Sociedad Española de Estadística e Investigación Operativa

Test

Volume 12, Number 1. June 2003

Resampling-based Multiple Testing for Microarray Data Analysis

Yongchao Ge
Department of Statistics
University of California, Berkeley.
Sandrine Dudoit
Division of Biostatistics
University of California, Berkeley.
Terence P. Speed
Division of Genetics and Bioinformatics,
The Walter and Eliza Hall Institute of Medical Research, Australia.
and Department of Statistics
University of California, Berkeley.

Sociedad de Estadística e Investigación Operativa
Test (2003) *Vol.* **12**, *No.* 1, pp. 1–77

Sociedad de Estadística e Investigación Operativa
Test (2003) *Vol.* **12**, *No.* 1, pp. 1–77

Resampling-based Multiple Testing for Microarray Data Analysis

Yongchao Ge*
Department of Statistics
University of California, Berkeley.
Sandrine Dudoit
Division of Biostatistics
University of California, Berkeley.
Terence P. Speed
Division of Genetics and Bioinformatics,
The Walter and Eliza Hall Institute of Medical Research, Australia.
and Department of Statistics. University of California, Berkeley.

Abstract

The burgeoning field of genomics has revived interest in multiple testing procedures by raising new methodological and computational challenges. For example, microarray experiments generate large multiplicity problems in which thousands of hypotheses are tested simultaneously. Westfall and Young (1993) propose resampling-based p-value adjustment procedures which are highly relevant to microarray experiments. This article discusses different criteria for error control in resampling-based multiple testing, including (a) the family wise error rate of Westfall and Young (1993) and (b) the false discovery rate developed by Benjamini and Hochberg (1995), both from a frequentist viewpoint; and (c) the positive false discovery rate of Storey (2002a), which has a Bayesian motivation. We also introduce our recently developed fast algorithm for implementing the minP adjustment to control family-wise error rate. Adjusted p-values for different approaches are applied to gene expression data from two recently published microarray studies. The properties of these procedures for multiple testing are compared.

Key Words: multiple testing, family-wise error rate, false discovery rate, adjusted p-value, fast algorithm, minP, microarray.

AMS subject classification: 62J15, 62G09, 62P10.

1 Introduction

The burgeoning field of genomics has revived interest in multiple testing procedures by raising new methodological and computational challenges. For example, microarray experiments generate large multiplicity problems

Received: February 2003; Accepted: February 2003

Address for correspondence:Yongchao Ge, Department of Statistics, University of California, Berkeley. 367 Evans Hall, #3860, Berkeley, CA 94720-3860. Tel: (510) 642-2781 Fax: (510) 642-7892, E-mail: gyc@stat.berkeley.edu .

in which thousands of hypotheses are tested simultaneously. Although the methods described in this paper are applicable in any multiple testing situation, particular emphasis is placed on the use of adjusted p-values for the identification of differentially expressed genes in microarray experiments.

DNA microarrays are a new and promising biotechnology which allow the monitoring of expression levels in cells for thousands of genes simultaneously. Microarrays are being applied increasingly in biological and medical research to address a wide range of problems, such as the classification of tumors or the study of host genomic responses to bacterial infections (Alizadeh et al. (2000); Alon et al. (1999); Boldrick et al. (2002); Golub et al. (1999); Perou et al. (1999); Pollack et al. (1999); Ross et al. (2000)). An important and common aim in microarray experiments is the identification of differentially expressed genes, i.e. of genes whose expression levels are associated with a response or covariate of interest. The covariates could be either polytomous (e.g. treatment/control status, cell type, drug type) or continuous (e.g. dose of a drug, time), and the responses could be, for example, censored survival times or other clinical outcomes. There are two issues in identifying differentially expressed genes: (a) from the biological viewpoint, the interest is simply to decide which genes are differentially expressed, while (b) from a statistical perspective, we might wish to quantify in some probabilistic manner the evidence concerning the possible differential expression of the genes.

Issue (a) can be addressed satisfactorily by ranking the genes using a suitable univariate test statistic or the associated p-values. Then the biologist can examine the genes in the top positions to decide whether they really are differentially expressed, using more accurate low-throughput experiments such as northern blots or one of the quantitative PCR-based techniques. The number of genes that can be investigated in this follow-up phase depends on the background and the aims of the experiment, and on the level of effort the investigator is willing to expend. However, some biologists may want a quantitative assessment of the likely differential expression of each gene, so that they do not have to follow-up genes with little prospect of being truly differentially expressed. To address this need, we consider the statistical issue (b). It can be addressed through multiple hypothesis testing, by carrying out a simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. Since a typical microarray experiment measures expression levels for several thousand genes simultaneously, we are faced with an extreme multiple testing problem. In any such testing situation,

two types of errors can occur: a false positive, or type I error, is committed when a gene is declared to be differentially expressed when it is not, and a false negative, or type II error, is committed when the test fails to identify a truly differentially expressed gene. Special problems arising from the multiplicity aspect include defining an appropriate type I error rate, and devising powerful multiple testing procedures which control this error rate and incorporate the *joint* distribution of the test statistics.

Westfall and Young (1993) propose resampling-based p-value adjustment procedures which are highly relevant to microarray experiments. In particular, these authors define adjusted p-values for multiple testing procedures which control the family-wise error rate and take into account the dependence structure between test statistics. However, due to the very large number of hypotheses in current applications, computational issues remain to be addressed. The present paper introduces a new algorithm for computing the Westfall and Young (1993) step-down minP adjusted p-values. A second line of multiple testing is developed by Benjamini and Hochberg (1995). They propose procedures to control the false discovery rate. This was further developed by Storey (2002a) with a new concept called *positive* false discovery rate, which has a Bayesian motivation.

Section 2 reviews the basic notions of multiple testing and discusses different criteria for controlling type I error rates. Section 3 presents procedures based on adjusted p-values to control family-wise error rates. Section 4 presents resampling algorithms for estimating the adjusted p-values of Section 3 and introduces a fast algorithm for computing the Westfall and Young (1993) step-down minP adjusted p-values. Section 5 presents procedures based on FDR adjusted p-values and the pFDR-based q-values. The multiple testing procedures of Sections 3, 4 and 5 are applied to gene expression data from two recently published microarray studies described in Section 6. The results from the studies are discussed in Section 7, and finally, Section 8 summarizes our findings and outlines open questions.

2 Multiple testing and adjusted p-values

2.1 Multiple testing in microarray experiments

Suppose we have microarray experiments which produce expression data on m genes (or variables) for n samples (corresponding to n individual

microarray experiments). Let the gene expression levels be arrayed as an $m \times n$ matrix $X = (x_{ij})$, with rows corresponding to genes and columns to individual microarray experiments [1]. In most cases, the additional data for sample j consists of one or more responses or covariates y_j. The gene expression levels x_{ij} might be either absolute (e.g. Affymetrix oligonucleotide arrays (Lockhart et al., 1996)) or relative with respect to the expression levels of a suitably defined common reference sample (e.g. two-color cDNA microarrays (DeRisi et al., 1997)). The y_j could be either polytomous or continuous. In the simplest case, the n samples would consist of n_1 control samples and n_2 treatment samples, in which case y_j would be treatment status (treatment or control). In the Apo AI experiment (Callow et al. (2000)), $m = 6{,}356, n_1 = n_2 = 8$ so that $n = n_1 + n_2 = 16$. This dataset will be described in Section 6.1. Let X_i denote the random variable corresponding to the expression level for gene i and let Y denote the response or covariate. If a single test is considered for each gene (variable), the null hypothesis for testing that the gene is not differentially expressed between the treatment and the control can be stated as:

$$H_i : \text{There is no association between } X_i \text{ and } Y.$$

If each H_i is tested separately, then nothing more than univariate hypothesis testing is needed. This kind of testing has been studied extensively in the statistical literature. In general, the appropriate test statistic for each gene will depend on the experimental design, the type of response or covariate and the alternative hypothesis in mind. For example, for binary covariates one might consider t- or Mann-Whitney statistics, for polytomous covariates one might use an F-statistic, and for survival data one might rely on the score statistic for the Cox proportional hazard model. We will not discuss the choice of statistic any further here, except to say that for each gene i the null hypothesis H_i will be tested using a statistic T_i, and t_i will denote a realization of the random variable T_i. To simplify matters, we further assume that the null H_i is rejected for large values of $|T_i|$, *i.e.* this will be a two-sided test. Our two examples both involve two-sample t-statistics, but the extensions to other statistics should be clear.

When testing $H_i, i = 1, \ldots, m$ simultaneously, we want to reject hy-

[1] Note that this gene expression data matrix is the transpose of the standard $n \times m$ design matrix. The $m \times n$ representation was adopted in the microarray literature for display purposes, since for very large m and small n it is easier to display an $m \times n$ matrix than an $n \times m$ matrix.

index	2271	5709	5622	4521	3156	5898	2164	5930	2427	5694
t-stat	4.93	4.82	-4.62	4.34	-4.31	-4.29	-3.98	3.91	-3.90	-3.88
p-value	0.0002	0.0003	0.0004	0.0007	0.0007	0.0007	0.0014	0.0016	0.0016	0.0017

Table 1: *The simulated results for 6000 independent not differentially expressed genes.*

potheses while controlling a suitably defined type I error rate (Dudoit et al. (2002b); Efron et al. (2000, 2001); Golub et al. (1999); Kerr et al. (2000); Manduchi et al. (2000); Tusher et al. (2001); Westfall et al. (2001)). Multiple testing is the subject of the present paper. Although this is by no means a new subject in the statistical literature, microarray experiments are a new and challenging area of application for multiple testing procedures because of the sheer number of comparisons involved.

Before moving on to the multiple testing problem, we summarize the results of a simple simulation based on the microarray experiments in Callow et al. (2000). Suppose that the elements of the array x_{ij} are independently and identically distributed $N(0,1)$, $i = 1, \ldots, 6000$, $j = 1, \ldots, 16$. Regard the first 8 columns of this array as corresponding to treatment units and the second 8 columns as corresponding to control units, just as in Callow et al. (2000). Table 1 lists the 10 genes with the largest two-sample t-statistics in absolute values. This table has three rows, the first giving the gene indices, ranging from 1 to 6000, the second giving the two-sample t-statistics, while the last row has the raw (*i.e.* unadjusted) p-values computed by the resampling algorithm described in Section 4. This table suggests that we cannot use the conventional 0.05 or 0.01 thresholds for p-values to find significantly differentially expressed genes, since by our simulation, the data have no genes differentially expressed between the treatment and control. Indeed, if the 0.05 threshold is used, about $6000 \times 0.05 = 300$ genes would be found differentially expressed, which would be quite misleading. We conclude that when testing thousands of genes, the use of conventional thresholds for p-values is inappropriate. The framework of multiple testing seeks to give guidance concerning what might be appropriate in such situations. In the remainder of this section, we review the basic notions and approaches to multiple testing.

	# not rejected	# rejected	
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	m_1
	W	R	m

Table 2: *Summary table for the multiple testing problem, based on Table 1 of Benjamini and Hochberg (1995).*

2.2 Type I error rates

Set-up. Consider the problem of simultaneously testing m null hypotheses H_i, $i = 1, \ldots, m$. Let $H_i = 0$ when the null hypothesis H_i is true, and $H_i = 1$ otherwise. In the frequentist setting, the situation can be summarized by Table 2, based on Table 1 of Benjamini and Hochberg (1995). The m specific hypotheses are assumed to be known in advance, and the sets $\mathcal{M}_0 = \{i : H_i = 0\}$ and $\mathcal{M}_1 = \{i : H_i = 1\}$ of true and false null hypotheses are unknown parameters, $m_0 = |\mathcal{M}_0|$, $m_1 = |\mathcal{M}_1|$. Note the complete set as $\mathcal{M} = \{1, 2, \cdots, m\} = \mathcal{M}_0 \cup \mathcal{M}_1$. The number R of rejected null hypotheses and $W = m - R$ are observable random variables, while S, T, U, and V in the table are unobservable random variables. In the microarray context, there is a null hypothesis H_i for each gene i and rejection of H_i corresponds to declaring that gene i is differentially expressed, in some suitable sense. In general, we would like to minimize the number V of *false positives*, or *type I errors*, and the number T of *false negatives*, or *type II errors*. The standard approach is to prespecify an acceptable type I error rate α and seek tests which minimize the type II error rate, i.e., maximize *power*, within the class of tests with type I error rate α.

Type I error rates. When testing a single hypothesis, H, say, the probability of a type I error, i.e., of rejecting the null hypothesis when it is true, is usually controlled at some designated level α. This can be achieved by choosing a critical value c_α such that $\Pr(|T| > c_\alpha \mid H = 0) \leq \alpha$ and reject-

ing H when $|T| > c_\alpha$. A variety of generalizations of type I error rates to the multiple testing situation are possible.

- *Per-comparison error rate* (PCER). The PCER is defined as the expected value of (number of type I errors/number of hypotheses), i.e.,

$$\text{PCER} = E(V)/m.$$

- *Per-family error rate* (PFER). Not really a rate, the PFER is defined as the expected number of type I errors, i.e.,

$$\text{PFER} = E(V).$$

- *Family-wise error rate* (FWER). The FWER is defined as the probability of at least one type I error, i.e.,

$$\text{FWER} = \Pr(V > 0).$$

- *False discovery rate* (FDR). The most natural way to define FDR would be $E(V/R)$, the expected proportion of type I errors among the rejected hypotheses. However, different methods of handling the case $R = 0$ lead to different definitions. Putting $V/R = 0$ when $R = 0$ gives the FDR definition of Benjamini and Hochberg (1995), i.e.,

$$\text{FDR} = E\left[\frac{V}{R}1_{\{R>0\}}\right] = E\left[\frac{V}{R} \mid R > 0\right]\Pr(R > 0).$$

  When $m = m_0$, it is easy to see that FDR = FWER.

- *Positive false discovery rate* (pFDR). If we are only interested in estimating an error rate when positive findings have occurred, then the pFDR of Storey (2002a) is appropriate. It is defined as the conditional expectation of the proportion of type I errors among the rejected hypotheses, given that at least one hypothesis is rejected,

$$\text{pFDR} = E\left[\frac{V}{R} \mid R > 0\right].$$

  Storey (2002a) shows that this definition is intuitively pleasing and has a nice Bayesian interpretation (cf. the remarks on page 11) below.

Comparison of type I error rates. Given the same multiple testing procedure, i.e. the same rejection region in the m-dimensional space of $(T_1, T_2, \ldots, T_m)$, it is easy to prove that

$$\begin{aligned} \text{PCER} &\leq \text{FDR} \leq \text{FWER} \leq \text{PFER}, \\ \text{FDR} &\leq \text{pFDR}. \end{aligned}$$

First, note that $0 \leq V \leq R \leq m$ and that $R = 0$ implies $V = 0$, whence

$$\frac{V}{m} \leq \frac{V}{R} 1_{\{R>0\}} \leq 1_{\{V>0\}} \leq V.$$

Taking expectations of the above proves these assertions. It is more difficult to describe the relations between pFDR and FWER. In microarray applications, we expect pFDR $\leq$ FWER, apart from the extreme case when $1 = \text{pFDR} \geq \text{FDR} = \text{FWER}$ when $m_0 = m$. This is unlikely to be the case with microarray experiments as it is generally expected that at least one gene will be differentially expressed. Also $\Pr(R > 0) \to 1$ as $m \to \infty$, in which case pFDR is identical to FDR. Therefore we expect the following inequality to hold generally,

$$\text{PCER} \leq \text{FDR} \leq \text{pFDR} \leq \text{FWER} \leq \text{PFER}. \tag{2.1}$$

Exact control, weak control and strong control. It is important to note that the expectations and probabilities above are *conditional* on the *true* hypothesis $H_{\mathcal{M}_0} = \bigcap_{i \in \mathcal{M}_0} \{H_i = 0\}$. Controlling an error rate in this case will be called *exact* control. For the FWER, exact control means control of $\Pr(V > 0 \mid H_{\mathcal{M}_0})$. Since the set $\mathcal{M}_0$ is unknown, in general, we turn to computing the error rate when all null hypotheses are true, i.e., under the *complete null* hypothesis $H_{\mathcal{M}} = \cap_{i=1}^{m} \{H_i = 0\}$, equivalently, when $m_0 = m$ or $\mathcal{M}_0 = \mathcal{M}$. Controlling an error rate under $H_{\mathcal{M}}$ is called *weak* control. For the FWER, weak control means control of $\Pr(V > 0 \mid H_{\mathcal{M}})$. *Strong* control means control for every possible choice $\mathcal{M}_0$. For the FWER, it means control of $\max_{\mathcal{M}_0 \subseteq \{1,\ldots,m\}} \Pr(V > 0 \mid H_{\mathcal{M}_0})$. In general, strong control implies exact control and weak control, but neither of weak control and exact control implies the other. In the microarray setting, where it is very unlikely that none of the genes is differentially expressed, it seems that weak control without any other safeguards is unsatisfactory, and that it is important to have exact or strong control of type I error rates. The advantage of exact control is higher power.

2.3 Adjusted p-values and q-values

Raw p-values. Consider first the test of a single hypothesis H with nested level α rejection regions Γ_α such that (a) $\Gamma_{\alpha_1} \subseteq \Gamma_{\alpha_2}$, for $0 \leq \alpha_1 \leq \alpha_2 \leq 1$, and (b) $\Pr(T \in \Gamma_\alpha \mid H = 0) \leq \alpha$, for $0 \leq \alpha \leq 1$. If we are interested in using the statistic $|T|$ to carry out a two-sided test, the nested rejection regions $\Gamma_\alpha = [-\infty, -c_\alpha] \cup [c_\alpha, \infty]$ are such that $\Pr(T \in \Gamma_\alpha \mid H = 0) = \alpha$. The p-value for the observed value $T = t$ is

$$p\text{-value}(t) = \min_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \Pr(T \in \Gamma_\alpha \mid H = 0). \tag{2.2}$$

In words, the p-value is the minimum type I error rate over all possible rejection regions Γ_α containing the observed value $T = t$. For a two sided test, $p\text{-value}(t) = \Pr(|T| \geq |t| \mid H = 0) = p$, say. The smaller the p-value p, the stronger the evidence against the null hypothesis H. Rejecting H when $p \leq \alpha$ provides control of the type I error rate at level α. The p-value can also be thought of as the level of the test at which the hypothesis H would just be rejected. Extending this concept to the multiple testing situation leads to the very useful definition of adjusted p-value. In what follows we will call the traditional (unadjusted) p-value associated with a univariate test a *raw* p-value.

Adjusted p-values. Let t_i and $p_i = \Pr(|T_i| \geq |t_i| \mid H_i = 0)$ denote respectively the test statistic and p-value for hypothesis H_i (gene i), $i = 1, \ldots, m$. Just as in the single hypothesis case, a multiple testing procedure may be defined in terms of critical values for the test statistics or the p-values of individual hypotheses: e.g. reject H_i if $|t_i| > c_i$ or if $p_i \leq \alpha_i$, where the critical values c_i or α_i are chosen to control a given type I error rate (FWER, PCER, PFER, or FDR) at a prespecified level α. Alternately, the multiple testing procedure may be defined in terms of adjusted p-values. Given any test procedure, the *adjusted p-value* corresponding to the test of a single hypothesis H_i can be defined as the level of the entire test procedure at which H_i would just be rejected, given the values of all test statistics involved (Shaffer (1995); Westfall and Young (1993); Yekutieli and Benjamini (1999)). If interest is in controlling the FWER, the FWER adjusted p-value for hypothesis H_i is:

$$\tilde{p}_i = \inf \{\alpha : H_i \text{ is rejected at FWER} = \alpha\}.$$

Hypothesis H_i is then rejected, i.e., gene i is declared differentially expressed, at FWER α if $\tilde{p}_i \leq \alpha$. Note that this definition is dependent on the rejection procedure used. If that procedure is very conservative, such as the classical Bonferroni procedure, then the corresponding adjusted p-values will also be very conservative. For the stepwise procedures to be discussed in Section 3 and Section 5.1, the adjusted p-value for gene i depends on not only the magnitude of the statistic T_i, but also on the rank of gene i among all the genes. Adjusted p-values for other type I error rates are defined similarly (Yekutieli and Benjamini (1999)), *e.g.*

$$\tilde{p}_i = \inf \{\alpha : H_i \text{ is rejected at FDR} = \alpha\}.$$

As in the single hypothesis case, an advantage of reporting adjusted p-values, as opposed to only rejection or not of the hypotheses, is that the level of the test does not need to be determined in advance. Some multiple testing procedures are most conveniently described in terms of their adjusted p-values, and for many these can in turn be determined easily using resampling methods.

q-values. The positive false discovery rate pFDR cannot be strongly controlled in the traditional sense as $\text{pFDR} = E(V/R \mid R > 0) = 1$ when $m_0 = m$. However, an analogue of adjusted p-value termed the q-value can be defined in this context, although we emphasize that Storey (2001) does not view it as a form of adjusted p-value. The notion of q-value is approached by recalling the definition of p-value in equation (2.2), considering the minimum of the type I error rates for all possible rejection regions Γ_α containing the observed $T = t$. Let $\text{pFDR}(\Gamma_\alpha)$ be the pFDR when each hypothesis is rejected by the same rejection region Γ_α. The *q-value* is defined analogously, namely

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \text{pFDR}(\Gamma_\alpha). \tag{2.3}$$

Note that the above definition requires the T_i to be identically distributed across genes. Alternatively, if observed p-values are used to reject the test, then the nested rejection region $\Gamma_\gamma = [0, \gamma]$, abbreviated by γ leads to

$$q\text{-value}(p) = \inf_{\{\gamma \geq p\}} \text{pFDR}(\gamma). \tag{2.4}$$

Remark 2.1. Firstly, no procedures can give strong or weak control for pFDR, as pFDR=1 when $m_0 = m$. However, $m_0 = m$ is extremely unlikely

with microarray data, and pFDR can be conservatively estimated under the unknown true hypothesis $H_{\mathcal{M}_0}$. One such method doing so will be given in Section 5.2. The use of q-value(p) provides a way to adjust p-values under $H_{\mathcal{M}_0}$ which leads to control of pFDR.

Secondly, Storey (2001) argues that a q-value is not a "pFDR adjusted p-value". This is because adjusted p-values are defined in terms of *a particular procedure*, i.e. a sequential p-value method, such as those to be discussed in Section 3 and Section 5.1, while pFDR can not be controlled by such procedure. Our view is that q-value(p) gives us the minimum pFDR that we can achieve when rejecting H_j whenever $p_j \leq p, j = 1, \ldots, m$. Therefore q-values are analogous to the single step adjustments for controlling FWER to be discussed in Section 3.1. Indeed, the notion of q-value is similar to the concept of "p-value correction" in Yekutieli and Benjamini (1999). The only difference between q-values and single step adjusted p-values is that q-values consider only the true but unknown $\mathcal{M}_0$ (exact control), while single step adjustments consider every possible choice of $\mathcal{M}_0, \mathcal{M}_0 \subseteq \{1, 2, \ldots, m\}$ (strong control). In what follows, we will use the terms q-value and adjusted p-value interchangeably for pFDR.

The q-value definition has an appealing Bayesian interpretation. Suppose that the $T_i \mid H_i$ are independently distributed as $(1 - H_i) \cdot F_0 + H_i \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and that the H_i are independently and identically distributed $Bernoulli(\pi_1)$, where $\pi_1 = 1 - \pi_0$, π_0 being the *a priori* probability that a hypothesis is true. Theorem 1 of Storey (2001) states that for all i

$$\text{pFDR}(\Gamma_\alpha) = \Pr(H_i = 0 \mid T_i \in \Gamma_\alpha). \tag{2.5}$$

Since the left-hand side does not depend on i, we drop it from the right hand side. Using the definition of q-value,

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \Pr(H = 0 \mid T \in \Gamma_\alpha).$$

Comparing this formula to the one for p-value(t) given in equation (2.2), it can be seen that the difference between a p-value and a q-value is that the role of $H = 0$ and $T \in \Gamma_\alpha$ have been switched. The q-values are thus Bayesian version of p-values, analogous to the "Bayesian posterior p-values" of Morton (1955). Details of a Bayesian interpretation can be found in Storey (2001).

3 Procedures controlling the family-wise error rate

There are three distinct classes of multiple testing procedures commonly used in the literature: single-step, step-down and step-up procedures. In *single-step* procedures, equivalent multiplicity adjustments are performed for all hypotheses, regardless of the ordering of the test statistics or raw p-values. Improvements in power, while preserving type I error rate control, may be achieved by *stepwise procedures*, in which rejection of a particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the tests of other hypotheses. *Step-down* procedures order the raw p-values (or test statistics) starting with the most significant, while *step-up* procedures start with the least significant.

3.1 Single-step procedures

For strong control of the FWER at level α, the Bonferroni procedure, perhaps the best known in multiple testing, rejects any hypothesis H_i with p-value less than or equal to α/m. The corresponding *Bonferroni single-step adjusted p-values* are thus given by

$$\tilde{p}_i = \min\bigl(mp_i, 1\bigr). \tag{3.1}$$

Control of the FWER in the strong sense follows from Boole's inequality, where the probabilities in what follows are conditional on $H_{\mathcal{M}_0} = \bigcap_{i\in\mathcal{M}_0}\{H_i = 0\}$.

$$\begin{aligned} \text{FWER} &= \Pr(V > 0) \leq \Pr\left[\bigcup_{i=1}^{m_0}\{\tilde{P}_i \leq \alpha\}\right] \leq \sum_{i=1}^{m_0}\Pr\bigl(\tilde{P}_i \leq \alpha\bigr) \\ &\leq \sum_{i=1}^{m_0}\alpha/m = m_0\alpha/m \leq \alpha. \end{aligned} \tag{3.2}$$

Bonferroni-adjusted p-values are not, strictly, adjusted p-values in the sense of the definition given earlier. Rather, they are conservative lower bounds to adjusted p-values which are difficult if not impossible to calculate without further assumptions. Closely related to the Bonferroni procedure is the Šidák procedure which is exact for protecting the FWER when the raw p-values are independently and uniformly distributed over $[0, 1]$. By a simple computation, the *Šidák single-step adjusted p-values* are given by

$$\tilde{p}_i = 1 - (1 - p_i)^m. \tag{3.3}$$

We sketch the easy proof that this procedure provides strong control. Note that

$$\begin{aligned} \Pr(V=0) &= \Pr(\bigcap_{i=1}^{m_0}\{\tilde{P}_i \geq \alpha\}) = \prod_{i=1}^{m_0}\Pr(\tilde{P}_i \geq \alpha) \\ &= \prod_{i=1}^{m_0}\Pr(P_i \geq 1-(1-\alpha)^{1/m}) = \{(1-\alpha)^{1/m}\}^{m_0}. \quad (3.4) \end{aligned}$$

Therefore,

$$\text{FWER} = \Pr(V>0) = 1-\Pr(V=0) = 1-(1-\alpha)^{m_0/m} \leq \alpha. \quad (3.5)$$

In many situations, the test statistics and hence the p-values are correlated. This is the case in microarray experiments, where groups of genes tend to have highly correlated expression levels due to co-regulation. Westfall and Young (1993) propose adjusted p-values for less conservative multiple testing procedures which take into account the dependence structure between test statistics. Their *single-step minP adjusted p-values* are defined by

$$\tilde{p}_i = \Pr\left(\min_{1\leq l\leq m} P_l \leq p_i \mid H_{\mathcal{M}}\right), \quad (3.6)$$

where $H_{\mathcal{M}}$ denotes the complete null hypothesis and P_l the random variable for the raw p-value of the lth hypothesis. Alternately, we may consider procedures based on the *single-step maxT adjusted p-values* which are defined in terms of the test statistics T_i themselves, namely

$$\tilde{p}_i = \Pr\left(\max_{1\leq l\leq m} |T_l| \geq |t_i| \mid H_{\mathcal{M}}\right). \quad (3.7)$$

The following points should be noted regarding these four procedures.

1. If the raw p-values $P_1,\ldots,P_m$ are independent, the minP adjusted p-values are the same as the Šidák adjusted p-values.

2. The Šidák procedure does not guarantee control of the FWER for arbitrary distributions of the test statistics, but it does control the FWER for test statistics that satisfy an inequality known as Šidák's inequality: $\Pr(|T_1| \leq c_1,\ldots,|T_m| \leq c_m) \geq \prod_{i=1}^{m}\Pr(|T_i| \leq c_i)$. This

inequality was initially derived by Dunn (1958) for $(T_1, \ldots, T_m)$ having a multivariate normal distribution with mean zero and certain types of covariance matrix. Šidák (1967) extended the result to arbitrary covariance matrices, and Jogdeo (1977) showed that the inequality holds for a larger class of distributions, including the multivariate t- and F-distributions. When the Šidák inequality holds, the minP adjusted p-values are less than the Šidák adjusted p-values.

3. Computing the quantities in equation (3.6) under the assumption that $P_l \sim U[0,1]$ and using the upper bound provided by Boole's inequality yields the Bonferroni p-values. In other words, procedures based on minP adjusted p-values are less conservative than the Bonferroni or Šidák (under the Šidák inequality) procedures. Again, in the case of independent test statistics, the Šidák and minP adjustments are equivalent.

4. Procedures based on the maxT and minP adjusted p-values control the FWER weakly under all conditions. Strong control of the FWER also holds under the assumption of subset pivotality (Westfall and Young, 1993, p. 42). The distribution of raw p-values $(P_1, \ldots, P_m)$ is said to have the *subset pivotality* property if for all subsets $\mathcal{K}$ of $\{1, \ldots, m\}$ the joint distributions of the sub-vector $\{P_i : i \in \mathcal{K}\}$ are identical under the restrictions $H_{\mathcal{K}} = \cap_{i \in \mathcal{K}} \{H_i = 0\}$ and $H_{\mathcal{M}} = \cap_{i=1}^{m} \{H_i = 0\}$. This property is required to ensure that procedure based on adjusted p-values computed under the complete null provide strong control of the FWER. A practical consequence of it is that resampling for computing adjusted p-values may be done under the complete null $H_{\mathcal{M}}$ rather than the unknown partial null hypotheses $H_{\mathcal{M}_0}$. For the problem of identifying differentially expressed considered in this article, the subset pivotality property is usually satisfied. Here is the proof. Let T_i be the statistic for gene i, *e.g.* the two-sample t-statistic or one of the other statistics defined in Section 8. For any subset $\mathcal{K} = \{i_1, i_2, \cdots, i_k\}$, let its complement set be $\{j_1, j_2, \cdots, j_{m-k}\}$. Since T_i is computed only from the data on gene i (the i-th row of the data matrix X), and not from any data from other genes, the joint distribution of $(T_{i_1}, T_{i_2}, \cdots, T_{i_k})$ is not going to depend on $(H_{j_1}, H_{j_2}, \cdots, H_{j_{m-k}})$ given the same specification of $(H_{i_1}, H_{i_2}, \cdots, H_{i_k})$. This proves subset pivotality.

5. The maxT p-values are easier to compute than the minP p-values, and

are equal to the minP p-values when the test statistics T_i are identically distributed. However, the two procedures generally produce different adjusted p-values, and considerations of balance, power, and computational feasibility should dictate the choice between the two approaches. When the test statistics T_i are not identically distributed (e.g. t-statistics with different degrees of freedom), not all tests contribute equally to the maxT adjusted p-values and this can lead to unbalanced adjustments (Beran 1988, Westfall and Young 1993, p. 50). When adjusted p-values are estimated by permutation (Section 4) and a large number of hypotheses are tested, procedures based on the minP p-values tend to be more sensitive to the number of permutations and more conservative than those based on the maxT p-values. Also, the minP p-values require more computation than the maxT p-values, because the raw p-values must be computed before considering the distribution of their successive minima.

3.2 Step-down procedures

While single-step procedures are simple to implement, they tend to be conservative for control of the FWER. Improvement in power, while preserving strong control of the FWER, may be achieved by step-down procedures. Below are the step-down analogues, in terms of their adjusted p-values, of the four procedures described in the previous section. Let $p_{r_1} \leq p_{r_2} \leq ... \leq p_{r_m}$ denote the *ordered raw p-values*. For control of the FWER at level α, the Holm (1979) procedure proceeds as follows. Starting from $i = 1$, then $i = 2$, until $i = m$, let i^* be the first integer i such that $p_{r_i} > \frac{\alpha}{m-i+1}$. If no such i^* exists, reject all hypotheses; otherwise, reject hypotheses H_{r_i} for $i = 1, \ldots, i^* - 1$. The *Holm step-down adjusted p-values* are thus given by

$$\tilde{p}_{r_i} = \max_{k=1,\ldots,i} \left\{ \min\big((m-k+1)\, p_{r_k}, 1\big) \right\}. \tag{3.8}$$

Holm's procedure is less conservative than the standard Bonferroni procedure, which would multiply the p-values by m at each step. Note that taking successive maxima of the quantities $\min\big((m-k+1)\, p_{r_k}, 1\big)$ enforces monotonicity of the adjusted p-values. That is, $\tilde{p}_{r_1} \leq \tilde{p}_{r_2} \leq ... \leq \tilde{p}_{r_m}$, and one can only reject a particular hypothesis provided all hypotheses with smaller raw p-values were rejected beforehand. Similarly, the *Šidák step-down adjusted p-values* are defined as

$$\tilde{p}_{r_i} = \max_{k=1,\ldots,i} \left\{ 1 - (1 - p_{r_k})^{(m-k+1)} \right\}. \tag{3.9}$$

The Westfall and Young (1993) *step-down minP adjusted p-values* are defined by

$$\tilde{p}_{r_i} = \max_{k=1,\ldots,i} \left\{ \Pr\Big(\min_{l=k,\ldots,m} P_{r_l} \leq p_{r_k} \mid H_{\mathcal{M}} \Big) \right\}, \tag{3.10}$$

and the *step-down maxT adjusted p-values* are defined by

$$\tilde{p}_{s_i} = \max_{k=1,\ldots,i} \left\{ \Pr\Big(\max_{l=k,\ldots,m} |T_{s_l}| \geq |t_{s_k}| \mid H_{\mathcal{M}} \Big) \right\}, \tag{3.11}$$

where $|t_{s_1}| \geq |t_{s_2}| \geq \ldots \geq |t_{s_m}|$ denote the ordered test statistics.

Note that computing the quantities in (3.10) under the assumption that the P_i are uniformly distributed on the interval [0,1], and using the upper bound provided by Boole's inequality, we obtain Holm's p-values. Procedures based on the step-down minP adjusted p-values are thus less conservative than Holm's procedure. For a proof of strong control of the FWER for the maxT and minP procedures assuming subset pivotality we refer the reader to Westfall and Young (1993, Section 2.8).

4 Resampling algorithms to control FWER

In many situations, the joint (and marginal) distribution of the test statistics is unknown. Bootstrap or permutation resampling can be used to estimate raw and adjusted p-values while avoiding parametric assumptions about the joint distribution of the test statistics. In the microarray setting, the joint distribution under the complete null hypothesis of the test statistics $T_1, \ldots, T_m$ can be estimated by permuting the columns of the gene expression data matrix X. Permuting entire columns of this matrix creates a situation in which the response or covariate Y is independent of the gene expression levels, while preserving the correlation structure and distributional characteristics of the gene expression levels. Depending on the sample size n it may be infeasible to consider all possible permutations, in which case a random subset of B permutations (including the observed) is considered. The manner in which the responses/covariates are permuted depends on the experimental design. For example, with a two-factor design, one can permute the levels of the factor of interest within the levels of

Box 1. Permutation algorithm for raw p-values
For the bth permutation, $b = 1, \ldots, B$:

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.

After the B permutations are done, for two-sided alternative hypotheses, the permutation p-value for hypothesis H_i is

$$p_i^* = \frac{\#\{b : |t_{i,b}| \geq |t_i|\}}{B} \qquad \text{for } i = 1, \ldots, m.$$

the other factor. Next, we present permutation algorithms for estimating adjusted p-values.

4.1 Raw p-values

Box 1 describes how to compute raw p-values from permutations. Permutation adjusted p-values for the Bonferroni, Šidák and Holm procedures can then be obtained by replacing p_i by p_i^* in equations (3.1), (3.3), (3.8), and (3.9).

4.2 Step-down maxT adjusted p-values

For the step-down maxT adjusted p-values of Westfall and Young (1993), the null distribution of successive maxima $\max_{l=i,\ldots,m} |T_{s_l}|$ of the test statistics needs to be estimated. (The single-step case is simpler and omitted here as we only need the distribution of the $\max_{l=1,\ldots,m} |T_{s_l}|$.) The details of the algorithm are presented in Box 2.

4.3 The traditional double permutation algorithm for step-down minP adjusted p-values

The single-step and step-down minP adjusted p-values of Westfall and Young (1993) are in general harder to compute as they require the joint null

Box 2. Permutation algorithm for step-down maxT adjusted p-values - based on Westfall and Young (1993, Algorithm 4.1, p. 116–117)

For the original data, order the observed test statistics such that $|t_{s_1}| \geq |t_{s_2}| \geq ... \geq |t_{s_m}|$. For the bth permutation, $b = 1, \ldots, B$:

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.
3. Next, compute $u_{i,b} = \max_{l=i,\ldots,m} |t_{s_l,b}|$ (see equation (3.11)), the successive maxima of test statistics by

$$\begin{aligned} u_{m,b} &= |t_{s_m,b}| \\ u_{i,b} &= \max\Big(u_{i+1,b}, |t_{s_i,b}|\Big) \qquad \text{for } i = m-1, \ldots, 1. \end{aligned}$$

The above steps are repeated B times and the adjusted p-values are estimated by

$$\tilde{p}^*_{s_i} = \frac{\#\{b : u_{i,b} \geq |t_{s_i}|\}}{B} \qquad \text{for } i = 1, \ldots, m$$

with the monotonicity constraints enforced by setting

$$\tilde{p}^*_{s_1} \leftarrow \tilde{p}^*_{s_1}, \qquad \tilde{p}^*_{s_i} \leftarrow \max\big(\tilde{p}^*_{s_{i-1}}, \tilde{p}^*_{s_i}\big) \qquad \text{for } i = 2, \ldots, m.$$

distribution of $P_1, \ldots, P_m$. The traditional double permutation algorithm for computing these p-values is described in Box 3.

When the raw p-values themselves are unknown, additional resampling at step 2 for estimating these p-values can be computationally infeasible. This algorithm is called a *double permutation algorithm* because of the two rounds of resampling procedures. For a typical microarray experiment, such as the one described in Section 6.1, all possible $B = 12{,}870$ permutations are used to estimate raw and adjusted p-values for $m = 6{,}356$ genes. A double permutation algorithm would require $O(mB^2 + m \log m) \approx O(10^{12})$ computations (cf. Table 3 p. 24). As the time taken for generating one set of raw p-values for all genes is about 2 minutes, our estimate of the compu-

Box 3. The traditional double permutation algorithm for step-down minP adjusted p-values - based on Westfall and Young (1993, Algorithm 2.8, p. 66-67.)

For the original data, use the algorithm in Box 1 to compute the raw p-values $p_1^*, \ldots, p_m^*$ and then order the raw p-values such that $p_{r_1}^* \leq p_{r_2}^* \leq \cdots \leq p_{r_m}^*$.
For the bth permutation, $b = 1, \ldots, B$:

1. Permute the n columns of the data matrix X.
2. Compute raw p-values $p_{1,b}, \ldots, p_{m,b}$ for each hypothesis from the permuted data.
3. Next, compute $q_{i,b} = \min_{l=i,\ldots,m} p_{r_l,b}$ (see equation (3.10)), the successive minima of the raw p-values.

$$\begin{aligned} q_{m,b} &= p_{r_m,b} \\ q_{i,b} &= \min\Big(q_{i+1,b}, p_{r_i,b}\Big) \qquad \text{for } i = m-1, \ldots, 1. \end{aligned}$$

The above steps are repeated B times and the adjusted p-values are estimated by

$$\tilde{p}_{r_i}^* = \frac{\#\{b : q_{i,b} \leq p_{r_i}^*\}}{B} \qquad \text{for } i = 1, \ldots, m.$$

with the monotonicity constraints enforced by setting

$$\tilde{p}_{r_1}^* \leftarrow \tilde{p}_{r_1}^*, \qquad \tilde{p}_{r_i}^* \leftarrow \max\big(\tilde{p}_{r_{i-1}}^*, \tilde{p}_{r_i}^*\big) \qquad \text{for } i = 2, \ldots, m.$$

tation time for such an algorithm is approximately 400 hours (2×12,000/60) on a Sun 200Mhz Ultrasparc workstation,

One way around the computational problem is to turn to procedures based on maxT adjusted p-values, which may be estimated from a single permutation using the algorithm in Box 2. However, as mentioned in Section 2.3, if the test statistics are not identically distributed across hypotheses, the maxT adjusted p-values may be different from the minP

adjusted p-values, and may give different weights to different hypotheses. For example, if the test statistic T_i for one particular hypothesis H_i has a heavy-tailed distribution, it will tend to be larger than other test statistics and hence H_i will tend to have smaller adjusted p-value than other hypotheses. In such cases it will be better to compute minP rather than maxT adjusted p-values. We now present a new resampling algorithm for estimating minP adjusted p-values without the double resampling step of Box 3. Note that this algorithm produces the *same* p-values as the double permutation algorithm in Box 3.

4.4 A new algorithm for step-down minP adjusted p-values

This algorithm allows the minP adjusted p-values to be obtained within a single permutation analysis. The main idea is to proceed one hypothesis (gene) at a time, instead of one permutation at a time, and to compute the B raw p-values for each hypothesis by sorting the B test statistics using the quick sort algorithm. To see this, first compute the permutation raw p-values p_i^* and assume without loss of generality that $p_1^* \leq p_2^* \leq \cdots \leq p_m^*$. Consider the following three key $m \times B$ matrices: a matrix of test statistics

$$T = \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,b} & \cdots & t_{1,B} \\ \vdots & \vdots & & \vdots & & \vdots \\ t_{i,1} & t_{i,2} & \cdots & t_{i,b} & \cdots & t_{i,B} \\ \vdots & \vdots & & \vdots & & \vdots \\ t_{m,1} & t_{m,2} & \cdots & t_{m,b} & \cdots & t_{m,B} \end{bmatrix},$$

a matrix of raw p-values

$$P = \Big[\ p_{i,b}\ \Big],$$

and a matrix of minima of raw p-values

$$Q = \Big[\ q_{i,b}\ \Big],$$

where $q_{i,b} = \min_{l=i,\ldots,m} p_{l,b}$ and the bth column of these matrices corresponds to a data matrix X_b, say, with permuted columns. In this matrix representation, the double permutation algorithm in Box 3 would compute the columns of matrices T, P, and Q one at a time. The permutation p-values in column b of P would be obtained by considering B permutations

Box 4. A new permutation algorithm for step-down minP adjusted p-values

0. Compute raw p-values for each hypothesis. Assume $p_1^* \leq p_2^* \leq \cdots \leq p_m^*$ without loss of generality, otherwise sort the rows of the data matrix X according to the ordered p_i^*. Initialize $q_{m+1,b} = 1$ for $b = 1, \ldots, B$. Initialize $i = m$.

1. For hypothesis H_i (row i), compute the B permutation test statistics $t_{i,1}, \ldots, t_{i,B}$ and use the quick sort algorithm to get the B raw p-values $p_{i,1}, \ldots, p_{i,B}$ as in Section 4.4.1.

2. Update the successive minima $q_{i,b}$

$$q_{i,b} \leftarrow \min(q_{i+1,b},\, p_{i,b}), \quad b = 1, \ldots, B.$$

3. Compute the adjusted p-values for hypothesis H_i

$$\tilde{p}_i^* = \frac{\#\{b : q_{i,b} \leq p_i^*\}}{B}.$$

4. Delete $p_{i,1}, \ldots, p_{i,B}$ [row i of P].
   Delete $q_{i+1,1}, \ldots, q_{i+1,B}$ [row $i+1$ of Q].

5. Move up one row, i.e., $i \leftarrow i - 1$.
   If $i = 0$, go to step 6, otherwise, go to step 1.

6. Enforce monotonicity of $\tilde{p}_i^*$

$$\tilde{p}_1^* \leftarrow \tilde{p}_1^*, \qquad \tilde{p}_i^* \leftarrow \max\bigl(\tilde{p}_{i-1}^*, \tilde{p}_i^*\bigr) \qquad \text{for } i = 2, \ldots, m.$$

of the columns of X_b and computing the matrix T all over again (with different order of the columns). Our new algorithm computes the matrix T only once and deals with the rows of T, P, and Q sequentially, starting with the last.

4.4.1 Use of order statistics to compute the raw p-values

To avoid the double permutation for the algorithm in Box 3, one could compute each row of T, P, and Q as follows. From the permutation distribution of T_i, $t_{i,1}$, $t_{i,2}, \ldots, t_{i,B}$, obtain the permutation distribution of P_i, $p_{i,1}$, $p_{i,2}, \ldots, p_{i,B}$, simultaneously from

$$p_{i,b} = \frac{\#\{b' : |t_{i,b'}| \geq |t_{i,b}|\}}{B}. \tag{4.1}$$

Although this method avoids the double permutation of the algorithm in Box 3, the computational complexity is the same, as the computing of each raw p-value needs B computations from equation (4.1). However, the idea of computing $p_{i,1}$, $p_{i,2}, \ldots, p_{i,B}$ simultaneously can be refined as follows. Order the ith row of matrix T and let r_b, $b = 1, \ldots, B$, be such that $|t_{i,r_1}| \geq |t_{i,r_2}| \geq \cdots \geq |t_{i,r_B}|$. Note that the r_b will in general vary from row to row, not to be confused with our general notation for the rank indices of the raw p-values. In our new algorithm, the computational time for estimating the $p_{i,b}$ for each row is reduced by using the quick sort algorithm, which requires $O(B \log B)$ computations compared to $O(B^2)$ for a crude bubble sorting algorithm.

No ties. If there are no ties, the B raw p-values may be obtained from

$$p_{i,r_j} = \frac{j}{B} \text{ for } j = 1, \ldots, m.$$

Ties. With small modifications, ties may be handled as follows. Let the statistics $t_1, t_2, \cdots, t_m$ be ordered as

$$\begin{array}{ccccccc} |t_{i,r_1^1}| & = \cdots = & |t_{i,r_1^{k_1}}| & > & |t_{i,r_2^1}| & = \cdots = & |t_{i,r_2^{k_2}}| \\ & & & & \vdots & & \vdots \\ & & & > & |t_{i,r_J^1}| & = \cdots = & |t_{i,r_J^{k_J}}|. \end{array}$$

and $\sum_{j=1}^{J} k_j = B$. Note that k_j, J, and r_j^k will in general vary from row to row. Then the B raw p-values may be obtained from

$$p_{i,r_j^1} = \cdots = p_{i,r_j^{k_j}} = \frac{\sum_{l=1}^{j} k_l}{B}, \qquad j = 1, \ldots, J.$$

4.4.2 Storage

Storing the entire T, P, and Q matrices requires $O(Bm)$ memory, which in the Apo AI experiment of Section 6.1 corresponds to $O(12{,}780 \times 6{,}356)$, that is, about 284 Megabytes ($12{,}780 \times 6{,}356 \times 4$, as each number needs 4 bytes to store). However, for the proposed algorithm in Box 4, only individual rows of the T, P, and Q matrices are required at any given time. The storage requirements of the algorithm are thus $O(B)$ for rows of T, P, and Q and $O(m)$ for the raw p-values $p_1^* \leq p_2^* \leq \cdots \leq p_m^*$, the data matrix X (assuming the number of experiments, n, is small).

4.4.3 Further remarks

1. As with the double permutation algorithm in Box 3, the algorithm in Box 4 can be used for any type of test statistic (t-, F-statistics, etc.), and allows for different test statistics to be used for different hypotheses. The algorithm in Box 4 can also be modified easily for one-sided hypotheses.

2. The algorithm in Box 4 requires the same permutation order to be kept for each row. When all possible permutations are considered, the same enumeration can be used for computing each row. When a random subset of B permutations is used, the B permutations can be stored in a number of ways, including the following two.

(a) For each row, reset the random seed at the same fixed value, and use the same function to generate the B random permutations.

(b) For a k class problem, where $k \geq 2$, recode each permutation as an integer corresponding to the binary representation of the permutation. For example, for n_1 observations from class 1, n_2 observations from class 2, ..., n_k observations from class k, $n = n_1 + n_2 + \cdots + n_k$, any given permutation can be represented as an n-vector $\mathbf{a} = (a_1, \ldots, a_n)$, where $a_j = c - 1$ if sample j is assigned to class c (c is dependent on the sample j). The vector a can be mapped to an integer by $f(\mathbf{a}) = \sum_{j=1}^{n} k^{j-1} a_j$.

3. The storage space for individual rows of T, P, and Q is $O(B)$ and the storage space for strategy (b) in comment (2) is also $O(B)$.

In summary, the computational complexity of the new algorithm for minP adjusted p-values is given in Table 3.

Note that we did not consider n, the sample size (number of arrays), as

	Running time	Space
Double permutation algorithm	$O(mB^2 + m \log m)$	$O(m)$
New algorithm	$O(mB \log B + m \log m)$	$O(m + B)$

Table 3: *Computational complexity of double permutation algorithm and new minP algorithms. The number of hypotheses (genes) is denoted by m and the number of permutations by B.*

it is typically very small compared to m and B. Obviously, the maximum number of permutations B depends on n, for example in the two-class case $B = \frac{n!}{n_1!n_2!}$.

5 Procedures to control FDR or pFDR

Recall the notation for the different type I error rates and the two definitions of false discovery rates given in Section 2.2. The latter arise by treating V/R differently in estimating $E(V/R)$ when $R = 0$. Benjamini and Hochberg (1995) suppose that $V/R = 0$ when $R = 0$, while Storey (2002a) uses the conditional expectation of V/R given $R > 0$, termed the positive false discovery rate. Earlier ideas related to FDR can be found in Seeger (1968) and Sorić (1989).

5.1 Frequentist approach

5.1.1 FDR with independent null hypotheses

Benjamini and Hochberg (1995) (*BH*) derived a step-up procedure for strong control of the FDR for independent null p-values, although the independence assumption under the alternative hypothesis is not necessary. FDR is there defined as $E\left(\frac{V}{R}1_{\{R>0\}}\right)$. Under the complete null hypothesis, i.e. when $m_0 = m$, FDR is equal to FWER, and so a procedure controlling FDR also controls FWER in the weak sense. Using notation from Section 3, let the observed raw p-values be $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$. Starting from $i = m$, and then taking $i = m - 1$, etc., until $i = 1$ (the step-up order), define i^* be the first integer i such that $p_{r_i} \leq \frac{i}{m}\alpha$. If i^* is not defined, then reject no hypothesis; otherwise, reject hypotheses H_{r_i} for $i = 1, \ldots, i^*$.

As with the definition of FWER adjusted p-values, the adjusted p-value corresponding to the *BH procedure* is

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m} \left\{ \min \left(\frac{m}{k} p_{r_k}, 1 \right) \right\}. \tag{5.1}$$

Benjamini and Hochberg (1995) proved that under the conditions stated in the previous paragraph,

$$E\left(\frac{V}{R} 1_{\{R>0\}} \right) \leq \frac{m_0}{m} \alpha \leq \alpha. \tag{5.2}$$

When $m_0 = m$, this procedure provides weak control of the FWER. Indeed, exactly this weak control was shown in Seeger (1968). Simes (1986) rediscovered this approach and also gave the proof. The proof by Benjamini and Hochberg (1995) giving strong control of FDR greatly expanded the popularity of this procedure.

5.1.2 FDR under general dependence

Benjamini and Yekutieli (2001) (BY) proved that the procedure based on equation (5.1) controls FDR under certain more general assumptions (*positive regression dependency*). In addition, they proposed a simple conservative modification of the original *BH* procedure which controls FDR under arbitrary dependence. For control of the FDR at level α, going from $i = m$, $i = m-1, \ldots$, until $i = 1$, define i^* the first integer i such that $p_{r_i} \leq \frac{i}{m \sum_{l=1}^{m} 1/l} \alpha$. If no such i^* exists, then reject no hypothesis; otherwise, reject hypotheses H_{r_i} for $i = 1, \ldots, i^*$. The adjusted p-values for the *BY procedure* can be defined by

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m} \left\{ \min(\frac{m \sum_{l=1}^{m} 1/l}{k} p_{r_k}, 1) \right\}. \tag{5.3}$$

For a large number m of hypotheses, the penalty of the BY procedure is about $\log(m)$ in comparison with the BH procedure of equation (5.1). This can be a very large price to pay for allowing arbitrary dependence.

5.2 Bayesian motivation

5.2.1 pFDR under independence or special dependence

Storey (2002a) defined the pFDR as $E(\frac{V}{R} \mid R > 0)$. We need to estimate the pFDR in order to estimate the q-value, which we regard as the pFDR analogue of adjusted p-values. From equation (2.5), it is easy to see that

$$\text{pFDR}(p) = \frac{\pi_0 \cdot \Pr(P \leq p \mid H = 0)}{\Pr(P \leq p)} = \frac{\pi_0 p}{\Pr(P \leq p)}.$$

Since $m\pi_0$ of the p-values are expected to be null, π_0 can be estimated from the largest p-values, say those greater than some prespecified p_0. The value of p_0 can be chosen as the median of all p-values, or 1/2, or an optimized choice for p_0 can be made, see Storey and Tibshirani (2001) where the notation λ is used. Given a suitable p_0, a conservative estimate of π_0 will be

$$\hat{\pi}_0 = \frac{W(p_0)}{(1 - p_0)m},$$

where $W(p) = \#\{i : p_i > p\}$, and $\Pr(P \leq p)$ can be estimated by

$$\widehat{\Pr}(P \leq p) = \frac{R(p)}{m},$$

where $R(p) = \#\{i : p_i \leq p\}$.

Since pFDR is conditioned on $R > 0$, a conservative estimate of $\Pr(R > 0)$ when the rejection region is $[0, p]$ and the p-values are independent is

$$\widehat{\Pr}(R > 0) = 1 - (1 - p)^m.$$

It follows that an estimate of pFDR at $[0, p]$ is

$$\begin{aligned} \widehat{\text{pFDR}}_{p_0}(p) &= \frac{\hat{\pi}_0(p_0) \cdot p}{\widehat{\Pr}(P \leq p) \cdot \widehat{\Pr}(R > 0)} \\ &= \frac{W(p_0) \cdot p}{(1 - p_0) \cdot (R(p) \vee 1) \cdot (1 - (1 - p)^m)}. \end{aligned} \tag{5.4}$$

Dropping the estimate of $\Pr(R > 0)$, we can estimate the FDR at $[0, p]$ by

$$\widehat{\text{FDR}}_{p_0}(p) = \frac{W(p_0) \cdot p}{(1 - p_0) \cdot (R(p) \vee 1)}. \tag{5.5}$$

Note that these expressions are estimated under the assumptions that either the null P_i are independently and identically distributed, or that they satisfy a special dependence condition, see Storey (2002a) for full details.

5.2.2 pFDR under more general dependence

Storey and Tibshirani (2001) (*ST*) extend the foregoing to apply under more general dependence assumptions involving certain ergodic conditions. We just sketch the ideas of the extension and the algorithm here, referring readers to the paper for fuller details.

First, equation (5.4) can also be written in terms of a general family of nested rejection regions $\{\Gamma_\alpha\}$ as

$$\begin{aligned}\widehat{\mathrm{pFDR}}_{\Gamma_{\alpha_0}}(\Gamma_\alpha) &= \frac{\hat{\pi}_0(\Gamma_{\alpha_0})\cdot\alpha}{\widehat{\Pr}(T\in\Gamma_\alpha)\cdot\widehat{\Pr}(R>0)} \\ &= \frac{W(\Gamma_{\alpha_0})\cdot\alpha}{(1-\alpha_0)\cdot(R(\Gamma_\alpha)\vee 1)\cdot\widehat{\Pr}(R>0)},\end{aligned} \tag{5.6}$$

where $R(\Gamma)=\#\{i: T_i\in\Gamma\}$ and $W(\Gamma)=\#\{i: T_i\notin\Gamma\}=m-W(\Gamma)$.

Note that the term $\widehat{\Pr}(R>0)$ is still retained. In this equation Γ_α is the level α rejection region. Now consider a general rejection region Γ, for example $[-\infty,-c]\cup[c,\infty]$ for a two-sided alternative, and let us estimate an analogue of the preceding formula by resampling. Take a region Γ_0 which is believed to contain mostly null hypotheses. If we denote B resamplings of null test statistics by $t_{i,b}, i=1,\ldots,m, b=1,\ldots,B$, then estimates of the quantities α, α_0 and $\Pr(R>0)$ in the preceding formula are as follows:

$$\hat{\alpha}=\frac{1}{Bm}\sum_{b=1}^{B}R_b(\Gamma)=\frac{\overline{R}(\Gamma)}{m},$$

$$\hat{\alpha}_0=\frac{1}{Bm}\sum_{b=1}^{B}R_b(\Gamma_0)=\frac{\overline{R}(\Gamma_0)}{m},$$

$$\widehat{\Pr}(R>0)=\frac{\#\{b: R_b(\Gamma)>0\}}{B}=\overline{I}_{\{R(\Gamma)>0\}},$$

where $R_b(\Gamma)=\#\{i: t_{i,b}\in\Gamma\}$, $\overline{R}(\Gamma)=\frac{1}{B}\sum_{b=1}^{B}R_b(\Gamma)$, and similarly for $W_b(\Gamma)$ and $\overline{W}(\Gamma)$. Similar quantities for the rejection region Γ_0 can be defined.

Putting these all together, a conservative estimate of pFDR(Γ), making use of Γ_0 is

$$\begin{aligned}\widehat{\mathrm{pFDR}}_{\Gamma_0}(\Gamma) &= \frac{W(\Gamma_0)\cdot\overline{R}(\Gamma)}{(m-\overline{R}(\Gamma_0))\cdot(R(\Gamma)\vee 1)\cdot\widehat{\Pr}(R>0)} \\ &= \frac{W(\Gamma_0)\cdot\overline{R}(\Gamma)}{\overline{W}(\Gamma_0)\cdot(R(\Gamma)\vee 1)\cdot\overline{I}_{\{R(\Gamma)>0\}}}. \end{aligned} \tag{5.7}$$

By dropping the estimate of $\Pr(R>0)$, we can have a conservative estimate of $FDR(\Gamma)$ as

$$\widehat{\mathrm{FDR}}_{\Gamma_0}(\Gamma) = \frac{W(\Gamma_0)\cdot\overline{R}(\Gamma)}{(m-\overline{R}(\Gamma_0))\cdot(R(\Gamma)\vee 1)} = \frac{W(\Gamma_0)\cdot\overline{R}(\Gamma)}{\overline{W}(\Gamma_0)\cdot(R(\Gamma)\vee 1)}. \tag{5.8}$$

5.2.3 Estimation of pFDR q-values

Using the definition of q-values given in equations (2.4) and (2.3), the estimates of the q-value corresponding to the ordered p-values $p_{r_1}\le p_{r_2}\le\cdots\le p_{r_m}$ are

$$\widehat{q}_{p_0}(p_{r_i}) = \min_{k=i,\ldots,m}\widehat{\mathrm{pFDR}}_{p_0}(p_{r_k}). \tag{5.9}$$

If our interest is in deriving q-values corresponding to the t-statistics, let us suppose that $|t_{s_1}|\ge|t_{s_2}|\ge\cdots\ge|t_{s_m}|$. Writing Γ_{s_k} be $[-\infty,-|t_{s_k}|]\cup[|t_{s_k}|,\infty]$, the q-values are then

$$\widehat{q}_{\Gamma_0}(t_{s_i}) = \min_{k=i,\ldots,m}\widehat{\mathrm{pFDR}}_{\Gamma_0}(\Gamma_{s_k}). \tag{5.10}$$

Remark 5.1. Storey (2002a) has already pointed out that the FDR estimate based on equation (5.5) gives a procedure to control FDR. To see how this occurs, note that $R(p_{r_k})=k$, for $k=1,\ldots,m$, and that $\hat{\pi}_0=\frac{W(p_0)}{(1-p_0)m}$. Substituting these into (5.5) and enforcing step-up monotonicity, FDR-based adjusted p-values can be estimated by

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m}\left\{\min\left(\frac{m}{k}p_{r_k}\hat{\pi}_0,1\right)\right\}. \tag{5.11}$$

We call this the *Storey procedure*. Equation (5.4) and enforced monotonicity can also be used to compute q-values for controlling pFDR, and we

call this the *Storey-q procedure*. Similarly, the *ST-procedure* uses equation (5.8) and enforced monotonicity for controlling FDR under quite general dependence satisfying ergodic conditions, while the *ST-q procedure* used equation (5.7) and monotonicity to control pFDR. Details of these procedures are given in Box 5.

Comparing equation (5.11) with equation (5.1), it is easy to see that the method proposed by Storey (2002a) has advantages over that of Benjamini and Hochberg (1995), since $\hat{\pi}_0$ is less than or equal to 1. This should be no surprise, since equation (5.1) controls the FDR in the strong sense, while equation (5.11) controls the FDR in the exact sense, with an estimated π_0. If we are only considering the FDR in the exact sense, then π_0 can be estimated, and by noting that $\frac{m_0}{m} = \pi_0$ in equation (5.2) the two procedures are seen to be the same. Thus we come to see that exact control might give improvements in power over strong control. Similarly, we can replace m_0 in equations (3.2) and (3.5) to get more powerful single-step Bonferroni and Šidák adjustments. Indeed, Benjamini and Hochberg (2000) proposed a different estimator of $\hat{\pi}_0$, but Storey (2002a) proved that his method leads to conservative control of FDR.

5.3 Resampling procedures

For the BH and BY adjustments we simply use the algorithm in Box 1 and equations (5.1) and (5.3). For the Storey and Storey-q procedures, we first use the algorithm in Box 1 to compute the raw p-values for each gene, and then use equations (5.5) for the Storey procedure and (5.4) for Storey-q procedure, lastly enforcing step-up monotonicity for each procedure.

A complete algorithm is outlined in Box 5 for the ST and ST-q procedures. Note that our algorithm is slightly different from the original one, for we do not pool the t-statistics across all genes as did Storey and Tibshirani (2001). The reason we have not pooled across genes here is that we have not done so elsewhere in this paper. We feel that more research is needed to provide theoretical and practical justification of the pooling strategy of Storey and Tibshirani (2001).

Box 5. A permutation algorithm for the ST-q and ST procedures - based on Storey and Tibshirani (2001) Algorithm 1
Choose a value τ_0 believed to contain most null hypotheses (for example, $\tau_0 = 0.2$). From the original data, compute the two-sample t-statistics, let $\tau_i = |t_i|$ and assume without loss of generality $\tau_1 \geq \cdots \geq \tau_m$; otherwise sort the rows of the data matrix according to the ordered τ_i.
Compute $R_i = \#\{k : |t_k| \geq \tau_i\}$, and $W_0 = \#\{k : |t_k| \leq \tau_0\}$.
For the bth permutation, $b = 1, \ldots, B$:

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.
3. Compute $R_{i,b} = \#\{l : |t_{l,b}| \geq \tau_i\}$ for $i = 1 \ldots, m$ and $W_{0,b} = \#\{i : |t_{i,b}| \leq \tau_0\}$

The above steps are repeated B times, and then for $i = 1, \ldots, m$ estimate

$$\overline{R}_i = \frac{1}{B}\sum_{b=1}^{B} R_{i,b}, \quad \overline{I}_i = \frac{1}{B}\sum_{b=1}^{B} I(R_{i,b} > 0), \quad \overline{W}_0 = \frac{1}{B}\sum_{b=1}^{B} W_{0,b}.$$

Then at τ_i the pFDR is

$$\text{pFDR}_i = \frac{W_0 \cdot \overline{R}_i}{\overline{W}_0 \cdot (R_i \vee 1) \cdot \overline{I}_i} \qquad \text{for } i = 1, \ldots, m,$$

and the FDR is

$$\text{FDR}_i = \frac{W_0 \cdot \overline{R}_i}{\overline{W}_0 \cdot (R_i \vee 1)} \qquad \text{for } i = 1, \ldots, m.$$

The q-values (for the ST-q procedure) and the FDR-based adjusted p-values (ST-procedure) can then be estimated by enforcing step-up monotonicity as follows:

$$q_m = \text{pFDR}_m, \qquad q_i = \min(q_{i+1}, \text{pFDR}_i), \qquad \text{for } i = m-1, \ldots, 1,$$

$$\tilde{p}_m = \text{FDR}_m, \qquad \tilde{p}_i = \min(\tilde{p}_{i+1}, \text{FDR}_i), \qquad \text{for } i = m-1, \ldots, 1.$$

5.4 Empirical Bayes procedures and the SAM software

Several papers (Efron et al. (2000, 2001); Efron and Tibshirani (2002)) connect empirical Bayes methods with false discovery rates. Also, the popular software `SAM` (Significance Analysis of Microarrays) (Efron et al. (2000); Tusher et al. (2001)) computes false discovery rates from a frequentist viewpoint. The empirical Bayes calculations and the SAM software provide estimates of the FDR, but it is not clear whether these procedures provide strong control of the FDR, i.e. whether $E(V/R \mid H_{\mathcal{M}_0}) \leq \alpha$ for any subset $\mathcal{M}_0$. More theoretical work would seem to be needed to address these issues, see e.g. Dudoit et al. (2002a), and for this reason we will not discuss them further.

6 Data

6.1 Apo AI experiment

The Apo AI experiment (Callow et al. (2000)) was carried out as part of a study of lipid metabolism and atherosclerosis susceptibility in mice. The apolipoprotein AI (Apo AI) is a gene known to play a pivotal role in HDL metabolism, and mice with the Apo AI gene knocked out have very low HDL cholesterol levels. The goal of this Apo AI experiment was to identify genes with altered expression in the livers of these knock-out mice compared to inbred control mice. The treatment group consisted of eight mice with the Apo AI gene knocked out and the control group consisted of eight wild-type C57Bl/6 mice. For each of these 16 mice, target cDNA was obtained from mRNA by reverse transcription and labeled using the red fluorescent dye, Cy5. The reference sample used in all hybridizations was prepared by pooling cDNA from the eight control mice and was labeled with the green fluorescent dye, Cy3. Target cDNA was hybridized to microarrays containing 6,356 cDNA probes, including 200 related to lipid metabolism. Each of the 16 hybridizations produced a pair of 16-bit images, which were processed using the software package *Spot* (Buckley (2000)). The resulting fluorescence intensities were normalized as described in Dudoit et al. (2002b). For each microarray $j = 1, \ldots, 16$, the base 2 logarithm of the Cy5/Cy3 fluorescence intensity ratio for gene i represents the expression response x_{ij} of that gene in either a control or a treatment mouse.

Differentially expressed genes were identified using two-sample Welch t-statistics (Welch (1938)) for each gene i:

$$t_i = \frac{\bar{x}_{2i} - \bar{x}_{1i}}{\sqrt{\frac{s^2_{2i}}{n_2} + \frac{s^2_{1i}}{n_1}}},$$

where $\bar{x}_{1i}$ and $\bar{x}_{2i}$ denote the average expression level of gene i in the $n_1 = 8$ control and $n_2 = 8$ treatment hybridizations, respectively. Here s^2_{1i} and s^2_{2i} denote the variances of gene i's expression level in the control and treatment hybridizations, respectively. Large absolute t-statistics suggest that the corresponding genes have different expression levels in the control and treatment groups. In order to assess the statistical significance of the results, we use the multiple testing procedures of Sections 3 and 5, estimating raw and adjusted p-values based on all possible $\binom{16}{8} = 12{,}870$ permutations of the treatment and control labels.

6.2 Leukemia study

Golub et al. (1999) were interested in identifying genes that are differentially expressed in patients with two type of leukemias, acute lymphoblastic leukemia (ALL, class 1) and acute myeloid leukemia (AML, class 2). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing $p = 6{,}817$ human genes. The learning set comprises $n = 38$ samples, 27 ALL cases and 11 AML cases (data available at `http://www.genome.wi.mit.edu/MPR`). Following Golub et al. (personal communication, Pablo Tamayo), three preprocessing steps were applied to the normalized matrix of intensity values available on the website: (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer respectively to the maximum and minimum intensities for a particular gene across mRNA samples; (iii) base 10 logarithmic transformation. Boxplots of the expression levels for each of the 38 samples revealed the need to standardize the expression levels within arrays before combining data across samples. The data were then summarized by a $3{,}051 \times 38$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level for gene i in mRNA sample j.

Differentially expressed genes in ALL and AML patients were identified by computing two-sample Welch t-statistics for each gene i as in Section 6.1.

In order to assess the statistical significance of the results, we considered the multiple testing procedures of Sections 3 and 5 and estimated raw and adjusted p-values based on B = 10,000, 100,000 and 1,000,000 random permutations of the ALL/AML labels.

7 Results

The Holm, step-down maxT, and step-down minP procedures described in Sections 3 and 4, the BH, BY, Storey and ST procedures to control FDR in Section 5, and the Storey-q and ST-q procedures to control pFDR in Section 5 were applied to the two microarray datasets of Section 6. Figure 1 gives the results for the Apo AI knock-out data described in Section 6.1. It consists of three panels corresponding to three type I error rate controlling procedures. The top panel is for FWER, the middle one is for FDR and the bottom one is for pFDR. For each panel, the x-axis is always the rank of p-values. Note, the rank of different adjusted p-values is always the same as the rank of the raw p-values apart from the maxT procedure. In that case, the adjusted p-values have the same ranks as the two-sample t-statistics.

Similarly, Figure 2 gives the results of applying these procedures to the Golub leukemia dataset described in Section 6.2. Note that for both datasets, the adjusted p-values for FWER are mostly higher than the adjusted p-values for FDR, which in turn are a little lower than the q-values for pFDR. This was to be expected by the inequalities in equation (2.1).

For the FWER procedures, the greatest difference between the maxT and minP procedures occurred for the Apo AI dataset and the leukemia dataset with the smallest number of permutations B = 10,000. In these two cases, the procedures only rejected hypotheses at FWER level less than 0.18 for the leukemia data and 0.53 for the Apo AI data. This was due to the discreteness of the permuted raw p-values used to compute the Holm, and minP adjusted p-values. For the Apo AI dataset, with sample sizes $n_1 = n_2 = 8$, the total number of permutations is only $\binom{16}{8} = 12{,}870$, and hence the two-sided raw p-values must be at least $2/12{,}870$. As a result, the Holm p-values can be no smaller than $6{,}356 \times 2/12{,}870 \approx 1$. This highlights the greater power of the maxT p-value procedure in comparison with the Holm and the minP procedure, when the number of permutations is small.

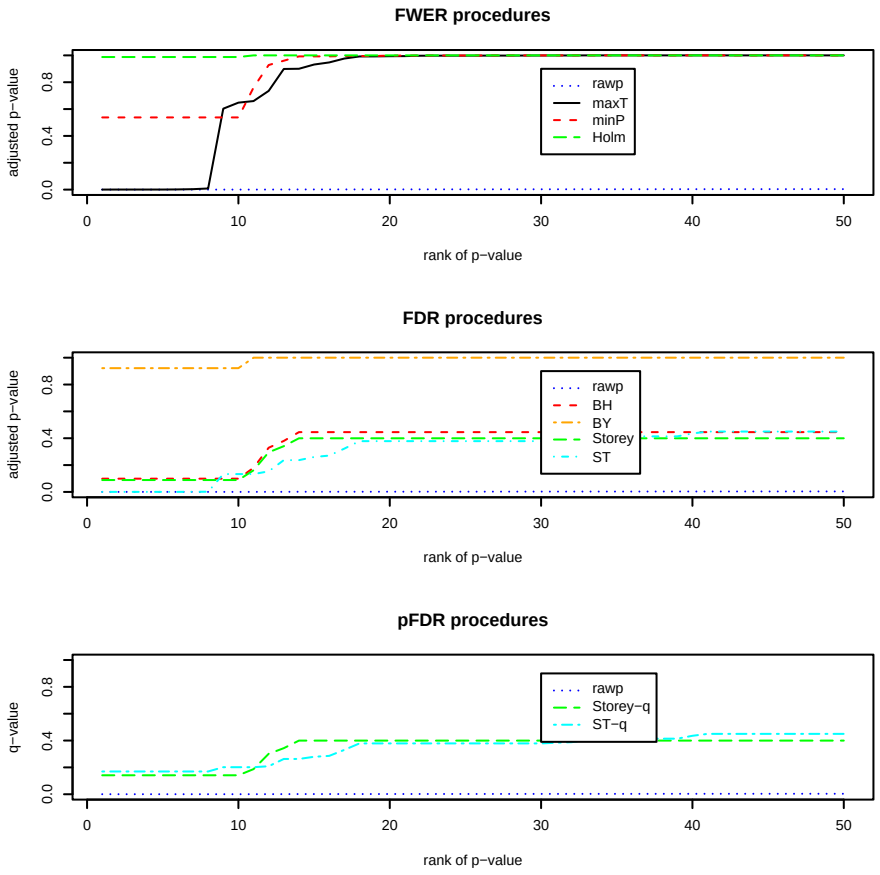

Figure 1: *Apo AI. Plot of adjusted p-values controlling different type I error rates against the rank of the p-values. p-values were estimated using all* $B = \binom{16}{8} =$ 12,870 *permutations. The top, middle and bottom panels are adjusted p-values controlling FWER, FDR and pFDR, respectively.*

To investigate the robustness of the Holm, maxT, and minP adjusted p-values to varying the number of permutations, we computed them for the leukemia dataset with B = 10,000, 100,000 and 1,000,000 permutations. Figure 3 showed that indeed the minP p-values were very sensitive to the number of permutations. After 100,000 permutations, the adjusted p-values become stable, while similar results for the Holm p-values are not shown.

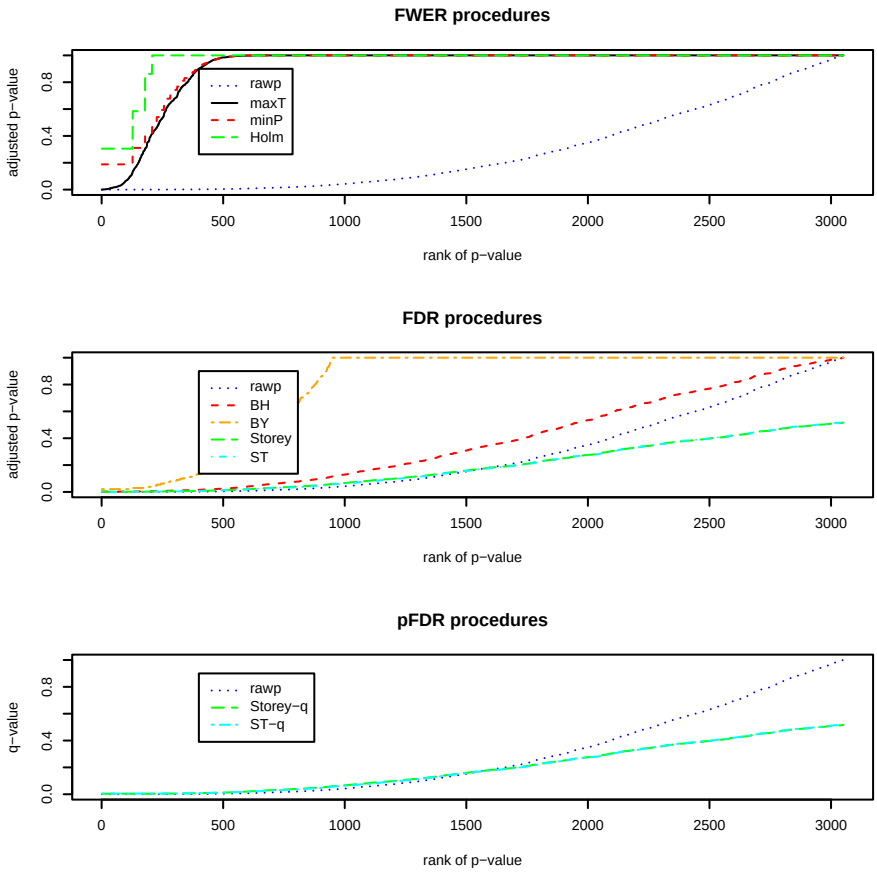

Figure 2: *Leukemia. Plot of adjusted p-values to controlling different type I error rates against the rank of the p-values. p-values were estimated using* $B = 10{,}000$ *random permutations. The top, middle and bottom panels are adjusted p-values controlling FWER, FDR and pFDR, respectively.*

On the other hand, the maxT adjustment was much more robust, for as seen in Figure 4 the adjusted p-values with $B = 10{,}000$, 100,000 and 1,000,000 are almost identical.

The FDR and pFDR procedures are also robust to the number of permutations, as they became stable for as few as $B = 1{,}000$ permutations. This is because these procedures use only a single round of permutations. The

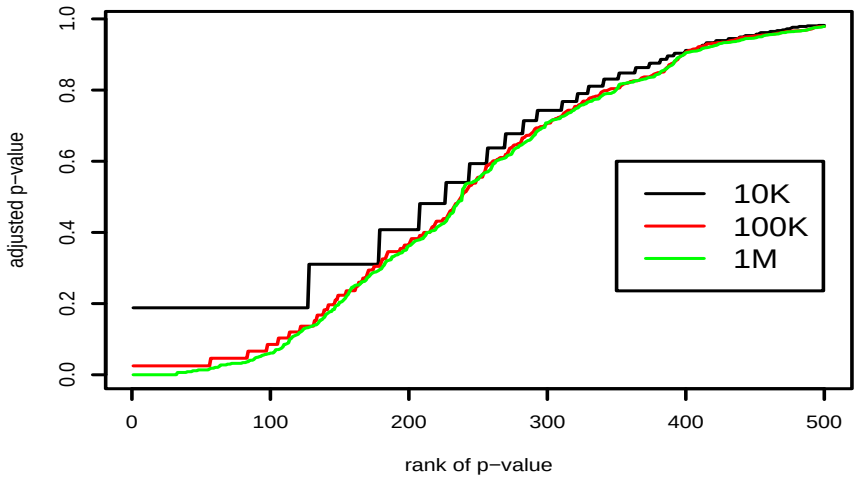

Figure 3: *Leukemia. Plot of minP adjusted p-values against the rank of adjusted p-values. p-values were estimated based on* $B = 10{,}000$, 100,000 *and* 1,000,000 *random permutations.*

BH adjusted values are very similar to the Storey and ST adjusted values, while the BY adjustments, trying to control FDR under arbitrary dependence, seem too conservative. The BY procedure gives adjusted p-values higher than those from the maxT procedure with the Apo AI dataset, and similar to them with the leukemia dataset. It seems that the BY procedure is not very useful in this context. The Storey-q and ST-q adjusted values are similar to each other, which could imply that the ability of ST-q to deal with dependence is not very great, or that there is not much dependence in the data.

Apo AI experiment. In this experiment, eight spotted DNA sequences clearly stood out from the remaining sequences and had maxT adjusted p-values less than 0.05. The ST procedures also pick the same 8, while all other procedures fail to pick them using a 0.05 cut-off. These eight probes correspond to only four distinct genes: Apo AI (3 copies), Apo CIII (2 copies), sterol C5 desaturase (2 copies), and a novel EST (1 copy). All changes were confirmed by real-time quantitative RT-PCR as described

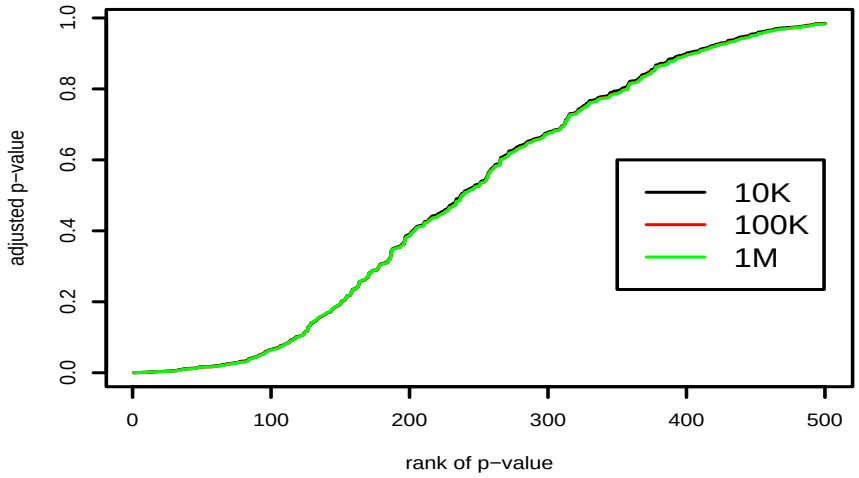

Figure 4: *Leukemia. Plot of maxT adjusted p-values against the rank of adjusted p-values estimated using* $B = 10{,}000$, 100,000 *and* 1,000,000 *random permutations.*

Dataset		Running times	
		Fast minP	maxT
Apo AI		9:38.89	3:4.23
Leukemia	$B = 10{,}000$	5:53.42	2:0.93
	$B = 100{,}000$	1:03:27.17	18:46.24
	$B = 1{,}000{,}000$	11:10:3.74	3:09:31.17

Table 4: *Running times of the fast minP and maxT algorithms for the Apo AI and leukemia datasets. Reported times are "user times" on Sun 200Mhz Ultrasparc workstations. The time is given in hours, minutes and seconds, e.g. 11:10:3.74 means 11 hours 10 minutes and 3.74 seconds*

in Callow et al. (2000). The presence of Apo AI among the differentially expressed genes is to be expected as this is the gene that was knocked out in the treatment mice. The Apo CIII gene, also associated with lipoprotein metabolism, is located very close to the Apo AI locus and Callow et al. (2000) showed that the down-regulation of Apo CIII was actually due to genetic polymorphism rather than absence of Apo AI. The presence of Apo AI and Apo CIII among the differentially expressed genes thus provides a

check of the statistical method, even if it is not a biologically interesting finding. Sterol C5 desaturase is an enzyme which catalyzes one of the terminal steps in cholesterol synthesis and the novel EST shares sequence similarity to a family of ATPases.

For Apo AI, we also considered adjusted p-values for non-parametric rank t-statistics. In this case, none of the procedures rejected any hypotheses at level less than 0.05. The poor performance of the maxT procedure using ranked data is likely due to the discreteness of the t-statistics computed from the ranks with a small sample size.

We also did a limited analysis to see how data selection affects adjusted p-values. For the Apo AI dataset, we selected the 10% of the 6,356 genes with the largest variances across the 16 samples, recomputed the step-down minP and maxT adjusted p-values. The adjusted p-values for the selected genes were always smaller or equal than the those for same genes within the complete data set (data not shown), and sometimes much smaller. This is reasonable, as a smaller number of hypotheses leads to smaller adjustments, but it highlights the fact that adjusted p-values will be affected by data preprocessing steps such as gene selection.

Leukemia study. Using the maxT adjustment, we found 92 (38) genes significant at the 0.05 (0.01) level, respectively. Among the 50 genes listed in Golub et al. (1999) (p.533 and Figure 3B), we found that 9 of those were not significant at the 0.05 level, and 27 of those were not significant at the 0.01 level. If we select 50 genes with the smallest adjusted p-values, 22 genes of Golub et al. (1999) (p.533 and Figure 3B) are not in our top 50 gene list. The results of minP were similar to those of maxT. We refer the reader to Golub et al. (1999) for a description of the genes and their involvement in ALL and AML. Note that this dataset is expected to have many genes differentially expressed between the two groups, and in this respect it is quite different from the Apo AI experiment, where we do not expect many genes to be differentially expressed. Since the Storey and ST procedures use information on the fraction of genes expected to be null, they can lead to adjusted p-values lower than the raw p-values, see the tail parts of the middle and bottom panels in Figure 2. In practice, we need not worry about this as only genes with small adjusted p-values (e.g. less than 0.05 or 0.10) are interesting, even in an exploratory analysis. A strategy to prevent this from happening would be to take the minimum of the raw p-values and the adjusted p-values. One final comment on this analysis:

the pre-processing for this dataset that was described in Section 6.2, in particular the filtering, would undoubtedly have an impact on the size of the adjusted p-values, perhaps reducing them considerably.

8 Discussion

8.1 Use of the new algorithm with the bootstrap and with other statistics.

In this paper, we gave a brief review of multiple testing procedures used in the analysis of microarray experiments. In particular we introduced a new and faster algorithm for calculating the step-down minP p-value adjustments. This algorithm not only makes it possible to analyze microarray data within the multiple testing framework, it also solves the general multiple testing problem described on page 114 of Westfall and Young's book as the double permutation problem. In brief, our algorithm reduces computational time from B^2 to $B \log B$, where B is the number of permutations. The idea of the algorithm can be extended to the bootstrap situation as well. The resampling-based test statistics simply need to be computed from samples with replacement rather than from permutations. We have described how to calculate adjusted p-values for two sample t-statistics, but the algorithm applies equally to other test statistics, such as the t with pooled variance, Wilcoxon, F, paired t, and block F-statistics.

In order to see this, let us focus on one gene only. Then we define the

(a) t-statistic with pooled variance: Let y_{ij} ($i = 1, 2, j = 1, 2, \ldots, n_i$ and $n_1 + n_2 = n$) be the observations from two treatments. Define $y_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, $i = 1, 2$. The t-statistic with pooled variance is:

$$t = \frac{y_{2\cdot} - y_{1\cdot}}{\sqrt{\frac{1}{n-2}\{\sum_{j=1}^{n_1}(y_{1j} - y_{1\cdot})^2 + \sum_{j=1}^{n_2}(y_{2j} - y_{2\cdot})^2\}(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

(b) Wilcoxon: The y_{ij} are defined as in (a). Rank all n observations, and denote the rank of observation y_{ij} by $s_{ij}, i = 1, 2, j = 1, 2, \ldots, n_i$. The rank sum statistic is $T = \sum_{j=1}^{n_2} s_{2j}$. As we have $E(T) = n_2(n+1)/2$, $Var(T) = n_1 n_2 (n+1)/12$, the normalized statistic is:

$$W = \frac{\sum_{j=1}^{n_2} s_{2j} - n_2(n+1)/2}{\sqrt{n_1 n_2 (n+1)/12}}.$$

(c) F-statistic: Let y_{ij} ($i = 1, 2, \ldots, k, j = 1, 2, \ldots, n_i$ and $\sum_{i=1}^{k} n_i = n$) be the observations from a one-way design. For treatment i, there are independent observations $y_{i1}, y_{i2}, \ldots, y_{in_i}$. Define $y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ and $y_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$. Then the F-statistic is

$$F = \frac{\sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2/(k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2/(n-k)}.$$

(d) Paired t-statistic: Let y_{ij} ($i = 1, 2 j = 1, 2, \ldots, n$) be n pairs of observations. If write $x_i = y_{2i} - y_{1i}$, then the paired t-statistic is

$$\text{paired } t = \frac{\bar{x}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)}}.$$

(e) Block F-statistic: Let y_{ij} ($i = 1, 2, \ldots, k, j = 1, 2, \ldots, n$) be the observations from a randomized block design with k treatments and n blocks. The observation on treatment i in block j is y_{ij}. Define $y_{i.} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}$, $y_{.j} = \frac{1}{k} \sum_{i=1}^{k} y_{ij}$ and $y_{..} = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$, then the block F-statistic is

$$\text{block } F = \frac{\sum_{i=1}^{k} n (y_{i.} - y_{..})^2/(k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - y_{i.} - y_{.j} + y_{..})^2/(n-1)(k-1)}.$$

Note that the t-statistic with pooled variance can be regarded as a special case of the F-statistic. Similarly, the paired t-statistic can be regarded as a special case of the block F-statistic. The Wilcoxon statistic is the nonparametric form of the t-statistic with pooled variance. Similarly, we can define other nonparametric statistics corresponding to the F, block F and paired t-statistics by replacing the observations y_{ij} with their corresponding ranks s_{ij}.

8.2 Which multiple testing procedure?

We have seen a bewildering variety of multiple testing procedures. How should we choose which to use? There are no simple answers here, but each procedure can be judged according to a number of criteria. Interpretation: does the procedure answer a question that is relevant to the analysis? Type of control: weak, exact or strong? Validity: are the assumptions under

which the procedure is valid definitely or plausibly true, or is their truth unclear, or are they most probably not true? And finally, computability: are the procedure's calculations straightforward to perform accurately, or is there substantial numerical or simulation uncertainty, or discreteness?

In this paper, we have learned a little about the procedures which control different type I error rates. From equation (2.1), the FWER is the most stringent, while FDR is the most relaxed, and pFDR is roughly in between. In microarray experiments, where we consider many thousands of hypotheses, FDR or pFDR are probably better criteria than FWER. Most people would be more interested in knowing or controlling the proportion of genes falsely declared differentially expressed, than controlling the probability of making one or more such false declarations. Most would not consider it a serious problem to make a few wrong decisions as long as the majority of the decisions are correct. The FDR and pFDR procedures promise to respond to this need, but there remain issues of validity.

It will take a while before we accumulate enough experience to know which approach leads to truer biological conclusions on a large scale, that is, which in truth better balances false positives and false negatives in practice. The FWER-based maxT procedure successfully identified the 8 differentially expressed genes in the Apo AI dataset which have been biologically verified, though the identical distribution assumption is highly doubtful, see below. For the Golub leukemia dataset, maxT gave a smaller number of differentially expressed genes than the FDR-based procedures, but no large scale validation has been done to determine the truth there. It seems possible that when just a few genes are expected to be differentially expressed, as with the Apo AI dataset, it might be a good idea to use FWER-based procedures, while when many genes are expected to be differentially expressed, as with the leukemia dataset, it might be better to use FDR or pFDR-based procedures.

Of all the procedures described in this paper, only Holm, minP and BY are essentially assumption-free. However, Holm and BY suffer from being far too conservative. On the other hand, minP is useful if we have enough experiments so that there are enough permutations to eliminate the discreteness of the p-values. While maxT is much less sensitive to the number of permutations, it does require the assumption of identical distributions. The strength of maxT and minP is that they are exploiting the dependence in the test statistics in order to improve power. By

contrast, the Šidák, BH, Storey and ST procedures are motivated by independence (and perhaps identical distribution) assumptions on the test statistics. They try to extend results valid under independence to more general situations by imposing special conditions: the Šidák inequality for the Šidák procedure, positive regression dependency for BH, and ergodic conditions for the Storey procedure. For a start, it is difficult to see how we can be sure that these conditions apply with microarray data. Further, it is hard for these procedures to improve power, as they do not fully exploit the information concerning dependence in the dataset. A potentially fruitful direction for future research is to develop variants of the FDR procedures similar to maxT or minP, which permit arbitrary dependence between the test statistics, and which automatically incorporate this dependence into the procedures in order to improve power.

Other issues are computational complexity and discreteness. We have seen that minP is the most computationally expensive procedure, and the computational burden can be substantial, even with our new algorithm. Table 4 shows the running time for minP and maxT giving different numbers B of permutations. Figure 5 shows the curves for minP and maxT comparing to the theoretical running time. It shows that for most practical applications, minP is about 3 times slower than the maxT procedures. The maxT procedure has the same complexity as computing the raw p-values. The other procedures, such as Holm, Šidák, BH, BY, Storey, and Storey-q are all based on computing the raw p-values, they should have the same running time as maxT. By contrast, ST and ST-q are more computationally expensive than maxT if the same number B of permutations is used. In practice, it is not necessary to run more than 1,000 permutations because ST and ST-q are quite robust to the number of permutations, so the computational burden will be at the same level as maxT. In summary, in terms of computational simplicity and speed, Holm, Šidák, BH, BY, Storey and Storey-q are good choices, followed by maxT, ST and ST-q, with the most computationally demanding being minP. We have discussed the issue of discreteness above, and noted that it can really affect minP, indeed it can be a real restriction on the use of minP if the number of permutations is not large enough.

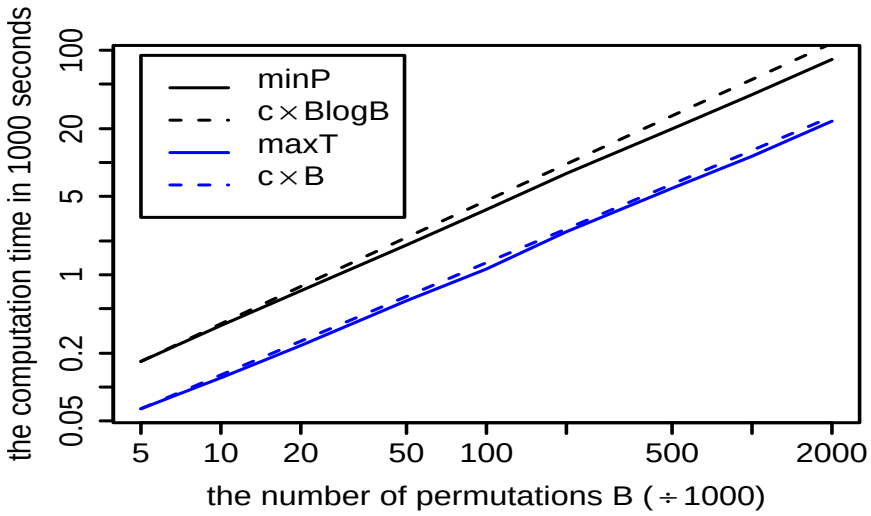

Figure 5: *Leukemia. Plot of computation time against the number of permutations B. The dashed lines are computed using the theoretical run time formulae shown in the legend. The constants c were computed from the timings for $B = 5{,}000$ permutations for maxT and minP separately.*

8.3 The C and R code available for different tests

The algorithms for the minP and maxT adjustment are implemented in C code, and incorporated in the R package *multtest*. R, Ihaka and Gentleman (1996), is free open source software similar to S/Splus. The C code and R package *multtest* may be downloaded from the Bioconductor website `http://www.bioconductor.org`. Currently, the package can deal with the t, t with pooled variance, F, paired t, Wilcoxon, and block F-statistics. It can also deal with the nonparametric forms of those statistics. The fixed random seed resampling method is implemented, and also the approach to store all of the permutations (see remarks 2(a) and 2(b) in Section 4.4.3) for most of these tests. The package also implements some FDR procedures such as BH and BY.

Acknowledgments

We thank Greg Grant, John Storey and Peter Westfall for their helpful discussions on the subject matter of this paper. A number of people have read this paper in draft form and made valuable comments, and we are grateful for their assistance. They include Julia Brettschneider, Darlene Goldstein, Greg Grant, Kathy Ruggiero, Yu Chuan Tai, Sue Wilson, Jean Yang and Xiaoyue Zhao. Finally, we would like to acknowledge support from NIH grants 8RO1GM59506A and 5RO161665-02.

DISCUSSION

Gary Glonek and Patty Solomon
University of Adelaide, Australia

Ge, Dudoit and Speed have made a welcome and impressive contribution in the formulation of an efficient algorithm for the calculation of the `minP` adjusted P-values, the provision of computer programs to implement the various adjustments and, perhaps equally importantly, in providing valuable insight into what is indeed a bewildering variety of methods for multiple testing.

Our first comment concerns the applicability of these methods to microarray experiments of the type that we often encounter in practice. A key element of any hypothesis test is the ability to obtain the null distribution of the test statistics and most of the methods in the present paper do this by means of permutation of the columns of the data matrix. However, in practice we often see experiments where this approach cannot be used. The reasons are either that there are too few slides to obtain a useful distribution from permutations or that the experimental design is such that the null hypothesis of interest cannot be specified in terms of invariance under a suitable set of permutations of the data. In such cases, it is very difficult to obtain even the raw p-values without undue reliance on parametric models and their assumptions. The obvious remedy of performing experiments with proper and adequate replication is not always easily applied. The increasing capacity to produce large numbers of slides often generates larger, more complicated experiments rather than simple highly replicated designs.

Thus it would appear that there is still an important class of problems that falls beyond the scope of presently available methods.

Our second remark relates to the purpose of multiple testing procedures in the context of microarray experiments. In this paper, the motivation given is to provide biologists with a quantitative assessment of the likely differential expression of each gene, so that they do not have to follow up genes with little prospect of being truly differentially expressed. The authors argue persuasively that, at least in principle, and particularly when many genes are expected to be differentially expressed, the FDR and pFDR are preferable to the more traditional approach of FWER. However, we are concerned that even these measures may not always fit with the stated purpose. Our difficulty lies with the fact that the pFDR associated with a particular rejection region applies to all genes within the region and does not discriminate between those that are close to the boundary and those that are not.

Suppose that as in Section 2.3 of Ge et al. the test statistics $Z_i|H_i$ are distributed independently as $(1 - H_i)F_0 + H_iF_1$ and that the H_i are independent Bernoulli(π_1), where $\pi_1 = 1 - \pi_0$.

Example 1. Assume that F_0 is the $N(0, 1)$ distribution, F_1 is the $N(6, 1)$ distribution and $\pi_1 = 0.1$, i.e., that a small proportion of genes are expected to be (independently) truly differentially expressed with a six-fold change in expression on the log scale. Consider the rejection region defined by $\Gamma = \{z : z \geq 2\}$. The pFDR in this case is $P(H = 0|Z \geq 2) = 0.17$ which for many purposes is acceptably low. Suppose now the test statistic for a particular gene is $z_i = 2$ and consider the question of how likely is it that such a gene is truly differentially expressed. The q-value for such a gene is then 0.17 but the posterior probability of no differential expression is $P(H = 0|Z = 2) = 0.99972$. In other words, although the pFDR for the procedure is 17%, the rate of false discovery for genes with $z \approx 2$ will be about 99.972%. The situation is illustrated more fully in Figure 1, by plotting both the q-value and the posterior probability of $H = 0$ against z. The posterior probabilities show the region $\{2 \leq z \leq 3\}$ to be barren territory for differential expression despite the fact that the pFDR remains low.

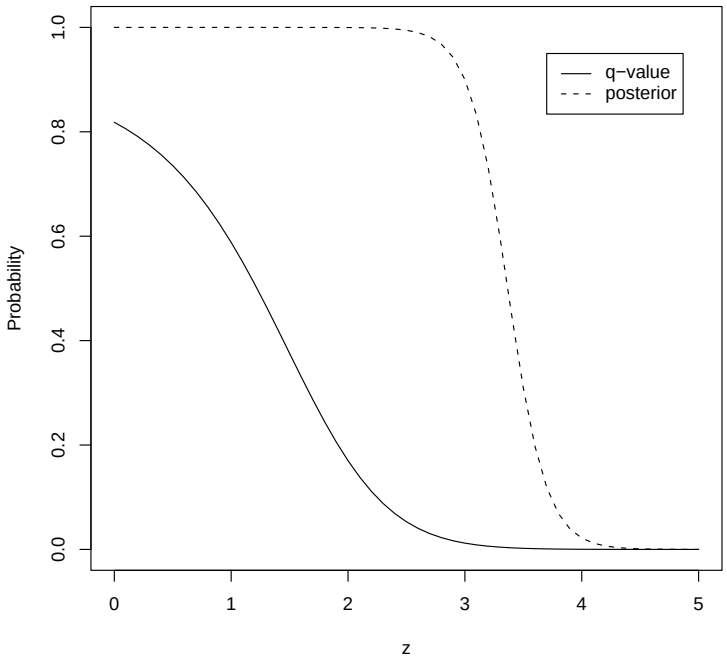

Figure 1: *q-values and posterior probabilities for* $\pi_1 = 0.1$, $F_0 = N(0,1)$ *and* $F_1 = N(6,1)$.

Example 2. Assume now that F_0 is the $N(0,1)$ distribution, F_1 is the $N(1,1)$ distribution and $\pi_1 = 0.4$ and consider again the rejection region defined by $\Gamma = \{z : z \geq 2\}$. Here, a larger proportion (40%) of genes is expected to be differentially expressed, but with only a two-fold change in expression under the alternative hypothesis. Suppose the test statistic for a particular gene is $z_i = 2$ and consider the question of how likely is it that such a gene is truly differentially expressed. The q-value for such a gene is then 0.177 and the posterior probability of no differential expression is $P(H = 0|Z = 2) = 0.251$. Figure 2 shows the q-values and posterior probabilities and in this case there is clearly an abundance of differential expression in the region $\{2 \leq z \leq 3\}$.

To summarise, in both examples using the rejection region $\Gamma = \{z : z \geq 2\}$ results in a pFDR of approximately 17%. This threshold is clearly too low in the first example but arguably not so in the second. On this

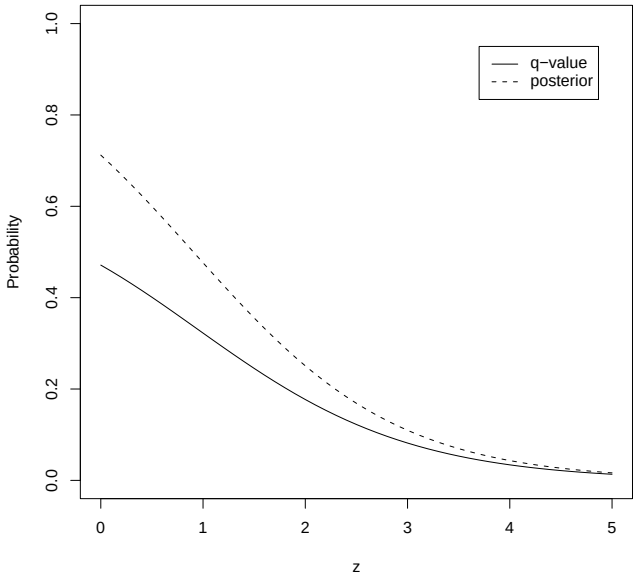

Figure 2: *q-values and posterior probabilities for* $\pi_1 = 0.4$, $F_0 = N(0,1)$ *and* $F_1 = N(1,1)$.

basis we conclude that consideration of the pFDR alone may not be enough to answer the question originally posed. Of course there are other issues, both practical and theoretical, that would need to be addressed for these ideas to be applied in practical situations involving the complexities of microarray experimentation. Such analysis is clearly beyond the scope of this discussion. Nevertheless we believe that our examples give cause to question whether the pFDR alone is a suitable criterion.

Gregory R. Grant
Computational Biology and Informatics Laboratory (CBIL)
University of Pennsylvania, Philadelphia

This is a very welcome organization of some of the current theory which has been developed for tackling the multiple testing issues which arise when using microarrays to help find differentially expressed genes between two conditions. The authors show where the field has arrived, and give a good feeling for where it is going. The focus is primarily on the theoretical basis of

the various methods and how to effectively compute them using re-sampling based approaches. Emphasis is on controlling type I errors (false-positives), and the authors have a fairly broad survey of the literature, including some of their own results, for theoretically sound non-parametric methods. There is still much to be done in the field to reduce the number of assumptions the algorithms make to maintain effective computation without excessive loss of power. The tradeoff is a loss of accuracy in the confidence measures themselves. The authors present an efficient way to implement a method which makes fewer assumptions than previously implemented methods, that is an efficient implementation of the minP statistic. The power, however, is probably not sufficient for most practical purposes.

The landscape of practical application of re-sampling based methods is not much discussed. The SAM method (Tusher et al., 2001) is disregarded as not having been sufficiently well investigated to date to consider, however it too provides a type of adjusted p-value should at least approximate strong control of the type I error in the FDR sense. SAM is perhaps being used in practice more than all of the other methods described in the paper in combination. It would be interesting to see how it compares with the other methods on the data sets in the results sections. Unfortunately there are no biological benchmarks for microarray data yet, not enough is known about any biological system to know exactly which are the differentially expressed genes. There are on the other hand empirical methods, simulations, which can help validate approaches, to varying degrees. Such validation methods however are not focussed on.

Which methods are being used by the end user community is mainly a function of which of them have been implemented in the most user-friendly fashion and give the least conservative results. SAM, for example, is an Excel plug-in. SAM also often gives researchers results they find reasonably satisfying, which usually are results substantially less conservative than traditional experiment-wise p-values. About half of the focus of Ge et al. is on the experiment-wise approach. The much more widely preferred and less conservative False Discovery Rate approach is given the other half of the attention, though the methods described to control it do not enjoy particularly wide use yet, particularly the pFDR.

Much of the foundation for the theory in this paper is inspired by Westfall and Young (1993). The less sophisticated reader might find it difficult to translate the theory as outlined in the paper, into the framework as

outlined in Westfall and Young. A parametric framework is required which microarray gene-expression data do not naturally fit, the review assumes instead the so-called non-parametric set-up, leaving all things unspecified except for issues about the equality of means across groups.

The paper proposes a new method to compute permutation distributions using p-values themselves as statistics instead of t-statistics. To date the t-statistic method has been how things have been implemented while the p-value approach has been considered too computationally intensive. Overcoming this is quite important because the t-statistic cannot be expected to be distributed identically for all genes, and so should not be compared directly across genes. The computational difficulty is outlined in Westfall and Young, page 114, however the possibilities of a space-consuming rather than time-consuming approach was not considered. Such an approach, requiring probably less than 1 gigabyte for microarrays, is also fairly natural and is not nearly as severe as the time-hit on today's computers. Happily the Ge et al. approach saves on both the time and space requirements, and as far as I know this is the first actual implementation of the minP statistic via re-sampling. This removes a very troublesome assumption, but is still among the conservative experiment-wise p-value methods. Even though it is a an experiment-wise approach which is not nearly as conservative as the other standard experiment-wise methods, such as the Bonferroni, which is very severe in the context of microarrays, it is still quite conservative, as can be seen in the results the authors obtain with it in the results sections.

It is likely that experiment-wise methods will one day be extinct in lieu of the much more popular FDR methods which allow mistakes while controlling the number of them. Already in practice experiment-wise methods are on the fringes. The low power of experiment-wise approach is reinforced by the results they have outlined in the results section of the paper. The minP method though a breakthrough in being the most theoretically satisfying, requires a greater number of permutations than the maxT method, and in fact requires a greater number of replicates per condition than the majority of groups are currently willing to perform, microarray data being extremely time-consuming and expensive. In fact performing more replicates sometimes introduces even more variability into the data as things like date of hybridization and numbers of technicians have a greater effect. One is tempted to opt for the maxT statistic because of its ability to function with fewer replicates, however it is not clear what effect this might have due to the violation of more assumptions. However given the new efficiency

of calculation, the authors hint upon application of minP type methods to the FDR approach. It's clear this is exactly where this field is going.

Christina M. Kendziorski
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison, USA

A number of statistical approaches have been developed to address the problem of identifying genes differentially regulated between two or among multiple conditions. Most approaches generally involve construction of some gene specific test statistic followed by identification of a rejection region and assessment of error. It is clear that two types of errors can be made at each gene; but the exact measure of error across multiple hypotheses has been a subject of study historically and a subject of much debate recently in the microarray literature. The authors provide a thorough and much needed overview of a number of methods for controlling type I error following multiple tests.

The relationship between the empirical Bayes approach and FDR is mentioned, but deserves additional comment. It has been known for some time that an empirical Bayes approach utilizing hierarchical models accounts *naturally* for type I errors following multiple tests (Berry, 1988). By *naturally*, I mean within the context of the empirical Bayes hierarchical modeling framework where posterior probabilities, the quantities of interest, depend on the number of tests and the entire set of realized data values. This dependence itself provides some adjustment of the posterior probabilities; additional adjustments are not always required. In fact, some simulation studies of microarray data have shown that additional adjustments of posterior probabilities derived within the context of a hierarchical modeling framework are not necessary to control the FDR. In Kendziorski et al. (2003), a simulation study shows that FDR is reasonably small (< 0.05) following rejection for posterior probabilities greater than 0.5 (this procedure is the Bayes rule which minimizes the posterior expected count of false positives and false negatives). Estimates obtained in the context of the simulation study are consistent with those provided by the Efron et al. (2000) empirical Bayes method. In addition to FDR, we also evaluated our approach in the context of other measures of error and performance. As expected, the utility depends on the sample size; and consideration of measures in addition to FDR could prove useful in sample size calculations.

Müller et al. (2003) consider FDR, false non-discovery rate (FNDR), and sensitivity to determine optimal sample sizes for microarray studies.

Of course, the results of any simulation study depend on the method used to simulate the data and it is for this reason that I point out a somewhat minor point mentioned by the authors that I strongly agree with: it will be a while before we accumulate enough experience to know which methods lead to more accurate biological conclusions. In the meantime, better methods to simulate microarray data could be useful in assessing methodologies and addressing a number of open questions.

John H. Maindonald
Centre for Bioinformation Science Mathematical Sciences Institute and John Curtin School of Medical Research
Australian National University, Canberra, Australia

It is highly useful to have these various approaches for handling multiplicity documented and compared, in the one place. The contrast with adjustments for multiplicity in more traditional statistical contexts is interesting. With four or five treatments to compare in a one-way design, is an adjustment for multiplicity really necessary? While many different forms of adjustment have been proposed, few of these have found their way into common use.

By contrast, there is a very short tradition of experience in the application of methods for adjusting for multiplicity in this microarray context, where the number of tests can be huge, and some form of adjustment for multiplicity is clearly necessary. There are many methods on offer, with more likely to appear in the next few years, and with a very limited tradition of experience in their application.

Misinterpretations of p-values and adjusted p-values are widespread, some of them incorporated into output from widely used commercial microarray packages. A recent complaint to those responsible for the output of one such package drew the response "that we attempt to strike a balance between statistical correctness and accuracy, and making concepts clear to non-statistical users"! I cringe at the possibilities for misinterpretation that are offered by the bewildering variety of motivations and methodologies that are described in the present paper.

We must hope, as the authors suggest, that practical experience will

in due course decide between these and other methods that will present themselves. This will take a long time to accumulate. While we accumulate such experience, the technology will continue to change, placing in question the usefulness of any except very recent experience.

Both for boostrap and for permutation methods, the distribution is a poor estimator of the population distribution for small sample sizes. Is it possible to build in parametric assumptions, perhaps assuming that the distribution is normal except in the tails, that will reduce the problem of loss of power relative to normal theory methods?

It would be interesting to do the same comparisons under conditions where the discreteness of the permutation distributions is more of an issue, for example with four or six mice per treatment. Also, why not use the permutation distribution to calibrate p-values that are derived from normal theory assumptions, interpolating between the discrete probabilities from the permutation distribution?.

Variation in the denominators of the sample t-statistics, and in the t-statistics themselves, will be more extreme than variation in the "true" unknown variances. Better ways than at present available are needed to use information on the distribution of variances across different genes to improve the crude variance estimates. A complication is that these variances can be, and in these data are, a function of hybridization intensity, of specific print-tip effects, and of order of printing effects. While most of the methods do not change the rank order of the genes (the sequential methods may change the order), changing the variance estimator will change the ranking.

John D. Storey
Department of Statistics
University of California, Berkeley, USA

Yongchao Ge, Sandrine Dudoit, and Terry Speed have written a lucid article on multiple testing in the context of DNA microarray studies. I have greatly benefitted from the interaction I have had with them, and it has greatly influenced my work in this area. Their presentation of the issues is thoughtful and careful, both in this article and in their previous work (Dudoit et al. (2002b)). This particular article will undoubtedly serve as a standard reference for those wanting to become acquainted with the research area.

Ge, Dudoit, and Speed (henceforth abbreviated by GDS), discuss both FWER and FDR. It seems that most microarray experiments will involve conditions where a reasonable fraction of the genes are differentially expressed. In such cases, the FDR is likely the more appropriate quantity. (Of course, one can create examples where FWER would be more applicable.) Therefore, my comments will be limited to false discovery rates and consist of four points. First, I will show that many of the different approaches to FDR they have presented do in fact become equivalent if one views p-value calculations from a "pooling" point of view. Second, I will review some recent results I have completed with Jonathan Taylor and David Siegmund that directly address some of the concerns they raise. Third, I will discuss where I think dependence is an issue in DNA microarray experiments, where it is not an issue, and how this relates to some of the methods they discuss. Fourth, I will argue that q-values provide a good gene-specific measure of significance as long as one considers them simultaneously in the appropriate way.

Connections Between Procedures

Suppose that m hypothesis tests are simultaneously tested with corresponding p-values $p_1, p_2, \ldots, p_m$. Benjamini and Hochberg (1995) propose the following algorithm for controlling the FDR at level α. Let $T_{BH} = \max\{p_i : p_i \leq \frac{i}{m}\alpha\}$. Then reject all null hypothesis corresponding to $p_i \leq T_{BH}$. When the null p-values are independent and uniformly distributed, this procedure strongly controls the FDR at level α. In Storey (2002a), I suggest the following estimate of FDR for a fixed p-value threshold t:

$$\widehat{FDR}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\frac{1}{m}\sum_{i=1}^{m} I(p_i \leq t)}, \tag{1}$$

where $\widehat{\pi}_0(\lambda)$ is an estimate of π_0, the proportion of true null hypotheses, with tuning parameter $0 \leq \lambda < 1$. The form of $\widehat{\pi}_0(\lambda)$ is

$$\widehat{\pi}_0(\lambda) = \frac{\sum_{i=1}^{m} I(p_i > \lambda)}{m(1-\lambda)}. \tag{2}$$

It is shown in Storey (2002a) under an i.i.d. mixture model that $\mathrm{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$, where $FDR(t)$ is the false discovery rate attained when thresholding the p-values for significance at t. This inequality holds

when the null p-values are independent and uniformly distributed (Storey et al. (2002)), the same conditions as in Benjamini and Hochberg (1995).

Even though the above form of "strong control" is from the opposite viewpoint of Benjamini and Hochberg (1995), it is tempting to form the threshold

$$T_\lambda = \max\{t : \widehat{FDR}_\lambda(t) \leq \alpha\} \tag{3}$$

in order to provide strong control of the FDR. It follows that $\widehat{\pi}_0(\lambda = 0) = 1$ so that $\widehat{FDR}_{\lambda=0}(t) = mt/\sum_{i=1}^{m} I(p_i \leq t)$. From this, it easily follows that $T_{BH} = T_{\lambda=0}$. Therefore, if one takes certain liberties with the procedure proposed in Storey (2002a), it can be viewed as a generalization of the BH procedure. In fact, we have shown in Storey et al. (2002) that T_λ strongly controls the FDR at level α (again under the same conditions as in Benjamini and Hochberg (1995)), under the constraint that $T_\lambda \leq \lambda$. The fact that the threshold occurs $\leq \lambda$ is a bit of a nuisance, but makes little difference in practice for wisely chosen λ. This constraint is unnecessary for large m, which I discuss later.

There has been much confusion in the literature recently over the differences between controlling the FDR via p-values or through permutation methods. In fact, GDS quickly dismiss the FDR method used in SAM (Tusher et al. (2001)) as being unconventional and not even worth discussing. For the case of detecting differential gene expression between two conditions, Tusher et al. (2001) define an asymmetric, data-dependent thresholding rule for significance, based on modified t-statistics and a quantile-quantile plot. The thresholding rule is indexed by $0 \leq \Delta < \infty$, where the larger Δ is, the fewer the number of significant genes there are. For a fixed Δ, Tusher et al. (2001) estimate the FDR by $\mathrm{E}[V^*(\Delta)]/R(\Delta)$, where $R(\Delta)$ is the number of significant genes at this threshold. $\mathrm{E}[V^*(\Delta)]$ is the average number of genes called significant under the permutation distribution obtained by scrambling the group labels, using the same asymmetric thresholding rule.

The following result shows that this method is in fact equivalent to the Benjamini and Hochberg (1995) method in the sense described above, as long as one calculates p-values by pooling across genes. Let $\widetilde{\Delta}_i$ be the largest Δ so that gene i is called significant, for $i = 1, 2, \ldots, m$. Then the p-value of gene i, when pooling across genes (i.e., assuming their null distributions are the same), is $p_i = \mathrm{E}[V^*(\widetilde{\Delta}_i)]/m$. This easily follows by the definition given in Lehmann (1986) and by considering the nested set

of significance regions indexed by Δ.

Theorem 1. Let $p_i = \mathrm{E}[V^*(\widetilde{\Delta}_i)]/m$, $\mathrm{E}[V^*(\Delta)]$, and $R(\Delta)$ be defined as above. Then the BH algorithm applied to $p_1, p_2, \ldots, p_m$ is equivalent to calling all genes significant by $\widehat{\Delta}$ in SAM where

$$\widehat{\Delta} = \min \left\{ \Delta : \frac{\mathrm{E}[V^*(\Delta)]}{R(\Delta)} \leq \alpha \right\}.$$

Therefore, one can use the SAM software in the above way to perform the BH method. We indirectly state this fact in Storey and Tibshirani (2001), but we do not explain it as thoroughly. Given this equivalence, I think that GDS have overlooked one potentially greater drawback of SAM. The rule defined by the quantile-quantile plot and Δ is determined from the same set of data on which the FDR estimates are made. It is clear that this can result in "over-fitting" and anti-conservative biases in FDR calculations. As an extreme example, suppose that we apply SAM to detecting differential gene expression in a single gene. It then uses right-sided or left-sided significance regions, depending on whether the observed statistic is respectively positive or negative. It is not hard to show that this results in a p-value that is $1/2$ of its actual size, and therefore the FDR estimates will be two times too small. As the number of genes increases, this bias decreases, but it is always present for the most significant genes.

By noting that any use of averaging over the number of statistics called significant under some simulated null distribution is equivalent to calculating p-values by pooling across genes (or tests), it can be seen that many of the re-sampling based FDR methods are simply p-value based methods with globally defined p-values. Moreover, because the expectation of the sum of indicator random variables is the same regardless of the dependence present between them, it is difficult to see how the re-sampling approach captures dependence in the FDR case. Because of this, I am slightly skeptical about how useful and novel the current re-sampling approaches are in false discovery rates (Of course the scenario is quite different for FWER where one is concerned with $\Pr(V \geq 1)$ (Westfall and Young, 1993)). The "Storey" and "ST" methods employed in GDS would have been completely equivalent if they had pooled across genes to form p-values. GDS argue that there is no reason to suspect that each gene has the same null distribution. Perhaps this is true, but their argument for "subset pivotality" also requires assumptions. Both sets of assumptions can be met with arguments

based on a large number of arrays. Finally, note that it takes B permutations when calculating p-values across genes to get the same resolution as mB permutations when calculating p-values within genes. Recall that m is usually on the order of 3000 to 30,000.

Next, I review several very recent results that do not depend on an independence assumption, nor on the assumption that each gene has the same null or alternative distributions.

Recent Results with Applicability to DNA Microarrays

For a large number of genes m, several results about $\widehat{FDR}_\lambda(t)$ and T_λ (see equations 1-3) have been shown that increase their applicability. Note that $V(t)/m_0 = \frac{\#\{\text{null } P_i \leq t\}}{m_0}$ and $S(t)/m_1 = \frac{\#\{\text{alt. } P_i \leq t\}}{m_1}$ are the empirical distribution functions of the null and alternative p-values, respectively. Almost sure convergence in the point-wise sense as $m \to \infty$ means that with probability 1:

$$\begin{aligned} \frac{V(t)}{m_0} &\to G_0(t) \text{ for each } t \in [0,1], \\ \frac{S(t)}{m_1} &\to G_1(t) \text{ for each } t \in [0,1], \end{aligned} \tag{4}$$

for some functions G_0 and G_1. The following results are proven in Storey et al. (2002). These are closely related to several results in Genovese and Wasserman (2001).

Theorem 2 (*Storey et al. 2002*) Suppose that $V(t)/m_0 = \frac{\#\{\text{null } P_i \leq t\}}{m_0}$ and $S(t)/m_1 = \frac{\#\{\text{alt. } P_i \leq t\}}{m_1}$ converge almost surely point-wise to continuous G_0 and G_1, respectively, where $G_0(t) \leq t$. Also suppose that $\lim_{m\to\infty} m_0/m = \pi_0$ exists. Then for each $\delta > 0$,

$$\lim_{m\to\infty} \inf_{t \geq \delta} \left[\widehat{FDR}_\lambda(t) - FDR(t) \right] \geq 0 \tag{5}$$

with probability 1. Also,

$$\lim_{m\to\infty} FDR(T_0) \leq \lim_{m\to\infty} FDR(T_\lambda) \leq \alpha. \tag{6}$$

Therefore, the estimate $\widehat{FDR}_\lambda(t)$ simultaneously dominates $FDR(t)$ over all thresholds t for large m. Also, the generalized thresholding proce-

dure T_λ asymptotically controls the FDR at level α. We have $\lim_{m\to\infty} FDR(T_0) < \lim_{m\to\infty} FDR(T_\lambda)$ for $\lambda > 0$, when G_0 and G_1 are strictly monotone. Under these conditions, the generalized procedure is more powerful than the BH procedure ($T_0 = T_{BH}$).

Estimates of the q-values for each p_i were given in Storey (2002a). We have also shown that under these conditions the q-values are simultaneously conservatively consistent. Therefore, for large m, one can examine all genes and their q-values simultaneously without inducing bias. This is explicitly stated in the following result.

Corollary 1 (*Storey et al. 2002*) For a given p-value p_i, let $\widehat{q}_\lambda(p_i)$ be its estimated q-value as defined in Storey (2002a). Then under the conditions of Theorem 8.3,

$$\lim_{m\to\infty} \inf_{t\geq\delta} [\widehat{q}_\lambda(t) - \text{q-value}(t)] \geq 0$$

for each $\delta > 0$.

These asymptotic results hold under the point-wise convergence of the empirical distribution functions. Note that we did not require each test to have the same null distribution, but rather the null distributions have to converge to some function. Many forms of weak dependence allow point-wise convergence of empirical distribution functions, for example ergodic dependence, blocks of dependent tests, and certain mixing distributions. This is a useful fact for certain applications, for example, when dealing with the dependence encountered in DNA microarrays.

Dependence in DNA Microarrays

I hypothesize that the most likely form of dependence between the genes encountered in DNA microarrays is weak dependence, and more specifically, "clumpy dependence"; that is, the measurements on the genes are dependent in small groups, each group being independent of the others. There are two reasons that make clumpy dependence likely. The first is that genes tend to work in pathways, that is, small groups of genes interact to produce some overall process. This can involve just a few to 50 or more genes. This would lead to a clumpy dependence in the pathway-specific noise in the data. The second reason is that there tends to be cross-hybridization in DNA microarrays. In other words, the signals between two genes can cross because of molecular similarity at the sequence level. Cross-hybridization

would only occur in small groups, and each group would be independent of the others. Typically microarrays measure the expression levels on 3000 to 30,000 genes, and each gene makes up a p-value. Therefore, given the clumpy dependence and large number of genes, I expect Theorem 2 and Corollary 1 to be relevant for the problem of detecting differential gene expression.

Many assumptions that have been made for modeling microarray data have yet to be verified. Hopefully evidence either for or against these assumptions will emerge. I have given a plausibility argument for the assumptions in Theorem 8.3 and Corollary 1. I have also provided numerical evidence in Storey et al. (2002) and Storey (2002b). GDS have stressed the dependence between the genes, not only in this article but in Dudoit et al. (2002b) as well. I leave it as a challenge to them to provide evidence from real microarray data that the aforementioned assumptions do not hold. I have not been able to find it myself. Keep in mind that one can cluster microarray data and see that many genes are in fact related, but this is very different than *stochastic* dependence, especially the type that would violate the assumptions I have argued are true.

Q-values Give Gene-specific Significance ... from a Global View

Given these results and arguments, I would like to suggest a useful way to use the information among all q-values. Since we can essentially consider all estimated q-values simultaneously, one can make various plots in order to find a useful q-value cut-off. The estimated q-values also give a gene-specific measures of significance. In Efron et al. (2001), we approached the problem of detecting differential gene expression from a Bayesian framework, where the posterior probability that a gene is not differentially expressed conditional on its observed statistic can be calculated. We called this quantity a "local false discovery rate" in the sense that it gives proportion of false positives among genes in a small neighborhood around the observed gene. We also used the main result from Storey (2001) to relate these posterior probabilities to the pFDR. This posterior probability is also a gene-specific measure of significance. Therefore, in a sense the qs-value and the traditional posterior probability are natural competitors for gene-specific measures of significance.

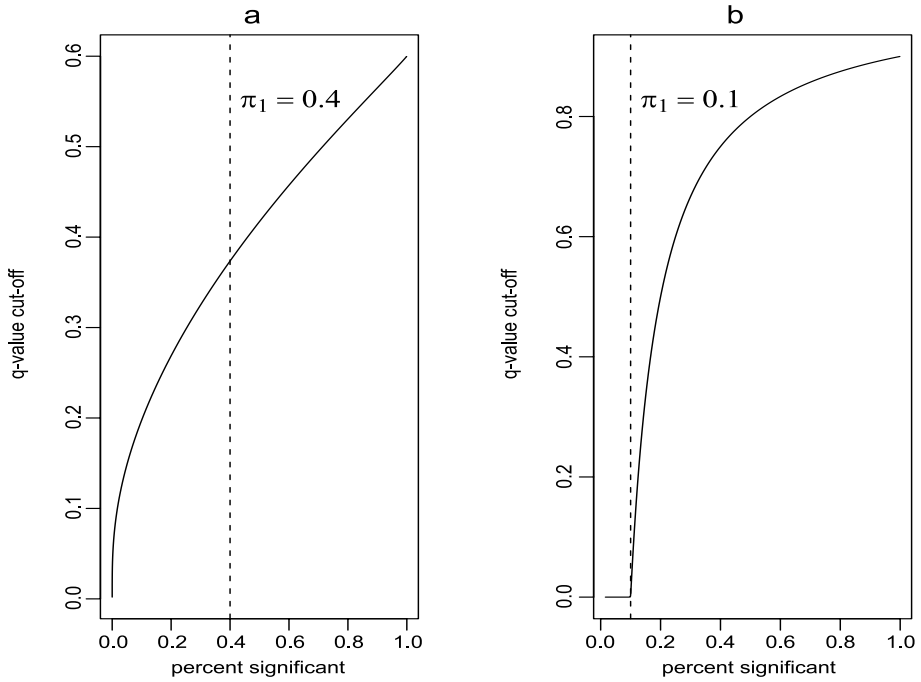

Figure 1: *A plot of the q-value cut-off versus the percent of data called significant with this cut-off when (a)* $\pi_1 = 0.4$, $\mu = 1$ *and (b)* $\pi_1 = 0.1$, $\mu = 6$.

Suppose we observe independent statistics $Z_1, Z_2, \ldots, Z_m$, distributed as a null $N(0,1)$ with probability π_0 and an alternative $N(\mu,1)$ with probability $\pi_1 = 1 - \pi_0$. Also suppose that we use a right-sided significance rule. Suppose $\pi_1 = 0.4$, $\mu = 1$, and we observe $z_i = 2$. Then its local false discovery rate is $\Pr(H_i = 0|Z_i = 2) = 0.25$ and its q-value is $\Pr(H_i = 0|Z_i \geq 2) = 0.18$. Now if we consider $\pi_1 = 0.1$ and $\mu = 6$ with the same observed statistic $z_i = 2$, we get a local false discovery rate of $\Pr(H_i = 0|Z_i = 2) = 0.9997$ and q-value of $\Pr(H_i = 0|Z_i \geq 2) = 0.17$. Therefore, by changing two important parameters we end up with totally different local false discovery rates, but very similar q-values. Clearly, we would not want to call $z_i = 2$ significant in the latter case, but perhaps it is reasonable to in the former case.

With this limited information, it appears that the q-value is not a very good gene-specific measure of significance. Is this a fair assessment? In

the context of looking at a single gene in the marginal sense, then this is a fair assessment. But in the context of the global problem of detecting differentially expressed genes, then this is absolutely not a fair assessment. It also appears to be useful to consider both the q-value and the posterior probability at the same time, but this is both difficult and it requires one to adopt the Bayesian framework.

We show a global use of the q-values does not require one to incorporate these Bayesian quantities. Consider Figure 1 where the percentage of statistics rejected has been plotted versus its corresponding q-value cut-off for both of the above scenarios. The π_1 values are denoted in each plot. From these plots, one can see the local information contained in the q-values is quite different in the two scenarios. Specifically, it can be seen in panel **b** that a q-value cut-off of 0.17 is completely unreasonable, whereas in panel **a** this is not as clear. In panel **b**, one can see that the q-value is virtually zero when 10% of the data have been rejected panel **b**. This information used in conjunction with that fact that $\pi_1 = 10\%$ makes it immediately clear that about 10% of the data being rejected is most reasonable. From panel **a**, this is not the case. Therefore, by considering all q-values simultaneously as well as π_1 in the spirit of Figure 1, one can see which cut-offs make sense. We can do this without being Bayesian and without introducing a totally new quantity.

In the methodology of Storey (2002a), one can obtain estimates of π_1. By generating only 3000 observations from each of these cases, I estimate $\widehat{\pi}_1 = 0.36$ when $\pi_1 = 0.4$ and $\mu = 1$. When $\pi_1 = 0.1$ and $\mu = 6$, I estimate $\widehat{\pi}_1 = 0.10$. The q-value plots are also very similar to the idealized versions in Figure 1. Therefore, it is not clear that the posterior probability (i.e., local false discovery rate) is always necessary.

Peter H. Westfall
Texas Tech University
Lubbock, Texas, USA

The Influence of John Tukey

Among his many other notable contributions to statistics, the late John Tukey also deserves credit for some of the ideas behind this article of Ge, Dudoit and Speed (hereafter GDS). In the late 1980's, several pharmaceu-

tical companies including Merck & Co. supported my development of the software that was eventually to become the SAS/STAT procedure PROC MULTTEST. It was at the urging of the Merck statisticians, partly because of their consultations with John Tukey, that the permutation-based "MinP" method was developed and incorporated.

The MinP and MaxT methods are FWER-controlling procedures. In the 1990's, Tukey strongly supported FDR-based methods, and he deserves some credit for this line of research as well.

While it would be ludicrous to compare myself to Tukey, I must confess that I, like he, have gravitated somewhat toward FDR-based methods in these golden years of my life (my 40s). Nevertheless, I will confine my comments to aspects of the FWER-controlling MinP and MaxT procedures, and hope that others will provide comments on the FDR and Bayesian methods.

History of PROC MULTTEST

Since the "R" software mentioned by GDS also goes by the name "MULTTEST" some brief comments concerning the origin the name seem to be in order. Originally, the procedure was to be applied to multivariate binary data only, and the procedure was dubbed "PROC MBIN" (Westfall et al., 1989). Then, in later updates, we realized that the methods were easily adaptable to continuous responses, and changed the name to "PROC MTEST" (Westfall et al., 1990). The ultimate goal of the pharmaceutical companies who supported the software development was for it to become a SAS-supported package, since SAS is so heavily used in pharmaceutical and regulatory environments. SAS adopted the procedure in 1992; however, the name "MTEST" was already in use, and we chose the name "PROC MULTTEST" instead.

GDS's MinP Method for Continuous Response

At the urging of statisticians Joseph Heyse, Keith Soper and Dror Rom at Merck in the 1980's, the permutation-based MinP approach described by GDS was made an integral part of PROC MULTTEST since its inception, but only for binary response data. The distributions of the marginal

(raw) p-values are calculated analytically from hypergeometric, multivariate hypergeometric, or analytically convolved multivariate hypergeometric distributions, depending upon the application, and these distributions are stored in tables, one for each variable. There is no need for "double permutation" in these cases, as the p-values for each permuted data set are computed via table look-ups. However, in cases where the binary totals are large, the evaluation of all configurations can be too time-consuming; in such cases continuity-corrected normal approximations are used and perform very well.

With continuous data, the permutation distributions do not collapse nicely to hypergeometric types, so the marginal distributions cannot be tabulated simply; we therefore encoded the MaxT method in PROC MULTTEST for this case. GDS are to be congratulated providing an excellent algorithm for the simultaneous evaluation of marginal and joint probabilities, the MinP method. As they indicate, there are many possible extensions, notably to bootstrap multiple testing, which can offer greater flexibility than the permutation-based methods.

Pesarin (2001, pp. 143–147) develops a similar "one-pass" permutation algorithm; however, he does not consider the computational improvements of GDS involving storage and sorting, both of which are very important for gene expression data where m is large.

Closure, Strong Control, and Exact Control

GDS point out differences in strong vs. exact control; the closure principle of Marcus et al. (1976) unifies these concepts. Using closure, one rejects a single hypothesis H_i at the FWER=α level if $H_{\mathcal{M}_0}$ is rejected for all $\mathcal{M}_0 \supseteq \{i\}$, $\mathcal{M}_0 \subseteq \mathcal{M}$; each $H_{\mathcal{M}_0}$ must be tested using a method that has "exact control" at the unadjusted α level to ensure strong FWER control for the individual hypotheses H_i. While closure generally requires evaluation of the $2^m - 1$ "exact" tests, there are enormous simplifications for MaxT- and MinP-based tests when subset pivotality holds. In the case of MinP-based tests, we have $P(\min_{i\in\mathcal{M}_0} P_i \le x|\mathcal{M}_0) = P(\min_{i\in\mathcal{M}_0} P_i \le x|\mathcal{M}) \le P(\min_{i\in\mathcal{M}_1} P_i \le x|\mathcal{M}) = P(\min_{i\in\mathcal{M}_1} P_i \le x|\mathcal{M}_1)$ when $\mathcal{M}_0 \subseteq \mathcal{M}_1$. Thus, significance of a test of $H_{\mathcal{M}_1}$, where $\min_{i\in\mathcal{M}_1} = p_1$, also implies significance of all tests of $H_{\mathcal{M}_0}$for which $\mathcal{M}_0 \subseteq \mathcal{M}_1$ and $\min_{i\in\mathcal{M}_0} = p_1$. This fact implies that only m tests (those involving the ordered p-values) are needed

for the closed testing procedure, rather than $2^m - 1$ tests. The closed testing procedure using MinP reduces to the step-down method described by GDS, and strong control of FWER follows. Similar arguments apply for the MaxT method and the ordered $|t_i|$-values.

Appeal to closure helps to answer some puzzling questions about the methodology:

- Why does MaxT control the FWER when inexact marginal tests are used?
- What happens if marginal distributions are identical but joint distributions are not?

The answer to the first question is simple. As long as each test of $H_{\mathcal{M}_0}$in the closure is exact, the procedure controls the FWER strongly. In the MaxT method, tests of $H_{\mathcal{M}_0}$use the permutation distribution of the statistic $\max_{i\in\mathcal{M}_0} |T_i|$ and are therefore exact permutation tests (Puri and Sen, 1971, pp. 66-70), provided that $H_{\mathcal{M}_0}$ refers to identical joint distributions. Since closure allows us to focus on individual tests of composite hypotheses, which are more widely studied and more readily understood than multiple testing procedures, the second question can be answered by studying composite tests, and conditions under which they are exact. To answer it, we must first answer the question, "What is the null hypothesis $H_{\mathcal{M}_0}$"? In GDS, the definition given is $H_{\mathcal{M}_0} = \cap_{i\in\mathcal{M}_0}\{H_i = 0\}$, where $\{H_i = 0\}$ could be interpreted to mean that the marginal distributions of X_i are identical for all levels of the covariate Y. While this is a precise definition of $\{H_i = 0\}$, the definition $H_{\mathcal{M}_0} = \cap_{i\in\mathcal{M}_0}\{H_i = 0\}$ is imprecise since the joint distributions are not mentioned. Two possible definitions consistent with this statement are (a) the $|\mathcal{M}_0|$-dimensional joint distributions of $\{X_i; i \in \mathcal{M}_0\}$ are identical for all levels of the covariate Y, or (b) the marginal distributions of the X_i are identical for all levels of the covariate Y, all $i \in \mathcal{M}_0$, but the joint distributions of $\{X_i; i \in \mathcal{M}_0\}$ are otherwise arbitrary for the various levels of Y. Under definition (a), the MinP and MaxT methods strongly control the FWER as mentioned above. However, under definition (b), the level of the test is the supremum of the probability of rejection over all multivariate distributions satisfying the marginal constraints, and in this situation the permutation test of $H_{\mathcal{M}_0}$ can yield excess type I errors. The problem is usually minor, but one may consider the

extreme case of two-sample data, where all gene expression measurements are standard normal (hence all nulls are true), but the measurements in group 1 are perfectly correlated, while the measurements in group 2 are independent. Excessive type I errors occur when the sample size in group 1 is large and that of group 2 is small, in the reverse case of imbalance the test is too conservative, and in the balanced case the test is almost exact.

Thus, strictly speaking, $H_{\mathcal{M}_0}$ must be defined as in (a) above to allow determination of affected genes with strong FWER control, using either MinPor MaxT, with permutation tests. The problem and its consequences are similar to the problem of heteroscedasticity in the two-sample univariate analysis: the maximum type I error rate of the two-sample t test is greater than α under heteroscedasticity, particularly with great imbalance. Further study is needed to assess both the extent and relevance of this problem for gene expression data, both in the case of FWER control and FDR control. This problem is a limitation of multiple permutation tests; an approximate solution is to use separate-sample bootstrap multiplicity adjustments; see Westfall and Young (1993, pp. 88–91) for the univariate case, and Pollard and van der Laan (2003) for the multivariate case with application to gene expression data.

Weighting and Serendipity

It is curious that the MaxT approach, with its apparent inconsistency of using inexact marginal tests in conjunction with exact joint tests, seems to work better than the MinP method in many cases. This phenomenon may be understood in terms of *implicit weighting* of the tests caused by different distributions. For a test to be significant using the single-step MinP method, we require $p_i \leq c_\alpha$, where c_α is the α quantile of the MinP distribution.

Using permutational p-values, the distributions of the p_i are close to uniform (with differences due to discreteness) under $\{H_i = 0\}$, thus $P(P_i \leq c_\alpha | H_i = 0) \approx c_\alpha$, for all i. Typically $c_\alpha \approx \alpha/m$ for FWER-controlling methods, and it may be difficult or sometimes impossible for the permutation distribution to achieve such a low critical point for large m.

In contrast, for a test to be significant using the single-step MaxTmethod, we require $|t_i| \geq d_\alpha$, where d_α is the $1-\alpha$ quantile of the null MaxT distri-

bution. In this case there is an imbalance that allows some hypotheses to be tested at higher marginal levels, thus giving possible significances with MaxT that are impossible with MinP. The marginal significance level for H_i, $P(|T_i| \geq d_\alpha \,|\, H_i = 0)$, may vary greatly from test to test, depending upon the distribution of the data X_i.

For example, suppose X_1 is double exponential and X_2 is uniform, independent of X_1, with samples of 5 bivariate observations in each of two groups. The 95% percentile of the distribution of $\max\{|T_1|, |T_2|\}$ is (via massive simulation) $d_\alpha = 2.63$, and (also by simulation) $P(|T_1| \geq 2.63) = 0.0157$ and $P(|T_2| \geq 2.63) = 0.0348$.

The MaxT method thus appears to "reward" distributions that are less outlier-prone. Whether this is a desired emphasis or not, the biologists may decide. To me it seems useful to give more weight to genes that consistently replicate, with no outliers. Of course, outliers themselves should be flagged to evaluate reliability of the experiment, and to identify unexpected gene activity, but these issues are separate from the problem at hand.

Something similar occurs when using the MinP method with binary response variables. Westfall and Wolfinger (1997) note that, when using the MinP method, variables with small marginal success rates are automatically discarded, thereby allowing higher power for variables with larger rates.

Again there is serendipity: if a variable is observed to have only 1 success across both treatment groups, it seems pointless to bother testing it, so why should it contribute to multiplicity adjustment?

While it is lucky that the MaxT method (and MinP for binary data) provide reasonable differential weightings, it is often desired to provide a more directed analysis, using *a priori* weightings (Westfall et al., 1998; Westfall and Soper, 2001) or weightings based on ancillary statistics (Westfall et al., 2003).

My final comment about balance concerns the use of $|t_i|$ as a test statistic. In cases where the distribution of T_i is skewed, the $|t_i|$-based test will reject more often in one of the tails. The logic that suggests balance *across* variables also suggests balance *within* a variable. It may be possible to use the methods of GDS with p-values $p_i = 2\min\{P(T_i \geq t_i \,|\, H_i = 0), P(T_i \leq t_i \,|\, H_i = 0)\}$ for better balance.

Rejoinder by Y. Ge, S. Dudoit and T. P. Speed

We are grateful to all the contributors for their stimulating and insightful comments. Together they make a substantial contribution to our understanding of algorithms for step-down minP adjustment, and more generally to the practice of multiple testing in microarray data analysis. We have arranged our response by topic rather than discussant, in the hope of concentrating on the common themes. Before doing so we remark that we did not attempt in our paper to give a comprehensive review of multiple testing in microarray data analysis. Rather, we focussed on those areas in which resampling played a central role. We apologize for not mentioning much important work in the wider field, some of which has come up in the discussion.

Family-wise error rates

We thank Dr. Westfall for his valuable summary of the history of multiple testing, especially for his outline of John W. Tukey's influence, which nicely complements the recent paper (Benjamini and Braun (2002)). We found his comments on the development of the SAS procedure PROC MULTTEST very interesting, as we did his mention of the fast algorithm for minP adjustment in the binary case. Note that our fast algorithm can also be used in binary case, and could be useful when the binary totals are large. Dr. Grant discussed a natural approach to implementing the algorithm in Box 3 without using double permutations, and Dr. Westfall also mentioned the "one pass" permutation algorithm developed by Pesarin (2001). As we indicated in Section 4.4 of our paper, analogous reasoning led independently to our new fast algorithm. We simply went a little further, speeding up the running time and reducing the storage space. Dr. Westfall explains clearly why maxT has a somewhat better performance than minP, and why it still controls FWER in the strong sense. He describes how maxT gives an "implicit weighting" to different genes, and this leads us to observe that in situations where *a priori* weighting is desirable, the algorithm of Box 4 can be easily modified to apply.

False discovery rates

We appreciate Dr. Storey's response to the point raised by Dr. Glonek and Dr. Solomon concerning the usefulness of q-values. FDR or pFDR need to be considered in the global sense, as these measures focus on the overall error rate, not on particular hypotheses. Indeed Finner and Roters (2001) discuss "cheating" with FDR. For example, if one wishes to reject a particular hypothesis, say H_1, one can take that hypothesis together with 99 other hypotheses which will be rejected with probability one, say $H_2, \cdots, H_{100}$. The FDR for rejecting the family of 100 hypotheses is no greater than 1%, independent of the value of the test statistic for H_1. In this contrived example, the 1% FDR applies to all 100 hypotheses (in the global sense), and says nothing about the hypothesis H_1. One practical consequence of this observation is that in real analyses, researchers should only include in multiple testing procedures genes whose differential expression status is unknown.

SAM, pooling, estimation, and strong control

Dr. Storey describes a nice connection between the BH and SAM procedures when computing raw p-values by $p_i = \mathrm{E}[V^*(\tilde{\Delta}_i)]/m$. This prompts us to make a few comments about his contribution, especially as it seems that we dealt with SAM too briefly. Firstly, we feel that the assumption of every gene having the same null distribution needs more careful scrutiny. To be sure, pooling will give higher resolution than computing p-values within genes, but is it justifiable? Pollard and van der Laan (2002) have shown evidence that the null t-statistics from different genes can have different distributions, and in our own experience, the assumption of identical distributions for every gene has been problematic. In Dr. Storey's view, "subset pivotality" requires the same null distribution of each gene, but we disagree. "Subset pivotality" says that joint distribution $(T_{i_1}, \cdots, T_{i_k}) \mid (H_{i_1}, \cdots, H_{i_k})$ is identical with the joint distribution $(T_{i_1}, \cdots, T_{i_k}) \mid H_{\mathcal{M}}$ for any subset $\{i_1, \cdots, i_k\} \subseteq \mathcal{M}$. If each T_i or P_i is computed within gene i, then surely subset pivotality holds. It is not something that depends on the data. The T_i may have different distributions and may be correlated. For example, correlated T_i can be found in example 2.1 on page 43 in Westfall and Young (1993). Using the argument in that

example, we can also show that subset pivotality can hold for T_i having different marginal distributions. A different concern with pooling is that it may disrupt subset pivotality, and hence our ability to demonstrate strong control. A related point is that the BH procedure is currently guaranteed to deliver strong control only under the assumption that the null p-values are independent, or that they satisfy a positive regression dependency condition. These assumptions may not be valid after pooling, as the p-values will no longer depend solely on data on the individual genes. The reasons for our concerns here should now be evident: we prefer to make assumptions about the data which are clearly reasonable ones, and we want to retain the ability to demonstrate strong validity so that any p-values can be taken seriously. Let us elaborate a little.

For us there are usually two steps in deriving a multiple testing procedure: first, use some heuristics, most of which come from estimation, to find a rejection procedure, and second, demonstrate strong control. For example, the BH procedure can perhaps be best explained using Seeger (1968) as motivation. His idea was to find genes for which (in the notation of Section 5.2) $V(p_i)/R(p_i) \leq \alpha$. When mp_i is used to estimate $V(p_i)$, we are led directly to the BH procedure. Historically, neither Simes (1986) nor Benjamini and Hochberg (1995) gave this kind of motivation. However, Benjamini and Hochberg's elegant proof of strong control FDR greatly accelerated the spread of the FDR concept. The SAM procedure involves estimation based on heuristics derived from the definition of FDR. But SAM computes its p-values after pooling across genes, which may modify the dependence structure among the null genes, and so SAM can be very different from the BH procedure, which is why we think its strong control of FDR still needs to be demonstrated, despite its close connection with BH.

In Section 3 of his comments, Dr. Storey presents some asymptotic results. However, it remains necessary for him to demonstrate strong control of the FDR there, i.e., to show that for any given α, rejecting the FDR-adjusted p-values no greater than α will lead to

$$\sup_{\mathcal{M}_0 \subseteq \mathcal{M}} \mathrm{E} \frac{\sum_{i \in \mathcal{M}_0} I(\tilde{p}_i \leq \alpha)}{\sum_{i \in \mathcal{M}} I(\tilde{p}_i \leq \alpha)} \leq \alpha. \tag{1}$$

Westfall & Young's step-down minP and maxT have been proved to deliver strong control in this sense for FWER. Dr. Storey and colleagues

have also demonstrated this kind of control of FDR when the null p-values are independent, see Storey et al. (2002). In that paper they also prove asymptotic control of FDR when $V(t)/m_0 = \#\{i \in \mathcal{M}_0 : P_i \leq t\}/m_0$ and $S(t)/m_1 = \#\{i \in \mathcal{M}_1 : P_i \leq t\}/m_1$ converge almost surely to $G_0(t)$ and $G_1(t)$ respectively, and $\lim_{m\to\infty} m_0/m = \pi_0$ exists. These conditions don't require the assumption that each gene has the same null or alternative distributions, but they do require that $V(t)/m_0$ and $S(t)/m_1$ should behave as if each gene has the same null (G_0) or alternative distribution (G_1). When these assumptions are valid, their asymptotic control of FDR can be written

$$\lim_{m\to\infty} \mathrm{E}\frac{\sum_{i\in\mathcal{M}_0} I(\tilde{p}_i \leq \alpha)}{\sum_{i\in\mathcal{M}} I(\tilde{p}_i \leq \alpha)} \leq \alpha \text{ when } m_0/m \to \pi_0. \tag{2}$$

This result may not mean strong control of FDR in the sense of equation (1) where we need to compute the maximum of FDR for all possible choices of m_0 or $\mathcal{M}_0$ in equation (2) before taking the limit. It would also be good to prove strong control of FDR here for dependent data. Again, if p-values are computed by pooling, then strong control may be still an issue as pooling may destroy their assumptions. For example, if the genes have a block independence structure and we are using two sample t-statistics, then pooling will remove the block-independence of the computed p-values, as every p_i will now involve the permuted distribution of all test statistics $T_1, \cdots, T_m$. Dr. Storey reminds us that the expectation of the sum of indicator variables stays the same, even the data are dependent, which is of course true, but the variance of that sum can be different under different dependence structures. That is the reason why we computed the individual p-values within genes for the Storey procedure. Despite all of the foregoing, we see all of the procedures discussed, including SAM, as having value, and also we will gain more confidence in their use from simulation results.

Test statistics, resampling and empirical Bayes

The two sample t-statistic is the most commonly used statistic for comparing the treatments and controls, and we used the simple $|t_i|$-based rejection rule. However, this may not be the best statistic for identifying differentially expressed genes: improvements may be possible. Dr. Westfall mentioned using balanced one-sided tests, computing $p_i = 2 \cdot \min(\Pr(T_i \geq t_i \mid H_i = 0), \Pr(T_i \leq t_i \mid H_i = 0))$. Mr. Maindonald suggested improving the crude

variance estimates of two sample t-statistic. In our view the test statistic needs to be carefully chosen, and we leave deciding the best one to future research. The same applies for the resampling strategy. We adopted the most commonly used one, namely, permutations. Other resampling strategies are possible, for example, the bootstrap. As long as the test statistics are computed and the resampling performed within genes, the algorithms in our paper should apply.

We thank Dr. Kendziorski's for her comments on the empirical Bayes approach. This approach is definitely a useful tool for identifying differentially expressed genes with microarray data. We look forward to theory demonstrating her feeling that a measure of control of FDR is automatically obtained using posterior probabilities. It will be good if it is true.

Dr. Glonek, Dr. Solomon, Mr. Maindonald, Dr. Grant, and many others are concerned that the FWER approach lacks power if the number of samples (microarrays) is limited, especially when adjustments are based on resampling. Also, the resampling-based approach may become more sophisticated and computationally intensive when the experimental design is more complicated than the simple replicated treatment and control used in this paper. In these cases, it does seem that some modeling of the data will be required to answer the questions of interest, possibly empirical Bayes modeling. If the number of permutations is quite limited, for example, when only three or four slides are available, a parametric form of resampling could be applied. Alternatively, graphical methods such as the use of normal qq-plots may be used to identify the genes of interest. We do feel, however, that it is probably unreasonable to expect essentially model-free p-value adjustments providing strong control in such situations.

New directions in multiple testing

We predict that FDR or one of its variants will come to enjoy wide application in microarray data analysis, for these procedures provide a good balance between the raw p-values and the stringent FWER adjusted p-values. The family-wise (also called experiment-wise) error rate is probably not very useful when a good proportion of the genes are expected to be differentially expressed, especially when testing many thousands of genes in an

exploratory analysis. FWER may be more useful in confirmatory analyses, when enough replicates are available. A potentially useful variant of FWER comes from allowing some number, say k falsely rejected hypotheses. Then we can attempt to control $\Pr(V > k)$, noting that when $k = 0$, this is equivalent to FWER. Korn et al. (2001) have begun work in this direction, and it seems promising. They also use the concept of false discovery proportion (FDP), this being defined as V/R, the observed FDR. The control of FDP is pleasing as it gives a procedure such that given one γ, say 0.1, $\Pr(V/R > \gamma) \leq \alpha$. In some ways this might be a better criterion than FDR, as FDR considers the overall expectation, not simply what is happening in the data we have. FDP relates directly to the information in the observed data, which is our primary interest.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96:6745–6750.

Benjamini, Y. and Braun, H. (2002). John W. Tukey's contriubtions to multiple comparisons. *The Annals of Statistics*, 30(6):1576–1594.

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

Benjamini, Y. and Hochberg, Y. (2000). The adaptive control of the false discovery rate in multiple hypotheses testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83(403):679–686.

Berry, D. (1988). Multiple comparisons, multiple tests, and data dredging: A bayesian perspective. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, eds., *Bayesian Statistics*, vol. 3, pp. 79–94. Oxford Univeristy Press.

Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O., and Relman, D. A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proceedings of the National Academy of Sciences*, 99(2):972–977.

Buckley, M. J. (2000). *The Spot user's guide*. CSIRO Mathematical and Information Sciences. `http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm`.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, 10(12):2022–2029.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–685.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2002a). Multiple hypothesis testing in microarray experiments. Submitted, available UC Berkeley, Division Biostatistics working paper series: 2002-110, http://www.bepress.com/ucbbiostat/paper110.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111—139.

Dunn, O. J. (1958). Estimation of the means of dependent variables. *The Annals of Mathematical Statistics*, 29:1095–1111.

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

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000). Microarrays and their use in a comparative experiment. Tech. Rep. 37B/213, Department of Statistics, Stanford University.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. *Biometrical Journal*, 8:985–1005.

Genovese, C. and Wasserman, L. (2001). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society, Series B*, 57:499–517.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Jogdeo, K. (1977). Association and probability inequalities. *Annals of Statistics*, 5(3):495–504.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. In press.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2001). Controlling the number of false discoveries: Application to high dimensional genomic data. Tech. Rep. 003, National Cancer Institute, Division of Cancer Treatment and Diagnosis. `http://linus.nci.nih.gov/~brb/TechReport.htm`.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Springer Verlag, New York, 2nd ed.

Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680.

Manduchi, E., Grant, G. R., McKenzie, S. E., Overton, G. C., Surrey, S., and Stoeckert Jr, C. J. (2000). Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, 16:685–698.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrics*, 63:655–660.

Morton, N. E. (1955). Sequential the tests for detection of linkage. *American Journal of Human Genetics*, 7:277–318.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2003). Optimal sample size for multiple testing: the case of gene expression microarrays. technical report, department of biostatistics. Tech. rep., The University of Texas M.D. Anderson Cancer Center.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96:9212–9217.

Pesarin, F. (2001). *Multivariate permutation tests with applications in biostatistics*. John Wiley and Sons, Chichester.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23:41–46.

Pollard, K. and van der Laan, M. (2002). Resampling-based methods for identification of significant subsets of genes in expression data. *U.C. Berkeley Division of Biostatistics Working Paper Series.* Working Paper 121, `http://www.bepress.com/ucbbiostat`.

Pollard, K. and van der Laan, M. (2003). Parametric and nonparametric methods to identify significantly differentially expressed genes. Manuscript.

Puri, M. and Sen, P. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24:227–234.

Seeger, P. (1968). A note on a method for the analysis of significance en masse. *Technometrics*, 10(3):586–593.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46:561–584.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62:626–633.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Sorić, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):608–610.

STOREY, J. D. (2001). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*. In press.

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

STOREY, J. D. (2002b). *False Discovery Rates: Theory and Applications to DNA Microarrays*. Ph.D. thesis, Department of Statistics, Stanford University.

STOREY, J. D., TAYLOR, J. E., and SIEGMUND, D. (2002). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*. In press.

STOREY, J. D. and TIBSHIRANI, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Tech. Rep. 2001-28, Department of Statistics, Stanford University.

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

WELCH, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

WESTFALL, P., KRISHEN, A., and YOUNG, S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine*, 17:2107–2119.

WESTFALL, P., KROPF, S., and FINOS, L. (2003). Weighted fwe-controlling methods in high-dimensional situations. Manuscript.

WESTFALL, P., LIN, Y., and YOUNG, S. (1989). A procedure for the analysis of multivariate binomial data with adjustments for multiplicity. In *Proceedings of the 14th Annual SAS® User's Group International Conference*, pp. 1385–1392.

WESTFALL, P. and SOPER, K. (2001). Using priors to improve multiple animal carcinogenicity tests. *Journal of the American Statistical Association*, 96:827–834.

WESTFALL, P. and WOLFINGER, R. (1997). Multiple tests with discrete distributions. *The American Statistician*, 51:3–8.

WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, New York.

WESTFALL, P. H., ZAYKIN, D. V., and YOUNG, S. S. (2001). Multiple tests for genetic effects in association studies. In S. Looney, ed., *Methods in Molecular Biology*, vol. 184: Biostatistical Methods, pp. 143–168. Humana Press, Toloway, NJ.

YEKUTIELI, D. and BENJAMINI, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196.