (1995). The FDR is defined formally as:

$$\mathrm{FDR} = E\left[\frac{N_{01}}{N_r \vee 1}\right] \tag{3}$$

where E refers to the expectation operator and $N_r \vee 1 = \max(N_r,1)$. Here N_r is the number of rejected hypotheses and N_{01} is the number of false positives among them. The positive FDR is equal to the FDR divided by the probability that N_r is greater than zero (Storey 2002).

The false non-discovery rate (FNR) is given by:

$$\mathrm{FNR} = E\left[\frac{N_{10}}{(N - N_r) \vee 1}\right] \tag{4}$$

where N_{10} is the number of false negatives.

Benjamini-Hochberg procedure

Benjamini and Hochberg (1995) proved by induction that the following procedure (referred to here as the BH procedure) controls the FDR at level α when the P-values following the null distribution are independent and uniformly distributed. The BP procedure is as follows:

Let $p_{(1)} \le \ldots \le p_{(N)}$ be the observed P-values.

Calculate

$$\hat{k} = \arg\max_{1 \le k \le N} \left\{k : p_{(k)} < \alpha k / N\right\}. \tag{5}$$

If $\hat{k}$ exists, then reject null hypotheses corresponding to $p_{(1)} \le \ldots \le p_{(\hat{k})}$. Otherwise, reject nothing.

Benjamini and Yekutieli (2001) showed that $\mathrm{FDR} \le \alpha N_0/N$ for positively dependent test statistics as well. Since the BH procedure controls the FDR at a level too low by a factor of N_0/N, it is natural to try to estimate N_0 and use $\alpha^* = \alpha(N/N_0)$ instead of α to gain more power.

Example of Bonferroni and BH tests

Suppose that 10 independent hypothesis tests are carried out leading to the following ordered P-values: 0.0001, 0.0035, 0.0069, 0.0083, 0.019, 0.3463, 0.3964, 0.5488, 0.6124, 0.9871. With $\alpha = 0.05$, the Bonferroni test rejects any hypothesis whose P-value is less than $\alpha/10 = 0.005$. Thus only the first 2 hypotheses are rejected. For the BH test, we find the largest k such that $P(k) < k\alpha/N$. Here $\hat{k} = 5$, and so we reject the first 5 hypotheses.

Recently, a number of key papers have been written on controlling the FDR (Genovese and Wasserman 2002: Storey 2002; Storey and Tibshirani 2003*a*, 2003*b*; Black 2004; Cox and Wong 2004; Storey *et al*. 2004). Hence methods for the detection of differentially expressed genes are still evolving.

Two-component mixture model framework

Definition of model

In this paper, we focus on a decision–theoretic approach to the problem of finding genes that are differentially expressed. We use a prediction rule approach based on a 2-component mixture model as formulated by Lee *et al*. (2000) and Efron *et al*. (2001). We let G denote the population of genes under consideration. It can be decomposed into G_0 and G_1, where G_0 is the population of genes that are not differentially expressed, and G_1 is the complement of G_0 that is, G_1 contains the genes that are differentially expressed.

We let the random variable Z_{ij} be defined to be 1 or 0 according as the jth gene belongs to G_i or not (i = 0,1; j = 1,…N). We have defined H_j to be 0 or 1 according as to whether the null hypothesis of no differential expression does or does not hold for the jth gene. Thus Z_{1j} is 0 or 1 according as to whether H_j is 0 or 1.

The prior probability that the ith gene belongs to G_0 is assumed to be π_0 for all j. That is, $\pi_0 = \mathrm{pr}\{H_j = 0\}$ and $\pi_1 = \mathrm{pr}\{H_j = 1\}$. Assuming that the test statistics W_j all have the same distribution in G_i, we let $f_i(w_j)$ denote the density of W_j in G_i(i = 1,2). The unconditional density $f(w_j)$ of W_j is given by the 2-component mixture model:

$$f(w_j) = \pi_0 f_0(w_j) + \pi_1 f_1(w_j). \tag{6}$$

Using Bayes Theorem, the posterior probability that the jth gene is not differentially expressed (that is, belongs to G_0) is given by:

$$\tau_0(w_j) = \pi_0 f_0(w_j)/f(w_j) \ (j = 1,\ldots N). \tag{7}$$

In this framework, the gene-specific posterior probabilities $\tau_0(w_j)$ provide the basis for optimal statistical inference about differential expression.

This approach is Bayesian in that it uses Bayes theorem, but it is not Bayesian in the estimation process. That is, we shall not make any prior assumptions about the mixing parameter π_0, nor about the parameters in the forms that we shall adopt for the densities $f_0(w_j)$ and $f(w_j)$.

The posterior probability $\tau_0(w_j)$ has been termed the local false discovery rate (local FDR) by Efron and Tibshirani (2002). As noted by Efron (2004), it can be viewed as an empirical Bayes version of BH methodology, using densities rather than tail areas.

Bayes decision rule

Let e_{01} and e_{10} denote the 2 errors when a rule is used to assign a gene as being differentially expressed or not, where e_{01} is the probability of a false positive and e_{10} is the probability of a false negative. Then, the risk is given by:

$$\mathrm{Risk} = (1 - c)\pi_0 e_{01} + c\pi_1 e_{10} \tag{8}$$

where $(1 - c)$ is the cost of a false positive. As the risk depends only on the ratio of the costs of misallocation, they have been scaled to add to 1 without loss of generality.

The Bayes rule, which is the rule that minimises the risk (equation 8), assigns a gene to G_1 if

$$\tau_0(w_j) \le c; \tag{9}$$

otherwise, the jth gene is assigned to G_0. In the case of equal costs of misallocation ($c = 0.5$), the cutoff point for the posterior probability $\tau_0(w_j)$ in equation 9 reduces to 0.5.

Estimated FDR

In practice, we do not know π_0, nor the density $f(w_j)$, and perhaps not $f_0(w_j)$. In some instances, the latter may be known as we may have chosen our test statistic so that its null distribution is known (or known to a good approximation). For example, we shall work with the one-way analysis of variance F-statistic, which can be so transformed that its null distribution is about the standard normal. In some situations, we might wish to estimate the null density. Efron (2004) has demonstrated that the theoretical null distribution in microarray experiments may not always be appropriate due to experimental noise.

Alternatively, null replications of the test statistic might be created, by the bootstrap or permutation methods. We shall estimate the population density $f(w)$ by maximum likelihood after its formulation using a mixture model. But it can be estimated also nonparametrically by its empirical distribution based on the observed test statistics w_j.

If $\hat{\pi}_0$, $\hat{f}_0(w_j)$ and $\hat{f}_1(w_j)$ denote estimates of π_0, $f_0(w_j)$ and $f_1(w_j)$, respectively, the gene-specific summaries of differential expression can be expressed in terms of the estimated posterior probabilities $\hat{\tau}_0(w_j)$, where:

$$\hat{\tau}_0(w_j) = \hat{\pi}_0 \hat{f}_0(w_j)/\hat{f}(w_j) \; (j = 1,\ldots,N) \tag{10}$$

is the estimated posterior probability that the jth gene is not differentially expressed. An optimal ranking of the genes can therefore be obtained by ranking the genes according to the $\hat{\tau}_0(w_j)$ ranked from smallest to largest. A short list of genes can be obtained by including all genes with $\hat{\tau}_0(w_j)$ less than some threshold c_0 or by taking the top N_0 genes in the ranked list.

Suppose that we select all genes with $\hat{\tau}_0(w_j) \le c_0$, then an estimate of the FDR rate is given by:

$$\hat{\mathrm{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j) I_{[0,c_0]}(\hat{\tau}_0(w_j))/N_r \tag{11}$$

where

$$N_r = \sum_{j=1}^{N} I_{[0,c_0]}(\hat{\tau}(w_j)) \tag{12}$$

is the number of the selected genes in the list. Here $I_A(w)$ is the indicator function that is 1 if w belongs to the interval A and is zero otherwise.

Thus we can find a data-dependent $c_0 \le 1$ as large as possible such that $\hat{\mathrm{FDR}} \le \alpha$. This assumes that there will be some genes with $\hat{\tau}_0(w_j) \le \alpha$ which will be true in the typical situation in practice. This bound is approximate due to the use of estimates in forming the posterior probabilities of nondifferential expression and so it depends on the fit of the densities $f_0(w_j)$ and $f(w_j)$.

Bayes risk in terms of estimated FDR and FNR

The Bayes prediction rule minimises the risk of an allocation defined by equation 8. We can estimate the error of a false positive e_{01} and the error of a false negative e_{10} by:

$$\hat{e}_{01} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} / \sum_{j=1}^{N} \hat{\tau}_0(w_j) \tag{13}$$

and

$$\hat{e}_{10} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} / \sum_{j=1}^{N} \hat{\tau}_1(w_j) \tag{14}$$

respectively, where $\hat{z}_{0j}$ is taken to be 0 or 1 according as to whether $\hat{\tau}_0(w_j)$ is less than or greater than c in equation 9, and $\hat{z}_{1j} = 1 - \hat{z}_{0j}$. Also, we can estimate the prior probability $\hat{\pi}_0$ as:

$$\pi_0 = \sum_{j=1}^{N} \hat{\tau}_0(w_j)/N \tag{15}$$

On substituting these estimates in equations 13–15 into the right-hand side of equation 9, the estimated risk can be written as:

$$\hat{\mathrm{Risk}} = (1-c)\hat{\omega}\hat{\mathrm{FDR}} + c(1-\hat{\omega})\hat{\mathrm{FNR}} \tag{16}$$

where

$$\hat{\mathrm{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} / \sum_{j=1}^{N} \hat{z}_{1j} \tag{17}$$

and

$$\hat{\mathrm{FNR}} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} / \sum_{j=1}^{N} \hat{z}_{0j} \tag{18}$$

are estimates of the FDR and FNR, respectively, and where

$$\hat{\omega} = \sum_{j=1}^{N} \hat{z}_{1j}/N = N_r/N \tag{19}$$

is an estimate of the probability that a gene is selected.

Thus unlike the tests or rules that are designed to control just the FDR, the Bayes rule approach in its selection of the genes can be viewed as controlling a linear combination of the FDR and FNR. The balance between the FDR and the FNR is controlled by the threshold c. An early reference on the Bayes rule in the context of hypothesis testing can be found in Lehmann (1959).

Estimation of posterior probabilities

Previous work

In previous work on this problem, Efron *et al*. (2001) adopted an empirical Bayes approach without any assumptions being made. The quantities $f_0(w_j)$ and $f(w_j)$

were estimated using the empirical distributions for the $w_{0j}^{(b)}$ and the w_j, where $w_{0j}^{(b)}$ denotes the value of w_j obtained on the bth random permutation of the class labels. In order to estimate π_0, which is not estimable in a parametric setting, they effected this using the inequality:

$$\pi_0 \le \min_w \{f(w)/f_0(w)\}. \quad (20)$$

Do *et al*. (2003) proposed an extension to the nonparametric approach of Efron *et al*. (2001) by adopting a fully model-based approach. While Efron's method (Efron *et al*. 2001) proceeds by plugging in point estimates, the fully model-based approach of Do *et al*. (2003) constructs a probability model for the unknown mixture, allowing investigators to deduce the desired inference about differential expression as posterior inference in that probability model. Dirichlet process mixture models are chosen to represent the probability model for the unknown distributions. Markov chain Monte Carlo (MCMC) posterior simulation was developed to generate samples from the relevant posterior and posterior predictive distributions. Newton *et al.* (2001), Kendziorski *et al*. (2003), Newton and Kendziorski (2003) and Newton *et al*. (2004) have adopted parametric empirical Bayes approaches to the problem of the detection of differential expression.

Previously, Pan (2002, 2003) and Zhao and Pan (2003) considered a nonparametric approach, which they called the mixture model method (MMM). They advocated modelling the densities $f_0(w_j)$ and $f(w_j)$ in the 2-component mixture model by normal mixtures. With this mixture model method approach, the likelihood ratio test statistic, $\lambda(W_j) = f_0(W_j)/f(W_j)$, can be used to test the null hypothesis that the jth gene is not differentially expressed.

Mixture model approach

We choose our test statistic W_j so that it has a normal distribution under the null hypothesis that the jth gene is not differentially expressed. For example, if F_j denotes the usual test statistic in a one-way analysis of variance with g classes, then we follow Broët *et al*. (2004) and transform the F_j statistic as:

$$W_j = \frac{(1 - \frac{2}{9(M-g)})F_j^{\frac{1}{3}} - (1 - \frac{2}{9(g-1)})}{\sqrt{\frac{2}{9(M-g)}F_j^{\frac{2}{3}} + \frac{2}{9(g-1)}}} \quad (21)$$

The distribution of the transformed statistic W_j is about a standard normal under the null hypothesis that the jth gene is not differentially expressed (that is, given its membership of population G_0). As noted in Broët *et al*. (2004), it is remarkably accurate for $(M - g) \ge 10$ (Johnson and Kotz 1970).

With this transformation, we can take the null density $f_0(w_j)$ to be the standard normal density (which has mean of 0 and unit variance). In order to estimate the mixing proportion π_0 and the mixture density $f(w_j)$, we postulate it to have the h-component normal mixture form:

$$f(w_j) = \sum_{i=0}^{h-1} \pi_i \phi(w_j; \mu_i, \sigma_i^2) \quad (22)$$

where we specify $\mu_0 = 0$ and $\sigma_0^2 = 1$. In equation 22, $\phi(w_j;\mu_i,\sigma_i^2)$ denotes the normal density with mean μ_i and unit variance σ_i^2. We suggest starting with $h = 2$, adding more components if considered necessary as judged using the Bayesian Information criterion (BIC).

Use of P-values

An alternative to working with the test statistic W_j, we could follow the approach of Allison *et al*. (2002) and use the associated P-value p_j. We can find these P-values using permutation methods whereby we permute the class labels. Using just the B permutations of the class labels for the gene-specific statistic W_j, the P-value for $W_j = w_j$ is assessed as

$$p_j = \frac{\#\{b : w_{0j}^{(b)} \ge w_j\}}{B} \quad (23)$$

where $w_{0j}^{(b)}$ is the null version of w_j after the bth permutation of the class labels. The distribution of p_j has support on the unit interval, and so its distribution can be represented by a mixture of β distributions of the first kind (Diaconis and Ylvisaker 1985). Under the null hypothesis of no differential expression for the jth gene, p_j will have a uniform distribution on the unit interval; that is the $\beta_{1,1}$ distribution.

The $\beta_{\alpha_1,\alpha_2}$ density is given by

$$f(u;\alpha_1,\alpha_2) = \{u^{\alpha_1-1}(1-u)^{\alpha_2-1}\}/B(\alpha_1,\alpha_2)I_{(0,1)}(u) \quad (24)$$

where

$$B(\alpha_1,\alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1+\alpha_2). \quad (25)$$

Allison *et al*. (2002) discusses the fitting of mixtures of $\beta_{\alpha_1,\alpha_2}$ components to the values of p_j for the N genes, including the caution that needs to be exercised in interpreting the existence of modes in the fitted mixture density as a consequence of the correlation between some of the p_j values. If the null distribution of W_j is calculated just on the data for the jth gene, as in the formation of p_j (equation 23), it suffers from a granularity problem. For example, there are only 10 ways to divide 6 microarrays into 2 equal sized groups. The null distribution has a resolution on the order of the number of permutations. If we perform B permutations, then the P-value will be estimated with a resolution of $1/B$. If we assume that each gene has the same null distribution and combine the permutations, then the resolution will be $1/(NB)$ for the pooled null distribution. Using the latter, the P-value for the jth gene can be estimated by:

$$p_j = \sum_{b=1}^{B} \#\{i: w_{0i}^{(b)} \geq w_j\ , i = 1,\ldots,N\}/NB \qquad (26)$$

where $w_{0i}^{(b)}$ is the value of w_j obtained on the bth permutation of the class labels ($i = 1,\ldots,N$).

The drawback of pooling the null statistics $w_{0i}^{(b)}$ across the genes to assess the null distribution of W_j is that the distributions used will be different, unless all H_j are zero. To illustrate this, we performed $B = 1000$ permutations of the class labels for the pooled 2-sample t-statistic. This was based on 2 samples consisting of $n_1 = 7$ and $n_2 = 8$ observations, each simulated from the standard normal distribution. It can be seen in Figure 1 that the histogram of the 1000 values of the t-statistic is close to its true null density given by the t-distribution with 13 degrees of freedom. To consider a case where the null hypothesis of no differential expression does not hold, we modified the simulated data so that the second sample of $n_2 = 8$ observations now had a mean of 10 and a variance of 9. The histogram of the 1000 values of the t-statistic obtained by permuting the class labels is given in Figure 2. On comparing these values with the true null density, it can be seen that they are more dispersed now as a consequence that the null hypothesis does not hold. Efron and Tibshirani (2002) suggest that the effect of this spurious component of variance be lessened by using only balanced permutations. X. Guo and W. Pan (unpublished data) suggest using a weighted permutation method to dampen the effect of permuted samples corresponding to genes that are judged to be differentially expressed.

Link with BH methodology

In this section, we consider the link of the approach based on the Bayes rule equation 9 with the tail-area methodology of Benjamini and Hochberg (1995). We shall first look at the link of the BH methodology with the posterior probability that the jth gene is not differentially expressed conditional on tail areas rather than on actual values as with the use of $\tau_0(w_j)$.

Using Bayes theorem, we have that:

$$\text{pr}\{H_j = 0 \,|\, w_j \geq w\} = \pi_0 \frac{1 - F_0(w)}{1 - F(w)} \qquad (27)$$

If we transpose the events on the left-hand side of equation 27, we have the usual definition for the P-value of the test that rejects the null hypothesis for sufficiently large w_j.

Suppose that we declare the jth gene to be differentially expressed if w_j is greater than w_o, where w_o is defined to be the minimum value of w such that the right-hand side of equation 27 is equal to α; that is:

$$w_o = \min\{w: \pi_0 \frac{1 - F_0(w)}{1 - F(w)} = \alpha\} \qquad (28)$$

Then the FDR of this rule is bounded by α (Efron *et al*. 2001; Genovese and Wasserman 2002; Wit and McClure 2004). It also has an interpretation in terms of the q-value of Storey (2003). Suppose that the right-hand side of equation 27 is monotonic (decreasing) in w. Then, as shown explicitly in Wit and McClure (2004), if we set π_0 equal to 1 and estimate $F(w)$ by its empirical distribution in the right-hand side of equation 28, the consequent rule is equivalent to the BH procedure.

Concerning the link with the rule (equation 9) based on $\tau_0(w_j)$ which uses densities rather than tail areas, it can be noted that the right-hand side of equation 27 is also the conditional expectation of $\tau_0(W_j)$ given that $W_j \geq w$ (Efron 2004). Further, if $\tau_0(w)$ is monotonic (decreasing in w), it is equivalent to declaring the jth gene to be differentially expressed if:

$$w \geq w_o \qquad (29)$$

where w_o is the value of w such that:

$$\tau_0(w) = c_o. \qquad (30)$$

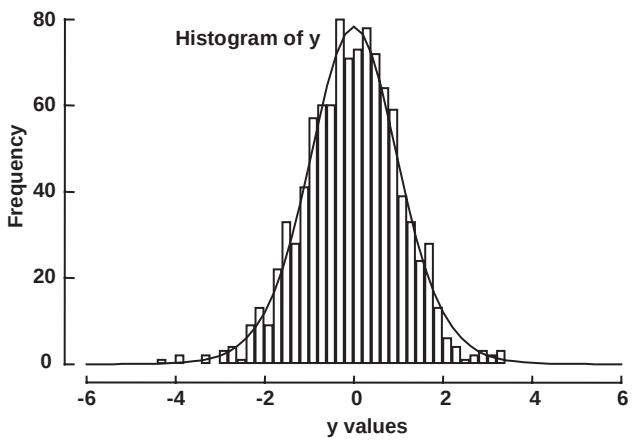

Figure 1. Histogram of the pooled 2-sample t-statistic under 1000 permutations of the class labels with t_{13} density superimposed. An example of a null case: with 7 N(0,1) points and 8 N(0,1) points.

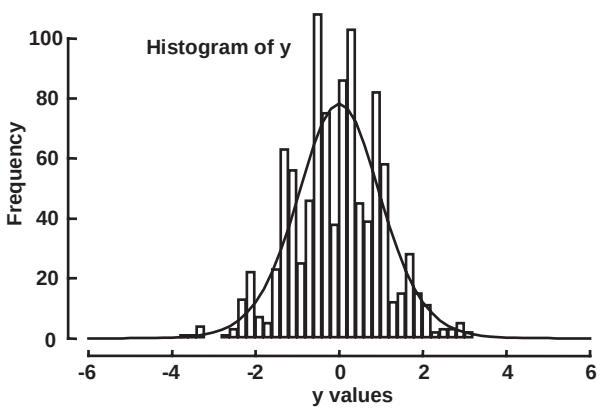

Figure 2. Histogram of the pooled 2-sample t-statistic under 1000 permutations of the class labels with t_{13} density superimposed. An example of a non-null case: with 7 N(0,1) points and 8 N(10,9) points.

Then its positive FDR is given by:

$$\pi_0 \frac{1 - F_0(w_o)}{1 - F(w_o)} \quad (31)$$

and so it, and hence its FDR, will not exceed α if w_o is chosen according to equation 28.

Thus if we were to use the modified form of the BH procedure with α replaced by α/π_0, then the Bayes rule approach is equivalent to the tail-area-based BH procedure. Of course in practice, we do not know π_0, the proportion of genes that are not differentially expressed. We have seen that the mixture model equation 6 provides a framework in which to estimate π_0. Also, it provides an estimate of the local FDR for each gene, namely the posterior probability $\tau_0(w_j)$ of nondifferential expression for the *j*th gene.

The estimate (eqn 17) of the implied (global) FDR of the Bayes rule-based approach can be viewed as being semi-parametric. In the case where $\tau_0(w)$ is monotonic in w, we can construct a fully parametric version by using the equation 31. Under the use of the standard normal for the null density and the formulation of the density as the normal mixture model, we can take $F_0(w_o) = \Phi(w_o)$ and:

$$\hat{F}(w_o) = \pi_0 \Phi(w_o) + \sum_{i=1}^{h-1} \hat{\pi}_i \Phi\left[\frac{w_o - \hat{\mu}_i}{\hat{\sigma}_i}\right] \quad (32)$$

in the right-hand side of equation 31. In our experience, this estimate is very similar to the semiparametric equation, as in the following example. In equation 32, $\hat{\mu}_i$ and $\hat{\sigma}_i$ denote the fitted mean and standard deviation of the *i*th normal component in the mixture model.

Example of the decision-theoretic approach

We consider the study of Hedenfalk *et al*. (2003), which consisted of $n_1 = 7$ BRCA1 arrays and $n_2 = 8$ BRCA2 arrays, along with some arrays from sporadic breast cancer. One goal of the study by Hedenfalk *et al*. (2003) was to find genes that are differentially expressed between BRCA1 and BRCA2-mutation-positive tumours by obtaining several microarrays from each cell type. In their analysis they computed a modified *F*-statistic and used it to assign a *P*-value to each gene. A threshold of $\alpha = 0.001$ was selected to find 51 genes from a total of $N = 3226$ that show differential gene expression. These authors subsequently used a threshold of $\alpha = 0.0001$ and they concluded that 9–11 genes are differentially expressed.

We applied our decision-theoretic approach to this dataset. In Table 1, we report the estimated values of the FDR, calculated using equation 18, for various levels of the threshold c_0.

It can be seen that if we were to declare the *j*th gene to be differentially expressed if $\tau_0(w_j) \leq 0.1$, then 175 genes would be selected as being significant, with an estimated FDR equal to 0.06. The prior probability π_0 of a gene not being differentially expressed was initialised to be 0.48. The estimates of the FDR in Table 1 are based on the semiparametric version (equation 17). We found that they were the same (to the second decimal place) as those calculated using the fully parametric estimate based on equation 31.

Table 1. Estimated false discovery rate ($\hat{\text{FDR}}$) for various levels of c_0

c_0^{A}	N_r^{B}	$\hat{\text{FDR}}$
0.5	1702	0.29
0.4	1235	0.23
0.3	850	0.18
0.2	483	0.12
0.1	175	0.06

[A]The cost threshold.
[B]The number of rejected hypotheses.

Of the 175 genes we found to be significant, 137 are overexpressed in BRCA1 tumours relative to BRCA2. Hedenfalk *et al*. (2003), and Storey and Tibshirani (2003*b*) in their further analysis of this dataset, found that a large block of genes are overexpressed in BRCA1. In particular, these included genes involved in DNA repair and cell death, such as MSH2 (DNA repair) and PDCD5 (induction of apoptosis), also identified by us.

For this dataset, Efron (2004) noted that the theoretical null distribution appears to be somewhat different from the empirical null. Hence it may not be wise to trust the theoretical null here, and in future work we will consider the estimation of the null component in the normal mixture model (equation 32).

Conclusion

In summary, we feel that a mixture model-based approach towards finding differentially expressed genes in microarray data can provide useful information beyond that of other methods. It gives a measure of the posterior probability that a specific gene is not differentially expressed (a local FDR). Standard methods for multiple hypothesis testing tend to focus on measures on the global FDR. The mixture model-based approach can also be used in the spirit of the *q*-value. A threshold can be obtained for the posterior probability of non-differential expression to ensure that the FDR is bounded at some desired level if all genes below the threshold are declared to be differentially expressed.

References

Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee C-K, Prolla TA, Weindruch R (2002) A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20. doi:10.1016/S0167-9473(01)00046-9

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series A (General)* **57**, 289–300.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate under dependency. *Annals of Statistics* **29**, 1165–1188. doi:10.1214/aos/1013699998

Black MA (2004) A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society. Series A (General)* **66**, 297–304. doi:10.1111/j.1369-7412.2003.05527.x

Broët P, Lewin A, Richardson S, Dalmasso C, Magdelenat H (2004) A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics (Oxford, England)* **20**, 2562–2571. doi:10.1093/bioinformatics/bth285

Cox DR, Wong MY (2004) A simple procedure for the selection of significant effects. *Journal of the Royal Statistical Society* **66**, 395–400. doi:10.1111/j.1369-7412.2004.05695.x

Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210–219. doi:10.1186/gb-2003-4-4-210

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686. doi:10.1126/science.278.5338.680

Diaconis P, Ylvisaker D (1985) Quantifying prior opinion. In 'Bayesian statistics 2'. (Eds JM Bernardo, MH DeGroot, DV Lindley, AFM Smith) pp. 133–156. (Wiley: New York)

Do K-A, Mueller P, Tang F (2003) A Bayesian mixture model for differential gene expression. Technical Report, Department of Biostatistics, University of Texas/MD Anderson Cancer Center, Houston, TX.

Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 0036.1–0036.21.

Dudoit S, Popper Shaffer J, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103. doi:10.1214/ss/1056397487

Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99**, 96–104.

Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86. doi:10.1002/gepi.1124

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160. doi:10.1198/016214501753382129

Genovese CR, Wasserman L (2002) Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series A (General)* **64**, 499–517. doi:10.1111/1467-9868.00347

Hedenfalk I, Ringnër M, Ben-Dor A, Yakhini Z, Chen Y, *et al*. (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2532–2537. doi:10.1073/pnas.0533805100

Johnson NL, Kotz S (1970) 'Continuous univariate distributions. Vol. 2.' (Wiley: New York)

Kendziorski CM, Newton MA, Lan H, Gould MN (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistical Methodology* **22**, 3899–3914.

Lee MT, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 9834–9838. doi:10.1073/pnas.97.18.9834

Lehmann EL (1959) 'Testing statistical hypotheses.' (Wiley: New York)

McLachlan GJ, Do KA, Ambroise C (2004) 'Analyzing microarray gene expression data.' (Wiley: New York)

Newton MA, Kendziorski C (2003) Parametric empirical Bayes methods for microarrays. In 'The analysis of gene expression data: methods and software'. (Eds G Parmigiani, ES Garrett, RA Irizarry, SL Zeger) pp. 254–271. (Springer: New York)

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52. doi:10.1089/106652701300099074

Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics (Oxford, England)* **5**, 155–176. doi:10.1093/biostatistics/5.2.155

Pan W (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics (Oxford, England)* **18**, 546–554. doi:10.1093/bioinformatics/18.4.546

Pan W (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics (Oxford, England)* **19**, 1333–1340. doi:10.1093/bioinformatics/btg167

Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics (Oxford, England)* **19**, 368–375. doi:10.1093/bioinformatics/btf877

Schena M, Shaon D, Heller R, Chai A, Brown P, Davis RW (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 10614–10619. doi:10.1073/pnas.93.20.10614

Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series A (General)* **64**, 479–498. doi:10.1111/1467-9868.00346

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035. doi:10.1214/aos/1074290335

Storey J, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series A (General)* **66**, 187–205. doi:10.1111/j.1467-9868.2004.00439.x

Storey JD, Tibshirani R (2003*a*) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In 'The analysis of gene expression data: methods and software'. (Eds G Parmigiani, ES Garrett, RA Irizarry, SL Zeger) pp. 272–290. (Springer: New York)

Storey JD, Tibshirani R (2003*b*) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445. doi:10.1073/pnas.1530509100

Tusher VG, Tibshirani R, 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. doi:10.1073/pnas.091062498

Wit E, McClure J (2004) 'Statistics for microarrays: design, analysis and inference.' (Wiley: Chichester)

Zhao Y, Pan W (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics (Oxford, England)* **19**, 1046–1054. doi:10.1093/bioinformatics/btf879

Received 14 February 2005, accepted 6 May 2005