BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 1987–1994
doi:10.1093/bioinformatics/bti301

Gene expression

Empirical Bayes screening of many *p*-values with applications to microarray studies

Susmita Datta[1] and Somnath Datta[2,*]

[1]Department of Mathematics and Statistics, Department of Biology, Georgia State University, Atlanta, GA 30303, USA and [2]Department of Statistics, University of Georgia, Athens, GA 30602, USA

Received on April 23, 2004; revised on December 9, 2004; accepted on January 27, 2005
Advance Access publication February 2, 2005

ABSTRACT

Motivation: Statistical tests for the detection of differentially expressed genes lead to a large collection of *p*-values one for each gene comparison. Without any further adjustment, these *p*-values may lead to a large number of false positives, simply because the number of genes to be tested is huge, which might mean wastage of laboratory resources. To account for multiple hypotheses, these *p*-values are typically adjusted using a single step method or a step-down method in order to achieve an overall control of the error rate (the so-called familywise error rate). In many applications, this may lead to an overly conservative strategy leading to too few genes being flagged.

Results: In this paper we introduce a novel empirical Bayes screening (EBS) technique to inspect a large number of *p*-values in an effort to detect additional positive cases. In effect, each case borrows strength from an overall picture of the alternative hypotheses computed from all the *p*-values, while the entire procedure is calibrated by a step-down method so that the familywise error rate at the complete null hypothesis is still controlled. It is shown that the EBS has substantially higher sensitivity than the standard step-down approach for multiple comparison at the cost of a modest increase in the false discovery rate (FDR). The EBS procedure also compares favorably when compared with existing FDR control procedures for multiple testing. The EBS procedure is particularly useful in situations where it is important to identify all possible potentially positive cases which can be subjected to further confirmatory testing in order to eliminate the false positives. We illustrated this screening procedure using a data set on human colorectal cancer where we show that the EBS method detected additional genes rela- ted to colon cancer that were missed by other methods.

This novel empirical Bayes procedure is advantageous over our earlier proposed empirical Bayes adjustments due to the following reasons: (i) it offers an automatic screening of the *p*-values the user may obtain from a univariate (i.e., gene by gene) analysis package making it extremely easy to use for a non-statistician, (ii) since it applies to the *p*-values, the tests do not have to be *t*-tests; in particular they could be *F*-tests which might arise in certain ANOVA formulations with expression data or even nonparametric tests, (iii) the empirical Bayes adjustment uses nonparametric function estimation techniques to estimate the marginal density of the transformed *p*-values rather than using a parametric model for the prior distribution and is therefore robust against model mis-specification.

Availability: R code for EBS is available from the authors upon request.

*To whom correspondence should be addressed.

Supplementary information: http://www.stat.uga.edu/∼datta/EBS/supp.htm

Contact: datta@stat.uga.edu

INTRODUCTION

In recent years, the problem of simultaneous testing of multiple hypotheses has seen new life mostly due to the recent microarray experiments where expression levels of thousands of genes are simultaneously measured and compared in two (or more) tissue types. This old problem has a new twist, namely, the number of comparisons to be made is extremely large, often running into the ten thousands, and there are only a limited number of replications. To complicate things further, there are design limitations, and also present are sources of potential bias so that some preprocessing of the data is typically necessary. As a result, numerous papers have come out in the last five years or so suggesting various methods, both frequentists, as well as, Bayesian, for the detection of differentially expressed genes in microarray experiments. Some of these are novel methods applicable to general multiple testing (Storey, 2002; Efron, 2004) while others (Kerr *et al*., 2000, 2002; Ideker *et al*., 2000; Newton *et al*., 2001; Tusher *et al*., 2001; Efron *et al*., 2001; Efron and Tibshirani, 2002; Dudoit *et al*., 2002; Lee *et al*., 2003; Storey and Tibshirani, 2003; Ge *et al*., 2003; Reiner *et al*., 2003; Zhao and Pan, 2003 and so on) are specifically designed for microarray studies, including adaptation of existing methods to suit microarray data. A comprehensive review up to 2002 can be found in Pan (2002); also see the literature review in Datta *et al*. (2004). See Dudoit *et al*. (2003) for a comparative review of various error rates of several commonly used multiple testing procedures.

Often times, biologists (practitioners) face the following frustrating situation in dealing with microarray assays involving a large number of genes but a very few replicates. If they attempt to correct for multiple testing using either a familywise error rate control procedure such as the Westfall and Young (1993) (WY hereafter) or an FDR control procedure such as Benjamini and Hochberg (1995) (BH hereafter) they hardly find any 'significant' genes. In other words, these procedures, although statistically correct, are often very conservative in practice. Recently, Datta *et al*. (2004) used the notion of empirical Bayes estimation in the context of microarray testing in a novel way. They adjusted a number of *t*-statistics in such a way that each of the modified statistics had a component that reflected their collective evidence against the complete null hypothesis. The procedure is however calibrated not to reflect a posterior probability, but rather to control the overall familywise error rate under the

complete null. They showed that this technique could greatly increase the sensitivity of the entire procedure at the cost of a modest increase in the false discovery rate.

The present proposal is a much more general attempt in using the idea of empirical Bayes for screening multiple cases. The empirical Bayes screening (EBS) method presented here has the following three distinct advantages over the procedure in Datta *et al*. (2004): (i) First and foremost, a simple resampling procedure can be implemented to carry out the EBS in its simplest form since the null (marginal) distribution of each p-value is uniform. Thus, EBS offers an automatic screening of the p-values a user may obtain, say, from an existing univariate (gene by gene) analysis package. Even though the simplest EBS ignores potential dependence between genes, it is shown to be quite robust with respect to cluster dependence in a simulation setting. (ii) It works directly with the p-values. As a result, the underlying tests do not have to be t-tests; in particular they could be F-tests which might arise in certain ANOVA formulations with expression data (Kerr *et al*., 2000). (iii) The empirical Bayes procedure uses nonparametric techniques to estimate the marginal density of the transformed p-values rather than using a parametric model for the prior distribution and is therefore robust against model mis-specification. A penultimate stage of this development was represented in Datta and Datta (2004).

The performance of the EBS procedure is compared with benchmark procedure of WY. In addition, we also compare two FDR control procedures, namely, the well-known adjustment due to BH and a relatively recent procedure called BUM (Pounds and Morris, 2003). We show that both WY and BH are potentially conservative. While BUM generally has very good sensitivity, it could be quite unstable in terms of FDR and the results could be unreliable for some data sets. Overall, the EBS procedure showed very good sensitivity while maintaining a reasonable FDR in the various simulation settings that we had considered.

Microarray studies are often regarded (and we are of this opinion) as a preliminary screening method in detecting interesting genes and any finding from such studies should be further validated by more rigorous laboratory procedures (such as the RT–PCR). On the other hand, a control of some global error rate (such as the familywise error rate, the false discovery rate, etc.) is important given the rather large number of hypotheses to be tested. We feel that the proposed EBS accomplishes both these objectives.

The rest of the paper is organized as follows. The development of the EBS procedure is provided in the next section. A number of simulation studies of various differential expression patterns are reported in the Simulation Results section. In all cases, the EBS led to an increase in the overall sensitivity compared to the benchmark WY and the other two competing p-value based methods. The Applications to a Cancer Data set section illustrates the EBS procedure using a data set on colorectal cancer. We show that the EBS was able to pick up additional relevant genes compared to the other three methods. The paper ends with a detailed Discussion section.

STATISTICAL METHODS

The empirical Bayes formulation

Suppose we have a number of tests of similar structure with associated p-values denoted $\widehat{p}_i$, $1 \le i \le M$. In microarray studies, M would equal the total number of genes (probe sets, etc.) on a microarray and for the ith gene $\widehat{p}_i$ might be the observed level of significance for a test that compares its average expression levels in two tissue types, say normal versus cancer cells. The p-values indicate evidence against the null hypotheses in the sense that the smaller a p-value, the more significant the evidence is that the gene is indeed differentially expressed. In general, it is defined as the chance of observing a value of the test statistic that is as extreme as (e.g., as large as) the value of the test statistic for the sample at hand, when indeed the gene is not differentially expressed. Thus, it is always a function of the sample test statistic and hence a random variable. Under the null hypothesis of no differential expression, $\widehat{p}_i$ is uniformly distributed on the interval $(0, 1)$ and therefore $z_i = \Phi^{-1}(\widehat{p}_i)$ is distributed as standard normal ($N(0, 1)$) where Φ is the standard normal c.d.f. (cumulative distribution function). In the empirical Bayes formulation, we embed these distributions in a larger family of parametric distributions which also support the alternative hypotheses. Because we are considering tests that are of similar structure, one such model would be to assume an $N(\theta_i, 1)$ distribution for z_i. Since our goal is to identify cases with 'small' p-values, we would test a new set of hypotheses $H_0^i{:}\theta_i = 0$ versus $H_1^i{:}\theta_i < 0$, in this model. Since we are faced with simultaneous testing of a (large) number of hypotheses, we might do better by combining evidence of all tests using an empirical Bayes approach (Robbins, 1964; Efron and Morris, 1975). To that end, assume a common, but unknown, prior distribution G, say, for each θ_i. A Bayes test of H_0^i against H_1^i would reject H_0^i for small values of $\widehat{\theta}_i^B$, that is, the posterior mean $E(\theta_i|z_i)$ of θ_i. Since the prior distribution G is unknown, this posterior mean needs to be estimated through nonparametric function estimation techniques from data across all tests that share this common prior distribution which we now pursue. The resulting estimated posterior mean would be called an empirical Bayes estimate (EBE) of θ_i, and is denoted $\widehat{\theta}_i^{\mathrm{EB}}$.

Construction of $\widehat{\theta}_i^{\mathbf{EB}}$

It follows by differentiation under the integral sign of the marginal density f_G of z_i that

$$f'_G(z_i) = -z_i f_G(z_i) + \int_{-\infty}^{\infty} \theta \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(z_i-\theta)^2}{2}\right\} \mathrm{d}G(\theta)$$

leading to the following well-known expression (see, e.g., Carlin and Louis, 2000, p. 86) of the posterior mean

$$\widehat{\theta}_i^B = z_i + \frac{f'_G(z_i)}{f_G(z_i)}.$$

Since f_G is the common marginal density of the z_i, we could estimate f_G by a nonparametric kernel density estimator based on the z_i,

$$\widehat{f_G}(t) = \frac{1}{Mh} \sum_{j=1}^{M} \phi\left(\frac{t-z_j}{h}\right), \quad t \ge 0,$$

where ϕ is the standard normal kernel (density) and h is a small positive number, called the bandwidth. The parameter h is user selectable and there are numerous methods such as likelihood cross validation, asymptotic minimization of the integrated mean squared errors etc. available in various statistical software packages for choosing h. Alternatively, a visual inspection ('eyeballing') of the resulting density plot may suffice in many applications. Since we want each p-value to borrow strength from the smallest p-values we want a longer left tail for the estimated density and therefore it might be better to oversmooth $\widehat{f_G}$ somewhat for greater sensitivity. We took this last approach for our application to the cancer data. A bandwidth of 0.7 led to a smooth left tail even though there were some large negative z values (Fig. 1).

Substitution of $\widehat{f_G}(t)$ in the above expression of the Bayes estimator leads to the following formula for the EBE of θ_i:

$$\widehat{\theta}_i^{\mathrm{EB}} = z_i - h^{-2} \frac{\left\{\sum_{j=1}^{M} (z_i - z_j)\phi\left(\frac{z_i-z_j}{h}\right)\right\}}{\left\{\sum_{j=1}^{M} \phi\left(\frac{z_i-z_j}{h}\right)\right\}}. \quad (1)$$

Step-down p-value calculation

We would calibrate our screening procedure such that a familywise error rate (FWER) of $\alpha \in (0, 1)$ is maintained. This represents the probability of

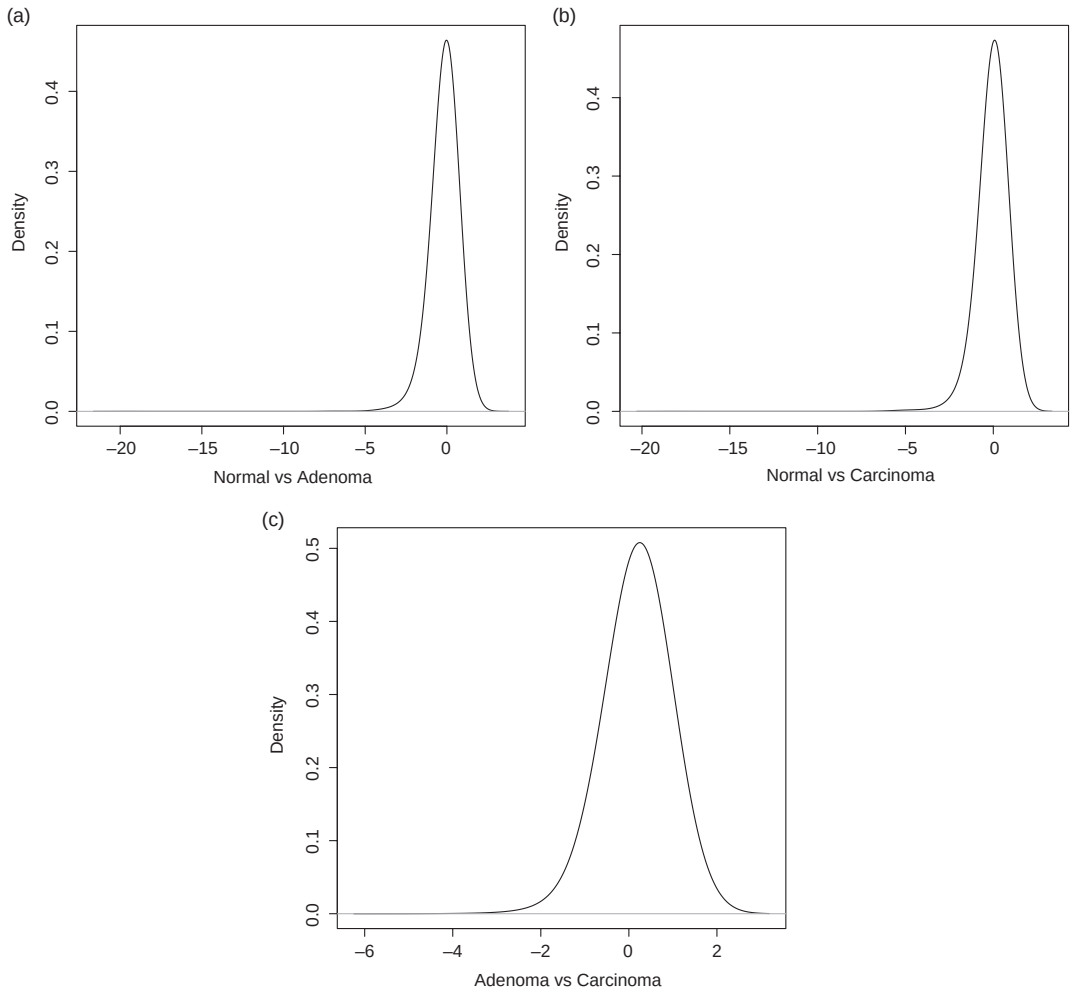

Fig. 1. Distributions of the normal transforms (Φ^{-1}) of the p-values in the colorectal cancer data.

reaching at least one significant conclusion when indeed the complete null hypothesis is true. To that end, we resort to the resampling-based step-down procedure of Westfall and Young (1993). Amongst many existing procedures of FWER control, this is generally regarded as one of the best (least conservative). After calculating the empirical Bayes estimates $\widehat{\theta}_i^{\mathrm{EB}}$ given by Equation (1), for all genes i, we compute the corresponding step-down adjusted p-values $\tilde{p}_i$ with the following algorithm.

Step 1. Find the rank orders r_i such that $\widehat{p}_{r_1} \leq \cdots \leq \widehat{p}_{r_M}$ and let $u_i = \widehat{\theta}_{r_i}^{\mathrm{EB}}$, $1 \leq i \leq M$.

Step 2. Generate a collection of random variables z_i^*, $1 \leq i \leq M$ from the (approximate) null distribution of the original z_i, $1 \leq i \leq M$.

Step 3. Convert the z_i^* to the corresponding EBEs $\widehat{\theta}_i^{*\mathrm{EB}}$ by Equation (1), with z_i^* in place of z_i throughout and let $u_i^* = \widehat{\theta}_{r_i}^{*\mathrm{EB}}$, $1 \leq i \leq M$ (note that the ordering r_i is not changed during resampling) and monotonize them by $u_i^* = \min(u_i^*, u_{i+1}^*)$, $i = (M-1), \ldots, 1$.

Step 4. Repeat Steps 2 and 3 a large number of times, say B, and denote the u_i^* values by $u_i^*(1), \ldots, u_i^*(B)$.

Step 5. Compute

$$\tilde{p}_{r_i} = B^{-1} \sum_{l=1}^{B} I(u_i^*(l) \leq u_i)$$

and monotonize them as $\tilde{p}_{r_i} = \max(\tilde{p}_{r_i}, \tilde{p}_{r_{i-1}})$, for $i = 2, \ldots, M$.

Finally, declare cases (e.g., genes) $r_1, \ldots, r_{k_\alpha}$ to be significant (e.g., differentially expressed), where

$$k_\alpha = \max\{1 \leq k \leq M : \tilde{p}_{r_k} \leq \alpha\}.$$

Step 2 above can be carried out in a variety of ways depending on the situation. In the simplest form, z can be generated by random sampling from $N(0, 1)$ and that is what we advocate in practice since it provides an automatic procedure. In essence, it assumes the tests (genes) are independent. Although it is not a correct assumption, we show through simulation studies in the next section that the performance of the EBS is quite robust even if this assumption is violated. However, more sophisticated choices are sometimes possible if the original data yielding the p-values are available. For example, in the context of a two-sample problem (e.g., pooled t-tests), z^* could be obtained by calculating the z-transforms of the observed level of significance of the test statistics calculated using randomly resampled or permuted vectors of the observations of all gene expressions from the original data (Dudoit *et al*., 2002). Datta *et al*. (2004) suggested creating pseudo data sets by resampling the residuals in an ANOVA model for the gene expression in multiple tissue types.

Working of the EBS

The 'sharing of evidence' or 'borrowing of strength' of the EBS procedure can be easily seen from the expression (1) of the EBE $\widehat{\theta}_i^{\mathrm{EB}}$. The first term z_i is a monotonic transformation of the ith p-value p_i and therefore comparing

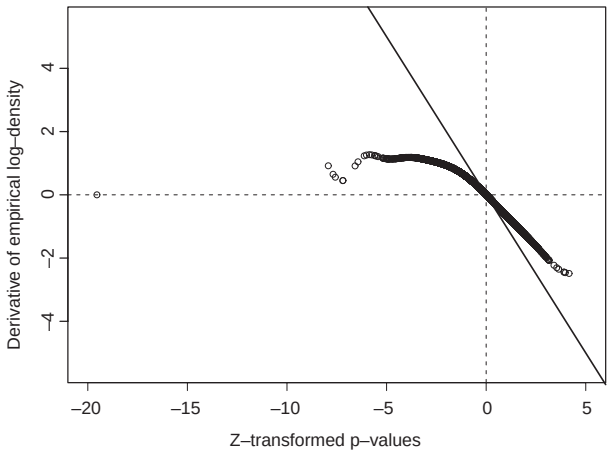

Fig. 2. Scatter plot of derivative of empirical log density against transformed p-values for the 'normal versus adenoma' comparison in the colorectal cancer data.

it with its null distribution is equivalent to the standard WY procedure. On the other hand, the second term in $\widehat{\theta}_i^{\mathrm{EB}}$ is an estimate of the derivative of the logarithmic marginal density $(\log f_G)'$ and its stochastic behavior under the complete null hypothesis will be different from its behavior under the alternative. Generally speaking, in microarray studies f_G will have a fatter left tail than the standard normal density (the null density). Thus, this term will tend to be smaller than a typical value under the complete null and, together, these terms will make $\widehat{\theta}_i^{\mathrm{EB}}$ stochastically even smaller than a corresponding value under the null. In other words, the degree of stochastic difference will be larger than that using z_i alone. The term $(\log \widehat{f_G})'$ represents an overall evidence against the complete null for such genes.

As a concrete example, consider the 'normal versus adenoma' comparison for the colorectal cancer data considered in the 'Applications to a Cancer Data Set' section. A typical alternative in microarray studies is represented in the shape of the empirical density $\widehat{f_G}$ in Figure 1. Clearly, $\widehat{f_G}$ has a tail shape as described in the previous paragraph. Figure 2 displays a scatter plot of $(\log \widehat{f_G})'(z_i)$ against z_i for this data set. In the same plot, the diagonal line $y = -z$ represents the null value of $(\log f_G)'$. We can see that for potentially informative genes (say, those corresponding to negative z_i), it tends to be below the diagonal line indicating an overall presence of 'differentially expressed genes' corresponding to a given z level.

SIMULATION RESULTS

In this section, we report the results of a number of simulation studies where we compute various performance measures for screening multiple p-values using both standard WY and EBS as well as the FDR control procedures BH and BUM. These are:

(i) *Sensitivity:* proportion amongst differentially expressed genes that were declared significant;

(ii) *Specificity:* proportion amongst non-differentially expressed genes that were not declared significant;

(iii) *False discovery rate (FDR):* proportion amongst genes declared significant that were not differentially expressed;

(iv) *False non-discovery rate (FNR):* proportion amongst genes declared not significant that were differentially expressed.

Two-sample paired comparisons

We consider a set of simultaneous paired t-tests which can be applicable, for example in studying gene expression levels between two samples hybridized on the same cDNA microarray. For simplicity we assume that the tests are independent. Although this is not likely to hold in the microarray setting, the results are still useful for understanding the utility of the empirical Bayes adjustment. Moreover, we subsequently study the performance of the EBS method based on this simple assumption in a dependent data setting.

We consider p-values arising from $M = 2000$ one-sample t-tests (or equivalently two-sample paired t-tests) with $r = 3, 5$ and 8 replicates, respectively, where four types of alternative hypotheses (differential gene expression) patterns are created in terms of means of normal data (with unit variance scale). The data can be thought of as the difference in the log-expression levels of gene expression in two tissue types in the microarray context. A graph of the non-zero means in the four simulated models is shown in Figure 3. As can be seen from this figure, the proportion of non-null hypotheses ranges between 2.5 and 13%. In each setting, we simulated data and computed 2000 t-statistics which were converted into p-values using $\widehat{p}_i = Pr\{|t(r-1)| > |t_{i,\mathrm{obs}}|\}$, where $t_{i,\mathrm{obs}}$ is the observed value of the ith t-test, $1 \le i \le M$. We let $z_i = \Phi^{-1}(\widehat{p}_i)$. The z_i^* were generated from i.i.d. $N(0,1)$ and $B = 500$ batches of resamples were used. The overall FWER was controlled at $\alpha = 5\%$. For each sample, $\tilde{p}$ were calculated in two ways: (i) using $\widehat{\theta}$, as described in the algorithmic steps in the previous section (which we refer to as 'EBS'), and (ii) using $\widehat{p}$ following the same algorithmic steps but with $u_i = \widehat{p}_{r_i}$, $u_i^* = p_{r_i}^*$, where $p_i^* = \Phi(z_i^*)$ (which we refer to as 'WY'). Note that WY is a standard step-down procedure that has been in use for maintaining FWER in multiple hypotheses tests (Westfall and Young, 1993). The results in Table 1 were all based on a somewhat arbitrarily chosen bandwidth of $h = 0.8$. Although we do not report the details, other nearby bandwidths were also investigated and sensitivity gains over WY were noticed in all cases. For the purpose of comparison we also study the performances of two FDR control procedures BH and BUM (one old, one new) where the FDR threshold level is set at 5%. Ideally, this would also ensure weak control of FWER at 5% (Benjamini and Hochberg, 1995). For each simulation setting, all four procedures were independently replicated 50 times and the four performance measures were calculated based on the average proportions based on these 50 runs.

The amount of specificity of all procedures under study here were at least 99% in all cases (with the exception of BUM in a few cases) and are not reported further in Table 1.

The results in Table 1, which are shown in percentages, clearly show that substantial sensitivity gain was achieved by employing the EBS over the benchmark WY, especially in the low sensitivity region. Overall, EBS compares extremely favorably amongst the competing methods. Even though BUM has decent sensitivity as well, it can break down completely in terms of controlling FDR in cases where a BUM model does not adequately reflect the empirical distribution of the p-values, as shown in Simulation 3. Basically, under this scenario, for $r = 3$, BUM incorrectly estimated the proportion of null hypothesis to be 1 for most of the 50 runs. EBS, on the other hand, managed to maintain an acceptable level of FDR in all cases that we studied even though it is not explicitly controlled in this procedure. The EBS appears to have the smallest FNR in all cases.

Next, we study the performance of the above EBS method that implicitly assumes that the tests are independent (as incorporated in the resampling stage) in the case of cluster-dependent data. Consider, for example, log-transformed data d_i for the ith gene

Table 1. Performance of various procedures for 2000 independent one-sample t-tests

Simulation	Replication size r	Sensitivity WY	EBS	BH	BUM	Sensitivity gain over WY (%)	FDR WY	EBS	BH	BUM	FNR WY	EBS	BH	BUM
1	3	0.2	7.3	0.2	4.8	4294	4.0	9.9	2.0	9.5	10.8	10.1	10.8	10.3
	5	12.1	81.5	78.7	78.0	572	0.0	5.8	4.5	4.3	9.6	2.2	2.5	2.6
	8	66.7	91.2	92.9	93.2	37	0.0	2.8	4.5	4.9	3.8	1.1	0.9	0.8
2	3	0.2	6.0	0.2	5.7	3162	4.0	9.5	2.0	10.3	13.0	12.4	13.0	12.4
	5	10.1	84.0	79.7	79.3	728	0.0	6.0	4.3	4.3	11.9	2.4	3.0	3.0
	8	62.4	93.6	94.6	94.9	50	0.0	3.2	4.4	4.8	5.3	1.0	0.8	0.8
3	3	0.8	16.4	0.8	96.4	2063	4.0	13.2	2.0	94.0	2.5	2.1	2.5	0.0
	5	55.9	100	100	100	79	0.2	4.5	4.8	3.8	1.1	0.0	0.0	0.0
	8	100	100	100	100	0	0.1	1	5.1	5.1	0.0	0.0	0.0	0.0
4	3	0.2	28.7	0.4	12.8	12370	4.0	8.1	2.0	8.3	10	7.3	10.0	8.8
	5	18	100	100	100	46	0.0	5.5	4.5	4.2	8.3	0.0	0.0	0.0
	8	99.6	100	100	100	0.4	0.0	1.3	4.5	5.0	0.0	0.0	0.0	0.0

The thresholds for FWER in case of WY and EBS and for FDR in case of BH and BUM were all taken to be 5%. The measures are reported in percentages.

Table 2. Performance of various procedures (at 5%) for dependent one-sample t-tests

Simulation	Replication size r	Number of clusters	Sensitivity WY	EBS	BH	BUM	FDR WY	EBS	BH	BUM	FNR WY	EBS	BH	BUM
1	3	10	0.2	11.1	0.3	7.1	6.0	11.1	2.3	10.5	10.8	9.7	10.8	10.1
		50	0.1	7.5	0.2	4.9	4.0	11.0	1.5	10.8	10.8	10.1	10.8	10.3
		200	0.2	8.5	0.4	5.5	4.0	10.1	4.8	9.7	10.8	10.0	10.8	10.3
	5	10	13.0	82.3	79.6	79.2	0.0	5.6	4.5	4.7	9.5	2.1	2.4	2.5
		50	12.9	81.4	78.4	77.6	0.0	5.6	4.2	4.0	9.5	2.2	2.6	2.7
		200	12.0	81.8	78.9	78.1	0.0	5.5	4.3	4.1	9.6	2.2	2.5	2.6
3	3	10	0.6	19.9	0.8	67.7	5.0	13.3	3.4	67.0	2.5	2.0	2.5	0.1
		50	0.3	17.9	0.3	85.0	6.0	14.2	6.0	83.3	2.5	2.1	2.5	0.0
		200	0.8	18.5	1.5	88.6	4.0	12.3	5.0	85.8	2.5	2.0	2.5	0.0
	5	10	59.9	100	100	100	0.0	4.6	5.1	10.3	1.0	0.0	0.0	0.0
		50	56.2	100	100	100	0.0	4.2	4.3	3.5	1.1	0.0	0.0	0.0
		200	56.6	100	100	100	0.1	5.0	5.1	3.9	1.1	0.0	0.0	0.0

generated by

$$d_{ij} = \mu_i + 2^{-1/2}\epsilon'_{I(i),j} + 2^{-1/2}\epsilon_{ij}, \quad 1 \le j \le r, \qquad (2)$$

where r is the replication size, $I = I(i)$ denotes the cluster-containing gene i; $1 \le i \le M = 2000$. The error terms ϵ and ϵ' are generated from independent $N(0,1)$ and the mean vector μ was the same as in the simulations above. Biologically speaking, we are envisioning that genes in a cluster act in consort resulting in correlated expression measures. Basically, Equation (2) ensures that a pair of genes belonging to the same cluster have non-zero correlation; however if they belong to different clusters their expression values are independent or uncorrelated. To see this, consider a pair of genes i and i'. If they both belong to the same cluster k, say then $I(i) = I(i') = k$ and the same error term $\epsilon'_{k,j}$ is used in generating the expression values d_{ij} and $d_{i'j}$ leading to $\mathrm{Corr}(d_{ij}, d_{i'j}) = 0.5$. As before, the following table was computed based on 50 runs and in each run we perform step-down calculation using 500 i.i.d. bootstrap replicates from standard normal. In other words, the identical procedure as before was followed without the knowledge that the data was generated this way. We chose three different numbers of clusters (of equal sizes), 10, 50 and 200. The results are reported in Table 2. For the sake of brevity, we only report the results for Simulations 1 and 3 with sample sizes 3 and 5. The FDR and the FNR of all the procedures appear to be insensitive to the number of clusters while the sensitivity is more variable in the low region. Overall, the performance of EBS appears to be quite robust with respect to dependent tests and it continues to enjoy its superior sensitivity property.

APPLICATIONS TO A CANCER DATA SET

We now illustrate our screening procedure with a real data set on colorectal cancer. This data set was featured in Datta *et al*. (2004). There were altogether nine Affymetrix chips corresponding to three different patients (individuals) and three tissue types, normal, adenoma and carcinoma. In this paper, we look at a subset of about 10 000 genes out of over 12 000 genes in the Affymetrix U95 chipset, whose expression levels were judged to be reliable. The initial p-values (unadjusted) were calculated from an ANOVA model described below applied to this data set.

ANOVA models for expression data

The ANOVA approach (linear models) has become a standard modeling tool to describe the (log-transformed) expression levels of genes in experiments involving multiple tissue types (Kerr *et al*., 2000, 2002; Kerr and Churchill, 2001; Datta *et al*., 2004; Datta and Datta, 2004; etc.). Consider an experiment in which we have measured the expression level X (appropriately normalized and transformed) for r individuals, g genes and J tissue types (varieties). Consider a design where a single microarray consisted of the expression levels of all genes for an individual in a given tissue type. We model X as

$$X_{ijk} = \mu + I_k + G_i + V_j + (IG)_{ik} + (GV)_{ij} + \epsilon_{ijk}; \qquad (3)$$

here $1 \le i \le g$, $1 \le j \le J$ and $1 \le k \le r$ index genes, tissue types (variety) and individuals, respectively. The ϵ_{ijk}'s denote i.i.d. mean zero normal errors. In this model, μ represents the overall or mean expression level; main effects I_k, G_i and V_j reflect the overall differences in the expression levels for individuals, genes and varieties, respectively and the interaction term $(IG)_{ik}$ accounts for the variability of expression of the ith gene among individuals. It is perhaps more reasonable to assume the individual effect I_k to be random. However for the sake of simplicity we treat these as fixed effects. Also the independence of the error terms across genes is a simplifying assumption. However, as demonstrated in the earlier simulation example, the procedures based on the independent error assumption continue to have reasonable performance for certain types of dependent errors (cluster dependence). Our primary interest lies in the gene×tissue-type interaction $(GV)_{ij}$ which measures the effect of gene i in tissue type j. The null hypothesis of no differential expression of gene i in two tissue types j_1 and j_2 is expressed as $H_0^{i;j_1,j_2} : (GV)_{ij_2} - (GV)_{ij_1} = 0$. The t-statistic testing $H_0^{i;j_1,j_2}$ is

$$t_{i;j_1,j_2} = \sqrt{\frac{gr}{2(g-1)}} \frac{\left(\overline{X}_{ij_2\cdot} - \overline{X}_{\cdot j_2\cdot} - \overline{X}_{ij_1\cdot} + \overline{X}_{\cdot j_1\cdot}\right)}{\widehat{\sigma}}$$

with

$$\widehat{\sigma}^2 = \sum_{ijk} \left(X_{ijk} - \overline{X}_{ij\cdot} - \overline{X}_{i\cdot k} + \overline{X}_{i\cdot\cdot}\right)^2 \Big/ \nu,$$

$\nu = g(r-1)(J-1)$. The p-value for testing differential expression of gene i between tissue types j_1 and j_2 is given by $\widehat{p}_{i;j_1,j_2} = 2\left[1 - \mathcal{P}^{t(\nu)}(|t_{i;j_1,j_2}|)\right]$, where $\mathcal{P}^{t(\nu)}$ is the cumulative distribution function of a central t distribution with ν degrees of freedom.

Figure 1 shows the smoothed empirical distributions (i.e., the estimated marginal density $\widehat{f_G}$) of the normal transforms (Φ^{-1}) of these p-values for 'normal versus adenoma', 'normal versus carcinoma' and 'adenoma versus carcinoma' comparisons for the colorectal data set.

For each tissue-pair comparison, we compute the p-values for all the genes using the above formulas and then feed them into the EBS procedure based on i.i.d. uniform resampling. We also ran the other three competing procedures on the same set of p-values. As in the simulation section, we take the overall FWER of $\alpha = 0.05$ for the WY and the EBS and FDR control at 5% for BH and BUM. In all cases, EBS has flagged more genes compared to other procedures. These are summarized in Table 3. Note from Figure 1 that the empirical distribution for the 'adenoma versus carcinoma' comparison is very close to the complete null distribution (i.e., standard normal). The BUM method broke down for this case since it estimated the proportion of null hypotheses to be one (see earlier comments about Simulation 3).

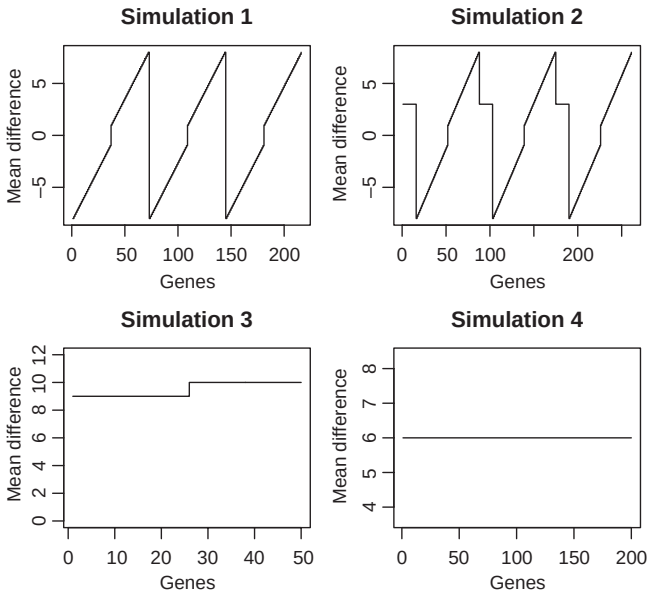

Fig. 3. Non-null mean differential expression (log-scale) of four simulation models.

Table 3. Number of genes that were declared to be differentially expressed for the cancer data

	Normal versus adenoma	Adenoma versus carcinoma	Normal versus carcinoma
WY	47	7	54
EBS	252	44	210
BH	211	23	183
BUM	223	NA	183

Validation

We inspected the results for the normal versus adenoma comparison in more detail. Figure 4 demonstrates the thresholds (indicated by vertical lines) for the four p-value-based procedures under consideration. In each case, genes whose p-values are to the left of the threshold are flagged (declared to be significant). Thus, EBS picked up 29 additional genes that were not flagged by any other procedure. Although we have demonstrated through simulation that indeed the EBS picks up more differentially expressed genes as evident from its superior sensitivity, we wanted to see whether these additional genes include any meaningful genes.

After searching the published literature, we found that indeed several of these genes have been linked to colorectal or other cancers in the past. We present a selected subset of five genes from this list in Table 4.

The first gene on this list, CDC2, is a well-known cancer gene (CG_ID 278) whose overexpression is colorectal adenocarcinoma is well documented (Kim *et al*., 1999). In a previous study, FUT4 gene was found to be expressed in human colorectal cancer tissues and colorectal cancer cell lines (see Nishihara *et al*., 1999) which is

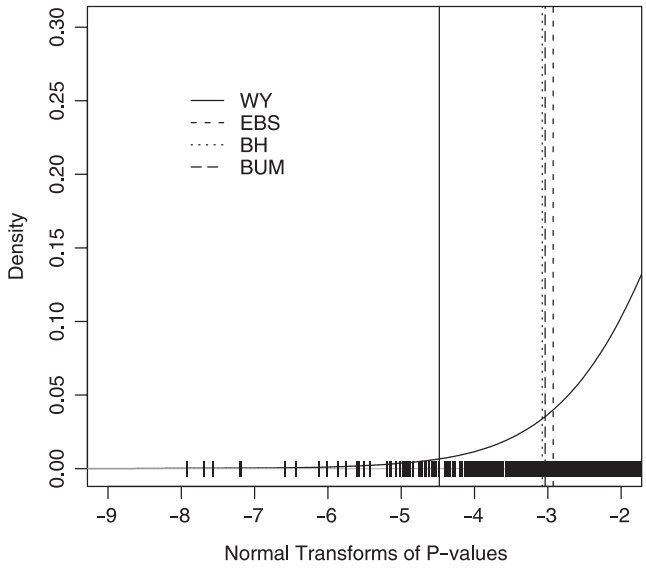

Fig. 4. Normal versus adenoma comparison: genes declared significant by the various methods are to the left of the respective thresholds.

Table 4. A selected list of genes that were only flagged by the EBS procedure for the normal versus adenoma comparison

Affy probe set	Gene name	Description
1803_at	CDC2	Cell cycle controller
39210_at	FUT4	Displays fucosyltransferase activity for type 2 (Gal beta1→GlcNac) containing oligosaccharides and neolactotetraosylceramide
36627_at	SPARCL1	Putative tumor-supressor
317_at	LGMN	Responsible for legumain activity
34319_at	S100P	S100 calcium binding protein P

further explained in Yazawa *et al*. (2002). SPARCL1 is a well-known cancer gene (CG_ID 1662) whose downregulation occurs in human non-small cell lung cancer (NSCLC), and also in prostate and colon carcinomas. It has been suggested (using a colon carcinoma model) that cancer gene LGMN (legumain, CG_ID 2178) could be a target for therapy (Liu *et al*., 2003). Cells overexpressing legumain possessed increased migratory and invasive activity *in vitro* and adopted an invasive and metastatic phenotype *in vivo*. S100P is a known cancer gene (CG_ID 2860) that has been linked with various cancers such as breast, pancreatic and prostate cancers (Guerreiro *et al*., 2000; Sato *et al*., 2004; Mousses *et al*., 2002).

DISCUSSION

In this paper, we propose a novel EBS procedure when one needs to decide about a large number of null hypotheses. Unlike the empirical Bayes adjustment of Datta *et al*. (2004) which only applies to studentized test statistics, the EBS procedure can work directly with a set of p-values produced by individual tests. Thus it offers an automatic screening of the p-values a user may obtain from his or her favorite gene-by-gene analysis software. In addition, the current procedure utilizes the p-values and not the test statistics; therefore, it has broader applicability to other types of tests such as the F-tests or rank tests. For example, in a microarray experiment involving multiple tissue types (e.g., normal, adenoma and carcinoma) one would be able to detect genes that are differentially expressed amongst the various types of tissues (without restricting attention to a particular tissue pair). The EBS procedure screens each p-value not only on its own magnitude but also on the basis of the totality of the p-values (or its empirical distribution). In that sense, each p-value may borrow evidence from other p-values leading to a detection of a greater number of 'interesting cases', when the complete null is false, while maintaining a control on the familywise error rate under the complete null.

As stated in the introduction, there are other global statistical approaches such as SAM (Tusher *et al*., 2001) and VERA (Ideker *et al*., 2000), for the detection of differentially expressed genes in microarray studies. In this paper, we specifically restrict our attention to methods that are based on gene-by-gene p-values. Three such existing methods have been compared with our proposed method. We conclude this section with a number of additional comments.

Strong versus weak control of FWER

The WY step-down procedure yields strong control of FWER under a subset pivotality condition. This means that even when the complete null is not true, the procedure will declare at least one of the component null hypotheses as positive with probability at most α. However, this will not be true for EBS which can only be calibrated at the complete null. Since philosophically (as well as algebraically) it uses shared or borrowed evidence from all hypotheses, the non-null distribution of one p-value affects the distribution of all the empirical Bayes estimates. Note that FDR control procedures such as BH or BUM also control the FWER in the weak sense. We feel that for many applications, the strong control requirement is unduly conservative and would recommend using the EBS nevertheless. However, if desired, a modified version of the EBS can be constructed as follows that would have better control of FWER under incomplete null hypotheses at the expense of lower sensitivity.

Choose a positive integer $1 \leq M_0 < M$, where $(M_0 + 1)$ represents one prior assessment of the maximum number of null hypotheses (for controlling the type 1 error rate protection). The original EBE procedure assumes $M_0 = M - 1$. Follow the same algorithmic steps as before, except for $1 \leq i \leq (M - M_0)$, estimate the derivative of the logarithmic derivatives using the empirical distribution of $z_{r_i}, \ldots, z_{r_M}$, and for $(M - M_0) < i \leq M$, use the empirical distribution of $z_{r_{M-M_0}}, \ldots, z_{r_M}$.

Modeling dependencies amongst genes

In this paper, a relatively straightforward analysis of a microarray data set is presented which does not account for gene to gene correlation. While this suffices for the illustration of the EBS procedure as a method of screening a large number of p-values, a more sophisticated data analysis would need to consider a correlation structure for the gene expressions. While the correct correlation structure may be too complicated and nearly impossible to formulate or to estimate on the basis of limited microarray data that is typically available, a good compromise would be to add a random effect term in the ANOVA model (3) corresponding to the cluster a gene belongs to. Of course, this approach would have to assume that the cluster memberships are known which can only be implemented in practice by

an initial clustering procedure. This would amount to assuming a constant correlation within each cluster. A parametric bootstrap, generating data from the appropriate normal distributions, will have to be employed in order to carry out the step-down procedure. On the other hand, simulation studies in this paper show that the EBS using p-value calculation based on independence assumption seems to have reasonable performance under this type of dependence as well.

Other applications

While we propose the EB screening procedure primarily for the detection of differentially expressed genes, it is applicable in any situation where one has to simultaneously decide about a large number of null hypotheses. This is particularly suitable in situations where it is not of a severe consequence to falsely reject some of the null hypotheses in case the complete null is false. In a sense, this can be viewed as an initial screening procedure where the positive results could be investigated further for confirmation. Proteomics (mass spectrometry) data is another example where the problem of feature (or m/z ratio) selection between competing tissue types can be thought of as a simultaneous testing problem. Perhaps one goal might be to use the selected features from a set of training samples to build a classifier. A classifier such as the Random Forest (Breiman, 2001) can handle 'extra' variables and therefore this might be a good application even if some m/z ratios that are not important in differentiating the spectra are selected as variables for the classifier along with the important m/z ratios.

ACKNOWLEDGEMENTS

Susmita Datta's research was supported in part by the Distinguished Cancer Clinicians and Scientists Program of Georgia Cancer Coalition. We thank the reviewers for their constructive comments which led to a much improved manuscript.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser B.*, **57**, 289–300.

Breiman,L. (2001) Random forests. Technical Report 567, Stat. Dept., UCB.

Carlin,B.P. and Louis,T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Chapman & Hall/CRC, Boca Raton.

Datta,S. and Datta,S. (2004) An empirical Bayes adjustment to multiple p-values for the detection of differentially expressed genes in microarray experiments. In Chen, Y-P. (ed.), *Bioinformatics 2004, Conferences in Research and Practice in Information Technology*—Second Asia-Pacific Bioinformatics Conference, Vol. 29, pp. 155–159.

Datta,S. *et al*. (2004) An empirical Bayes adjustment to increase the sensitivity of detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **20**, 235–242.

Dudoit,S. *et al*. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12**, 111–139.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in Microarray experiments. *Statist. Sci.*, **18**, 71–103.

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

Efron,B. and Morris,C. (1975) Data analysis using Stein's estimator and its generalization. *JASA*, **70**, 311–319.

Efron,B. *et al*. (2001) Empirical Bayes analysis of a microarray experiment. *JASA*, **96**, 1151–1160.

Efron,B. and Tibshirani,R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, **23**, 70–86.

Ge,Y. *et al*. (2003) Resampling based multiple testing for microarray data analysis. *TEST*, **12**, 1–44 (plus discussion pp. 44–77).

Guerreiro,D.S.I. *et al*. (2000) S100P calcium-binding protein overexpression is associated with immortalization of human breast epithelial cells *in vitro* and early stages of breast cancer development *in vivo*. *Int. J. Oncol.*, **16**, 231–240.

Ideker,T. *et al*. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comp. Biol.*, **7**, 805–817.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Kerr,M.K. *et al*. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statist. Sinica*, **12**, 203–217.

Kerr,M.K. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comp. Biol.*, **7**, 819–837.

Kim,J.H. *et al*. (1999) Amplified CDK2 and cdc2 activities in primary colorectal carcinoma. *Cancer*, **8**, 546–553.

Lee,K.E. *et al*. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

Liu,C. *et al*. (2003) Overexpression of legumain in tumors is significant for invasion/metastasis and a candidate enzymatic target for prodrug therapy. *Cancer Res.*, **63**, 2957–2964.

Mousses,S. *et al*. (2002) Clinical validation of candidate genes associated with prostate cancer progression in the CWR22 model system using tissue microarrays. *Cancer Res.*, **62**, 1256–1260.

Newton,M.A. *et al*. (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.*, **8**, 37–52.

Nishihara,S. *et al*. (1999) α1,3-Fucosyltransferase 9 (FUT9; Fuc-TIX) preferentially fucosylates the distal GlcNAc residue of polylactosamine chain while the other four α-1,3FUT members preferentially fucosylate the inner GlcNAc residue. *FEBS Lett.*, **462**, 289–294.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **12**, 546–554.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Robbins,H. (1964) The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, **35**, 1–20.

Sato,N. *et al*. (2004) Identification of maspin and S100P as novel hypomethylation targets in pancreatic cancer using global gene expression profiling. *Oncogene*, **23**, 1531–1538.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B.*, **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

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

Yazawa,S. *et al*. (2002) Tumor-related expression of α1,2fucosylated antigens on colorectal carcinoma cells and its suppression by cell-medicated priming using sugar acceptors for α1,2fucosyltransferase. *Glycobiology*, **12**, 545–553.

Zhao,Y. and Pan,W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **19**, 1046–1054.