# Resampling-based multiple testing for microarray data analysis

Yongchao Ge[1], Sandrine Dudoit[2], and Terence P. Speed[1,3]

Jan. 2003

Technical Report # 633

1. Department of Statistics, University of California, Berkeley
2. Division of Biostatistics, University of California, Berkeley
3. Division of Genetics and Bioinformatics,
The Walter and Eliza Hall Institute of Medical Research, Australia

*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`

## 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. In their 1993 book, Westfall & Young 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 & Young (1993) and (b) the false discovery rate developed by Benjamini & Hochberg (1995), both from a frequentist viewpoint; and (c) the positive false discovery rate of Storey (2002), 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.

**Keywords**: multiple testing; family-wise error rate; false discovery rate; adjusted $p$-value; fast algorithm; minP; microarray.

# 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 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.

In their 1993 book, Westfall & Young 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 & Young (1993) step-down minP adjusted $p$-values. A second line of multiple testing is developed by Benjamini & Hochberg (1995). They propose procedures to control the false discovery rate. This was further developed by Storey (2002) 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 & 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, 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

[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.

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 hypotheses while controlling a suitably defined type I error rate (Dudoit, Yang, Callow & Speed 2002, Efron et al. 2000, Efron et al. 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.

| 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.

## 2.2 Type I error rates

**Set-up.** Consider the problem of simultaneously testing $m$ null hypotheses $H_i$, $i = 1, \dots, m$. 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 & 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 $\alpha$ and seek tests which minimize the type II error rate, *i.e.*, maximize *power*, within the class of tests with type I error rate $\alpha$.

| | # 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 & Hochberg (1995).

**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 $\alpha$. This can be achieved by choosing a critical value $c_\alpha$ such that $\Pr(|T| > c_\alpha \mid H = 0) \leq \alpha$ and rejecting $H$ 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 & Hochberg (1995), *i.e.*,
$$\text{FDR} = E(\frac{V}{R} 1_{\{R>0\}}) = E(\frac{V}{R} \mid R > 0) \Pr(R > 0).$$
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 (2002) 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,
$$\text{pFDR} = E(\frac{V}{R} \mid R > 0).$$
Storey (2002) shows that this definition is intuitively pleasing and has a nice Bayesian interpretation (*cf.* the remarks on page 10) 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 most microarray applications, we expect pFDR $\leq$ FWER, apart from the extreme case when $1 =$ pFDR $\geq$ FDR $=$ 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,

$$\text{PCER} \leq \text{FDR} \leq \text{pFDR} \leq \text{FWER} \leq \text{PFER}. \tag{1}$$

**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 $\alpha$ rejection regions $\Gamma_\alpha$ 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

$$p\text{-value}(t) = \min_{\{\Gamma_\alpha :\, t \in \Gamma_\alpha\}} \Pr(T \in \Gamma_\alpha \mid H = 0). \tag{2}$$

In words, the $p$-value is the minimum type I error rate over all possible rejection regions $\Gamma_\alpha$ 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 $\alpha$. 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 $\alpha_i$ are chosen to control a given type I error rate (FWER, PCER, PFER, or FDR) at a prespecified level $\alpha$. 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 & Young 1993, Yekutieli & Benjamini 1999). If interest is in controlling the FWER, the FWER adjusted $p$-value for hypothesis $H_i$ is:

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

Hypothesis $H_i$ is then rejected, *i.e.*, gene $i$ is declared differentially expressed, at FWER $\alpha$ 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 & Benjamini 1999), *e.g.*

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

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), considering the minimum of the type I error rates for all possible rejection regions $\Gamma_\alpha$ containing the observed $T = t$. Let $\text{pFDR}(\Gamma_\alpha)$ be the pFDR when each hypothesis is rejected by the same rejection region $\Gamma_\alpha$. The *q-value* is defined analogously, namely

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha :\, t \in \Gamma_\alpha\}} \text{pFDR}(\Gamma_\alpha). \tag{3}$$

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 $\gamma$ leads to

$$q\text{-value}(p) = \inf_{\{\gamma \geq p\}} \text{pFDR}(\gamma). \tag{4}$$

**Remarks.**
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 & Benjamini (1999). The only difference between $q$-values and single step adjusted $p$-values is that $q$-values consider only the true but unknown $H_{\mathcal{M}_0}$ (exact control), while single step adjustments consider every possible choice of $H_{\mathcal{M}_0}, \mathcal{M}_0 \subseteq \{1, 2, \ldots, m\}$ (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 $\pi_1 = 1 - \pi_0$, $\pi_0$ being the *a priori* probability that a hypothesis is true. Theorem 1 of Storey (2001) states that for all $i$

$$\text{pFDR}(\Gamma_\alpha) = \Pr(H_i = 0 \mid T_i \in \Gamma_\alpha). \tag{5}$$

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), 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 $\alpha$, the Bonferroni procedure, perhaps the best known in multiple testing, rejects any hypothesis $H_i$ with $p$-value less than or equal to $\alpha/m$. The corresponding *Bonferroni single-step adjusted p-values* are thus given by

$$\tilde{p}_i = \min\bigl(mp_i, 1\bigr). \tag{6}$$

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

$$\text{FWER} = \Pr(V > 0) \le \Pr\bigl(\bigcup_{i=1}^{m_0}\{\tilde{P}_i \le \alpha\}\bigr) \le \sum_{i=1}^{m_0}\Pr\bigl(\tilde{P}_i \le \alpha\bigr) \le \sum_{i=1}^{m_0}\alpha/m = m_0\alpha/m \le \alpha. \tag{7}$$

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

$$\tilde{p}_i = 1 - (1 - p_i)^m. \tag{8}$$

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

$$\Pr(V = 0) = \Pr(\bigcap_{i=1}^{m_0}\{\tilde{P}_i \ge \alpha\}) = \prod_{i=1}^{m_0}\Pr(\tilde{P}_i \ge \alpha) = \prod_{i=1}^{m_0}\Pr(P_i \ge 1-(1-\alpha)^{1/m}) = \{(1-\alpha)^{1/m}\}^{m_0}.$$

Therefore,

$$\text{FWER} = \Pr(V > 0) = 1 - \Pr(V = 0) = 1 - (1 - \alpha)^{m_0/m} \le \alpha. \tag{9}$$

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 & 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

$$\tilde{p}_i = \Pr\big(\min_{1\leq l\leq m} P_l \leq p_i \mid H_{\mathcal{M}}\big), \tag{10}$$

where $H_{\mathcal{M}}$ denotes the complete null hypothesis and $P_l$ the random variable for the raw $p$-value of the $l$th 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

$$\tilde{p}_i = \Pr\big(\max_{1\leq l\leq m} |T_l| \geq |t_i| \mid H_{\mathcal{M}}\big). \tag{11}$$

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 (10) 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 & 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 & 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 $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m}$ denote the *ordered raw p-values.* For control of the FWER at level $\alpha$, 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

$$\tilde{p}_{r_i} = \max_{k=1,\ldots,i} \Big\{ \min\big( (m-k+1)\, p_{r_k}, 1 \big) \Big\}. \tag{12}$$

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, $\tilde{p}_{r_1} \leq \tilde{p}_{r_2} \leq \ldots \leq \tilde{p}_{r_m}$, 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

$$\tilde{p}_{r_i} = \max_{k=1,\ldots,i} \Big\{ 1 - (1 - p_{r_k})^{(m-k+1)} \Big\}. \tag{13}$$

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

$$\tilde{p}_{r_i} = \max_{k=1,\dots,i} \Big\{ \Pr\big( \min_{l=k,\dots,m} P_{r_l} \le p_{r_k} \mid H_{\mathcal{M}} \big) \Big\}, \tag{14}$$

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

$$\tilde{p}_{s_i} = \max_{k=1,\dots,i} \Big\{ \Pr\big( \max_{l=k,\dots,m} |T_{s_l}| \ge |t_{s_k}| \mid H_{\mathcal{M}} \big) \Big\}, \tag{15}$$

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

Note that computing the quantities in (14) 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 & 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 $T_1, \dots, T_m$ 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 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 (6), (8), (12), and (13).

**Box 1. Permutation algorithm for raw $p$-values**

For the $b$th 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

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

## 4.2 Step-down maxT adjusted $p$-values

For the step-down maxT adjusted $p$-values of Westfall & Young, 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 maximum $\max_{l=1,\ldots,m} |T_{s_l}|$.) 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 & Young are in general harder to compute as they require the joint null 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. 20). As the time taken for generating one set of raw $p$-values for all genes is about 2 minutes, our estimate of the computation 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

**Box 2. Permutation algorithm for step-down maxT adjusted $p$-values - based on Westfall & 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 $b$th 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 (15) ), 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

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

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

**Box 3. The traditional double permutation algorithm for step-down minP adjusted $p$-values - based on Westfall & 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 $b$th 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 (14) ), 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

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

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 $b$th 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 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.

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

### 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

$$p_{i,b} = \frac{\#\{b' : |t_{i,b'}| \geq |t_{i,b}|\}}{B}. \tag{16}$$

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 (16). However, the idea of computing $p_{i,1}, p_{i,2}, \ldots, p_{i,B}$ simultaneously can be refined as follows. Order the $i$th row of matrix $T$ and let $r_b$, $b = 1, \ldots, B$, be such that $|t_{i,r_1}| \geq |t_{i,r_2}| \cdots \geq |t_{i,r_B}|$. 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

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

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

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

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 × 6,356 × 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.

| | 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 $B$.

Note that we did not consider $n$, the sample size (number of arrays), as 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 & Hochberg (1995) suppose that $V/R = 0$ when $R = 0$, while Storey (2002) 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 & 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 $E(\frac{V}{R}1_{\{R>0\}})$. 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

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m}\{\min(\frac{m}{k}p_{r_k}, 1)\}. \tag{17}$$

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

$$E(\frac{V}{R}1_{\{R>0\}}) \leq \frac{m_0}{m}\alpha \leq \alpha. \tag{18}$$

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 & Hochberg (1995) giving strong control of FDR greatly expanded the popularity of this procedure.

### 5.1.2 FDR under general dependence

Benjamini & Yekutieli (2001) ($BY$) proved that the procedure based on equation (17) 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 $\alpha$, 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

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m}\{\min(\frac{m\sum_{l=1}^{m} 1/l}{k}p_{r_k}, 1)\}. \tag{19}$$

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 (17). 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 (2002) 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 (5), it is easy to see that

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

Since $m\pi_0$ of the $p$-values are expected to be null, $\pi_0$ 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 & Tibshirani (2001) where the notation $\lambda$ is used. Given a suitable $p_0$, a conservative estimate of $\pi_0$ will be

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

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

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

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

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

$$\widehat{\text{pFDR}}_{p_0}(p) = \frac{\hat{\pi}_0(p_0) \cdot p}{\widehat{\Pr}(P \le p) \cdot \widehat{\Pr}(R > 0)} = \frac{W(p_0) \cdot p}{(1 - p_0) \cdot (R(p) \vee 1) \cdot (1 - (1 - p)^m)}. \tag{20}$$

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

$$\widehat{\text{FDR}}_{p_0}(p) = \frac{W(p_0) \cdot p}{(1 - p_0) \cdot (R(p) \vee 1)}. \tag{21}$$

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 (2002) for full details.

### 5.2.2 pFDR under more general dependence

Storey & 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 (20) can also be written in terms of a general family of nested rejection regions $\{\Gamma_\alpha\}$ as

$$\widehat{\text{pFDR}}_{\Gamma_{\alpha_0}}(\Gamma_\alpha) = \frac{\hat{\pi}_0(\Gamma_{\alpha_0}) \cdot \alpha}{\widehat{\Pr}(T \in \Gamma_\alpha) \cdot \widehat{\Pr}(R > 0)} = \frac{W(\Gamma_{\alpha_0}) \cdot \alpha}{(1-\alpha_0) \cdot (R(\Gamma_\alpha) \vee 1) \cdot \widehat{\Pr}(R > 0)},$$

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 $\Gamma_\alpha$ is the level $\alpha$ rejection region. Now consider a general rejection region $\Gamma$, 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 $\Gamma_0$ 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 $\alpha$, $\alpha_0$ 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 $\Gamma_0$ can be defined.

Putting these all together, a conservative estimate of pFDR($\Gamma$), making use of $\Gamma_0$ is

$$\widehat{\text{pFDR}}_{\Gamma_0}(\Gamma) = \frac{W(\Gamma_0) \cdot \overline{R}(\Gamma)}{(m - \overline{R}(\Gamma_0)) \cdot (R(\Gamma) \vee 1) \cdot \widehat{\Pr}(R > 0)} = \frac{W(\Gamma_0) \cdot \overline{R}(\Gamma)}{\overline{W}(\Gamma_0) \cdot (R(\Gamma) \vee 1) \cdot \overline{I}_{\{R(\Gamma) > 0\}}}. \quad (22)$$

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

$$\widehat{\text{FDR}}_{\Gamma_0}(\Gamma) = \frac{W(\Gamma_0) \cdot \overline{R}(\Gamma)}{(m - \overline{R}(\Gamma_0)) \cdot (R(\Gamma) \vee 1)} = \frac{W(\Gamma_0) \cdot \overline{R}(\Gamma)}{\overline{W}(\Gamma_0) \cdot (R(\Gamma) \vee 1)}. \quad (23)$$

### 5.2.3 Estimation of pFDR q-values

Using the definition of $q$-values given in equations (4) and (3), the estimates of the $q$-value corresponding to the ordered $p$-values $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$ are

$$\widehat{q}_{p_0}(p_{r_i}) = \min_{k=i,\ldots,m} \widehat{\text{pFDR}}_{p_0}(p_{r_k}). \tag{24}$$

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

$$\widehat{q}_{\Gamma_0}(t_{s_i}) = \min_{k=i,\ldots,m} \widehat{\text{pFDR}}_{\Gamma_0}(\Gamma_{s_k}). \tag{25}$$

**Remarks.**
Storey (2002) has already pointed out that the FDR estimate based on equation (21) 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 (21) and enforcing step-up monotonicity, FDR-based adjusted $p$-values can be estimated by

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m} \{\min(\frac{m}{k} p_{r_k} \hat{\pi}_0, 1)\}. \tag{26}$$

We call this the *Storey procedure*. Equation (20) 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 (23) and enforced monotonicity for controlling FDR under quite general dependence satisfying ergodic conditions, while the *ST-q procedure* used equation (22) and monotonicity to control pFDR. Details of these procedures are given in Box 5.

Comparing equation (26) with equation (17), it is easy to see that the method proposed by Storey (2002) has advantages over that of Benjamini & Hochberg (1995), since $\hat{\pi}_0$ is less than or equal to 1. This should be no surprise, since equation (17) controls the FDR in the strong sense, while equation (26) controls the FDR in the exact sense, with an estimated $\pi_0$. If we are only considering the FDR in the exact sense, then $\pi_0$ can be estimated, and by noting that $\frac{m_0}{m} = \pi_0$ in equation (18) 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 (7) and (9) to get more powerful single-step Bonferroni and Šidák adjustments. Indeed, Benjamini & Hochberg (2000) proposed a different estimator of $\hat{\pi}_0$, but Storey (2002) 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 (17) and (19). 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 (21) for Storey procedure and (20) 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 & 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 & Tibshirani (2001).

**[Note, Box 5 is placed approximately here]**

### 5.4 Empirical Bayes procedures and the SAM software

Several papers (Efron et al. 2000, Efron et al. 2001, Efron & 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, Shaffer & Boldrick (2002), 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, Yang, Callow & Speed (2002). 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 $i$:

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

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 $s_{1i}^2$ and $s_{2i}^2$ 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, Westfall & Young step-down maxT and 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 (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.

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. 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 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

| 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 | 11:10:27.17 | 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.* 11:10:26.17 means 11 hours 10 minutes and 26.17 seconds

copies), sterol C5 desaturase (2 copies), and a novel EST (1 copy). All changes were confirmed by real-time quantitative RT-PCR as described 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 pre-processing 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. 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 $y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $i = 1, 2$. The $t$-statistic with pooled variance is:

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

(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 $E(T) = n_2(n+1)/2$, $Var(T) = n_1n_2(n+1)/12$, the normalized statistic is:

$$W = \frac{\sum_{j=1}^{n_2} s_{2j} - n_2(n+1)/2}{\sqrt{n_1n_2(n+1)/12}}.$$

(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 (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.

### 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 & 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.

## 9 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.

# 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. & 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. & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci.* **96**: 6745–6750.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statisit. Soc. B* **57**: 289–300.

Benjamini, Y. & Hochberg, Y. (2000). The adaptive control of the false discovery rate in multiple hypotheses testing with independent statistics, *J. Behav. Educ. Statis.* **25**(1): 60–83.

Benjamini, Y. & 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.

Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O. & Relman, D. A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria, *Proc. Natl. Acad. Sci.* **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. & 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. & 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. & Boldrick, J. C. (2002). 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. & Speed, T. P. (2002). 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**: 1095–111.

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

Efron, B., Tibshirani, R., Goss, V. & Chu, G. (2000). Microarrays and their use in a comparative experiment, *Technical report*, Department of Statistics, Stanford University.

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

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. & 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, *Scand. J. Statist.* **6**: 65–70.

Ihaka, R. & 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**: 495–504.

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

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. & 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. & Stoeckert Jr, C. J. (2000). Generation of patterns from gene expression data by assigning confidence to differentially expressed genes, *Bioinformatics* **16**: 685–698.

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

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. & Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proc. Natl. Acad. Sci.* **96**: 9212–9217.

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

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. & 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, *Technical Report 2001-12*, Department of Statistics, Stanford University.

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

Storey, J. D. & Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays, *Technical Report 2001-28*, Department of Statistics, Stanford University.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response, *Proc. Natl. Acad. Sci.* **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. H. & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, John Wiley & Sons.

Westfall, P. H., Zaykin, D. V. & Young, S. S. (2001). Multiple tests for genetic effects in association studies, *in* S. Looney (ed.), *Methods in Molecular Biology*, Vol. 184: Biostatistical Methods, Humana Press, Toloway, NJ, pp. 143–168.

Yekutieli, D. & 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.

**Box 5. A permutation algorithm for the ST-$q$ and ST procedures - based on Storey & Tibshirani (2001) Algorithm 1**

Choose a value $\tau_0$ 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 $\tau_i$.

Compute $R_i = \#\{k : |t_k| \geq \tau_i\}$, and $W_0 = \#\{k : |t_k| \leq \tau_0\}$.

For the $b$th 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

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

Then at $\tau_i$ 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:

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

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

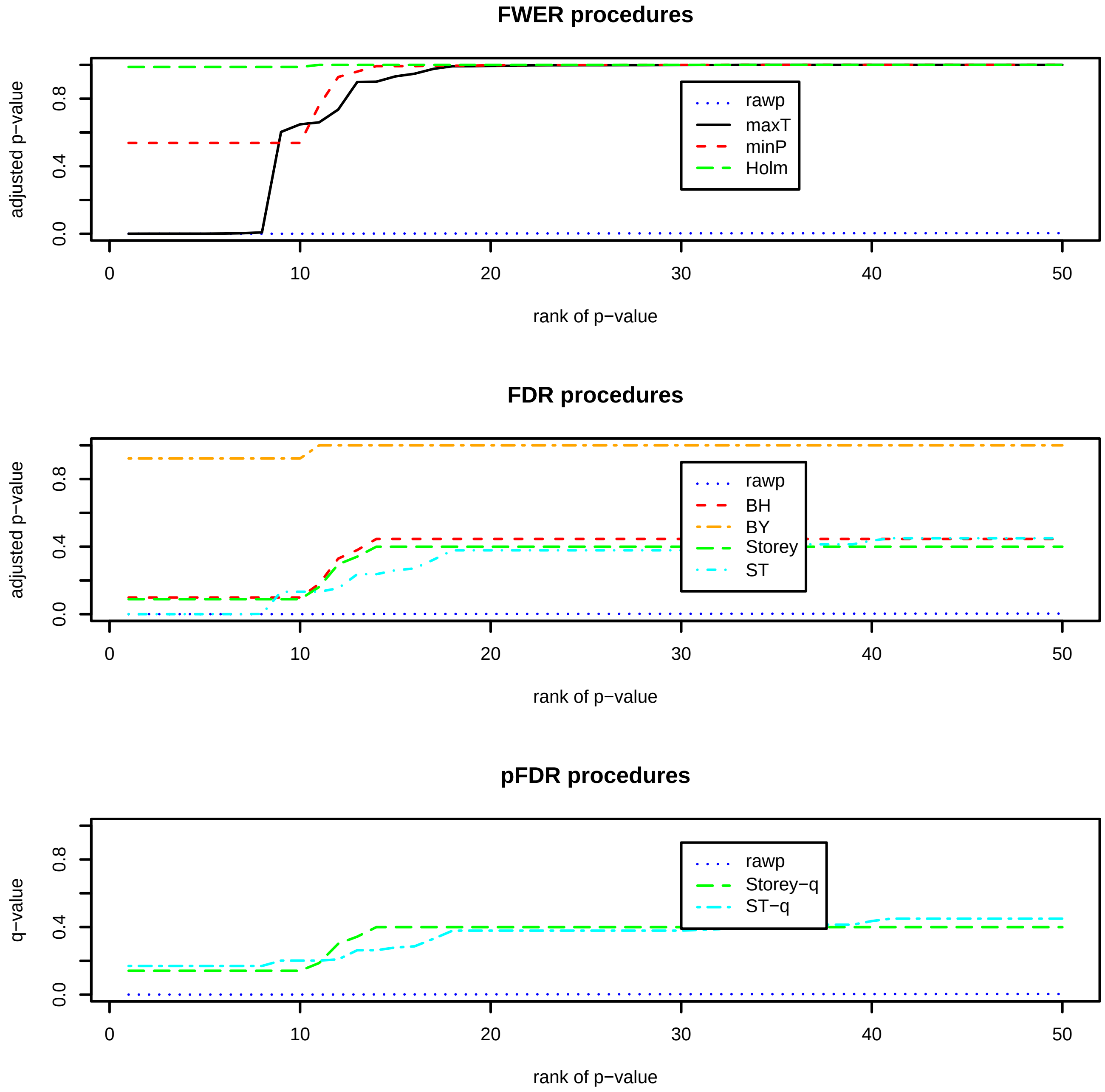

Figure 1: *Apo AI.* Plot of adjusted $p$-values controlling different type I error rates against the rank of the $p$-values. $p$-values were estimated using all $B = \binom{16}{8} = 12{,}870$ permutations. The top, middle and bottom panels are adjusted $p$-values controlling FWER, FDR and pFDR, respectively.

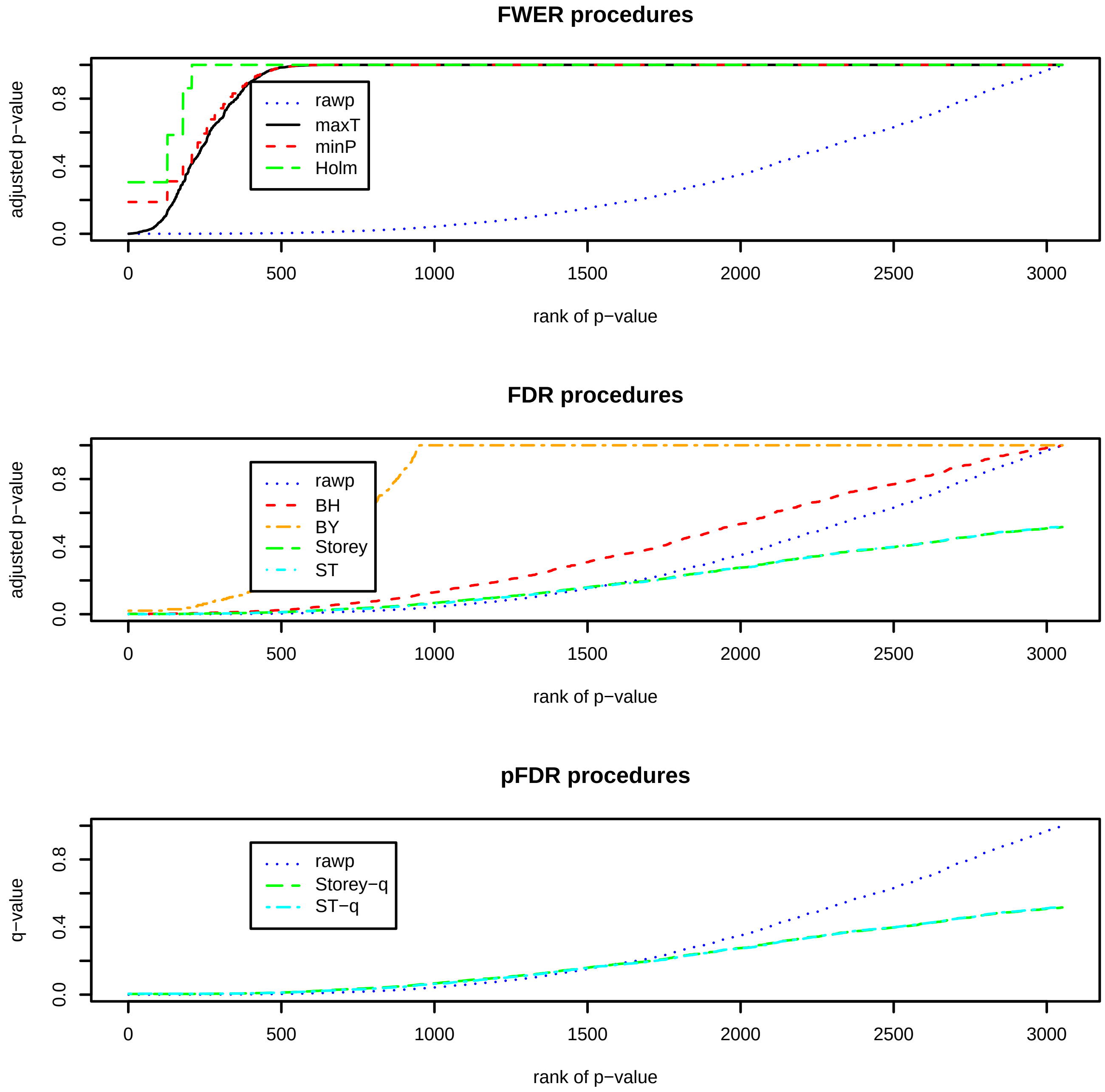

Figure 2: *Leukemia.* Plot of adjusted $p$-values to controlling different type I error rates against the rank of the $p$-values. $p$-values were estimated using $B = 10{,}000$ random permutations. The top, middle and bottom panels are adjusted $p$-values controlling FWER, FDR and pFDR, respectively.

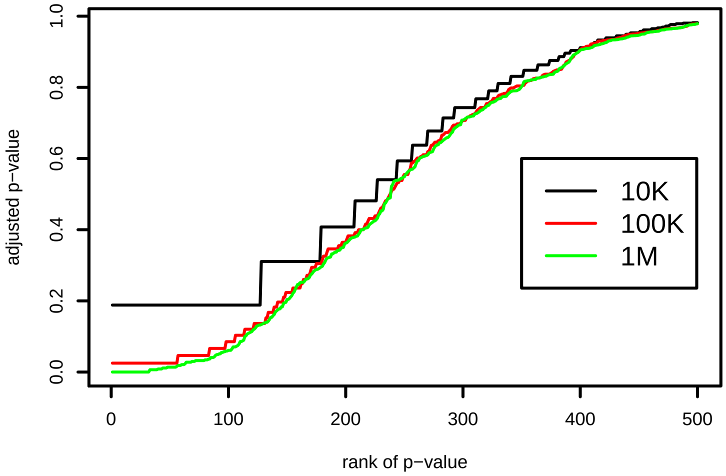

Figure 3: *Leukemia*. Plot of minP adjusted $p$-values against the rank of adjusted $p$-values. $p$-values were estimated based on $B$ = 10,000, 100,000 and 1,000,000 random permutations.

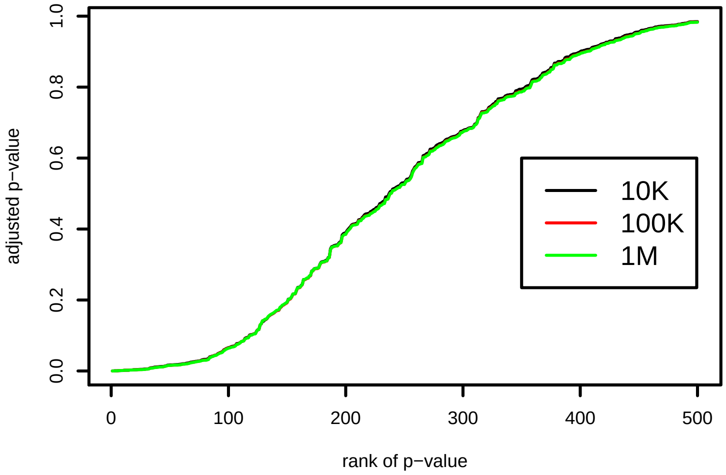

Figure 4: *Leukemia*. Plot of maxT adjusted $p$-values against the rank of adjusted $p$-values estimated using $B$ = 10,000, 100,000 and 1,000,000 random permutations.

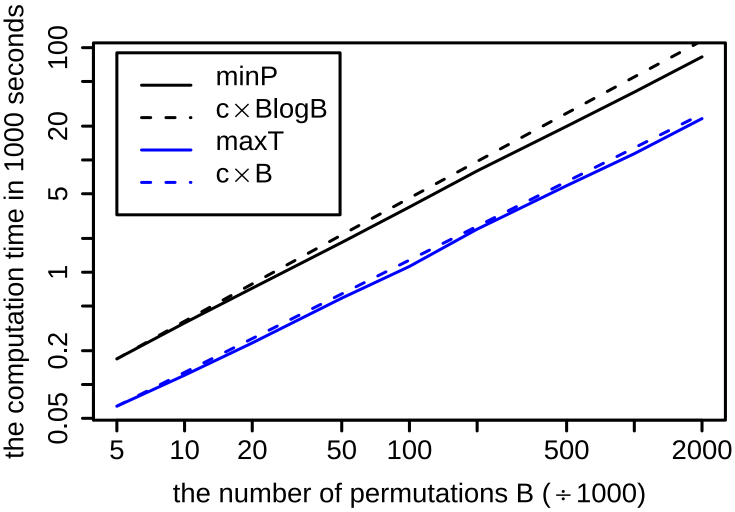

Figure 5: *Leukemia*. Plot of computation time against the number of permutations $B$. The dashed lines are computed using the theoretical run time formulae shown in the legend. The constants $c$ were computed from the timings for $B = 5{,}000$ permutations for maxT and minP separately.