University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 110

Multiple Hypothesis Testing in Microarray Experiments

Sandrine Dudoit* Juliet Popper Shaffer†
Jennifer C. Boldrick‡

*Division of Biostatistics, University of California, Berkeley, sandrine@stat.berkeley.edu
†Department of Statistics, University of California, Berkeley
‡Dept. of Microbiology & Immunology, Stanford University

Multiple Hypothesis Testing in Microarray Experiments

Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick

Abstract

DNA microarrays are a new and promising biotechnology which allows the monitoring of expression levels in cells for thousands of genes simultaneously. An important and common question in microarray experiments is the identification of differentially expressed genes, i.e., genes whose expression levels are associated with a response or covariate of interest. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. As a typical microarray experiment measures expression levels for thousands of genes simultaneously, large multiplicity problems are generated. This article discusses different approaches to multiple hypothesis testing in the context of microarray experiments and compares the procedures on microarray and simulated datasets.

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. DNA microarrays are a new and promising biotechnology which allows 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 question in microarray experiments is the identification of differentially expressed genes, *i.e.*, 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. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. As a typical microarray experiment measures expression levels for thousands of genes simultaneously, large multiplicity problems are generated. In any testing situation, two types of errors can be committed: a false positive, or Type I error, is committed by declaring that a gene is differentially expressed when it isn't, and a false negative, or Type II error, is committed when the test fails to identify a truly differentially expressed gene. When many hypotheses are tested and each test has a specified Type I error probability, the chance of committing some Type I errors increases, often sharply, with the number of hypotheses. In particular, a p–value of 0.01 for one gene among a list of several thousands will no longer correspond to a significant finding, as it is inevitable that such small p–values will occur by chance when considering a large enough set of genes. 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 account for the *joint* distribution of the test statistics. A number of recent papers have addressed the question of multiple testing in microarray experiments (Dudoit et al. 2002, Efron et al. 2000, Golub et al. 1999, Kerr et al. 2000, Manduchi et al. 2000, Tusher et al. 2001, Westfall et al. 2001). However, the proposed solutions have not always been cast in the standard statistical framework.

This article discusses different approaches to multiple hypothesis testing in the context of microarray experiments and compares the procedures on microarray and simulated datasets. Section 2 reviews basic notions and approaches to multiple testing, and discusses the recent proposals of Efron et al. (2000), Golub et al. (1999), and Tusher et al. (2001) within this framework. The microarray datasets and simulation models which are used to evaluate the different multiple testing procedures are described in Section 3 and the results of the comparison study are presented in Section 4. Finally, Section 5 summarizes our findings and outlines open questions. Although the focus is on the identification of differentially expressed genes in microarray experiments, some of the methods described in this article are applicable to any large–scale multiple testing problem.

2 Methods

2.1 Multiple testing in microarray experiments

Consider a microarray experiment which produces expression data on m genes (or variables) for n mRNA samples, and further suppose that a response or covariate of interest is recorded for each sample. Such data may arise, for example, from a study of gene expression in tumor biopsy specimens from leukemia patients (Golub et al. 1999): in this case, the response is the tumor type and the goal is to identify genes that are differentially expressed in the different types of tumors. The data for sample i consist of a response or covariate y_i and a gene expression profile $\mathbf{x}_i = (x_{1i}, \ldots, x_{mi})$, where x_{ji} denotes the expression level of gene j in sample i, $i = 1, \ldots, n$, $j = 1, \ldots, m$. The expression levels x_{ji} might be either absolute (*e.g.* Affymetrix oligonucleotide chips (Lockhart et al. 1996)) or relative with respect to the expression levels of a suitably defined common reference sample (*e.g.* Stanford two–color cDNA microarrays (Brown & Botstein 1999, DeRisi et al. 1997)). Note that the expression levels x_{ji} are in general highly processed data. The raw data in a microarray experiment consist of image files, and important preprocessing steps include image analysis of these scanned images and normalization (Yang, Buckley, Dudoit & Speed 2002, Yang et al. 2001, Yang, Dudoit, Luu, Lin, Peng, Ngai & Speed 2002). The gene expression data are conventionally stored in an $m \times n$ matrix $X = (x_{ji})$, with rows corresponding to genes and columns to individual mRNA samples [1]. In a typical experiment, the total number n of samples is anywhere between around ten and a few hundreds, and the number m of genes is several thousands. The gene expression levels, x, are continuous variables, while the response or covariate, y, could be either polytomous or continuous as described above. Let X_j denote the random variable corresponding to the expression level for gene j and let Y denote the response or covariate.

The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene j of the null hypothesis H_j of no association between X_j and Y. A standard approach to this problem consists of two aspects: (1) computing a test statistic T_j for each gene j, and (2) applying a multiple testing procedure to determine which hypotheses to reject while controlling a suitably defined Type I error rate (Dudoit et al. 2002, Efron et al. 2000, Golub et al. 1999, Kerr et al. 2000, Manduchi et al. 2000, Tusher et al. 2001, Westfall et al. 2001).

The univariate problem in (1) has been studied extensively in the statistical literature. In general, the appropriate test statistic will depend on the experimental design and the type of response or covariate. For example, for binary covariates one might consider a t– or a Mann–Whitney statistic, for categorical responses one might use an F–statistic, and for survival data one might rely on the score statistic for the Cox proportional hazard model. We won't discuss the choice of statistic any further here, except to say that for each gene j the null hypothesis H_j is tested based on a statistic T_j, where t_j denotes a realization of the random variable T_j. To simplify matters, and unless specified otherwise, we further assume that the null H_j is rejected for large values of $|T_j|$ (two–sided hypotheses). Question (2) is the subject of the present article. Although multiple testing is by no means a new subject in the statistical literature, microarray experiments present a new and challenging area of application for multiple testing procedures because of the sheer number of comparisons. In the remainder of this section, we review basic notions and approaches to

[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.

multiple testing, and discuss recent proposals for dealing with the multiplicity problem in microarray experiments.

2.2 Type I error rates

Set–up. Consider the problem of testing simultaneously m null hypotheses H_j, $j = 1, \ldots, m$, and denote by R the number of rejected hypotheses. In the frequentist setting, the situation can be summarized by the table below (Benjamini & Hochberg 1995). The specific m hypotheses are assumed to be known in advance, the numbers m_0 and $m_1 = m - m_0$ of true and false null hypotheses are unknown parameters, R is an observable random variable, and S, T, U, and V are unobservable random variables. In the microarray context, there is a null hypothesis H_j for each gene j and rejection of H_j corresponds to declaring that gene j is differentially expressed. In general, one 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 in a univariate setting is to prespecify an acceptable Type I error rate α and seek tests which minimize the Type II error rate, *i.e.*, maximize *power*, within the class of tests with Type I error rate α.

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

Type I error rates. When testing a single hypothesis, H_1, say, the probability of a Type I error, *i.e.*, of rejecting the null hypothesis when it is true, is usually controlled at some designated level α. This can be achieved by choosing a critical value c_α such that $pr(|T_1| \geq c_\alpha | \mathrm{H}_1) \leq \alpha$ and rejecting H_1 when $|T_1| \geq c_\alpha$. A variety of generalizations to the multiple testing situation are possible; the Type I error rates described next are the most standard (Shaffer 1995).

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

$$PCER = E(V)/m.$$

- *Per–family error rate (PFER).* The PFER is defined as the expected number of Type I errors, *i.e.*,

$$PFER = E(V).$$

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

$$FWER = pr(V \geq 1).$$

- *False discovery rate (FDR).* The FDR of Benjamini & Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses, *i.e.*,

$$FDR = E(Q),$$

where by definition

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

Strong *vs.* **weak control.** It is important to note that the expectations and probabilities above are *conditional* on which hypotheses are true. A fundamental, yet often ignored distinction, is that between strong and weak control of the Type I error rate. *Strong control* refers to control of the Type I error rate under any combination of true and false hypotheses, *i.e.*, any value of m_0. In contrast, *weak control* refers to control of the Type I error rate only when all the null hypotheses are true, *i.e.*, under the *complete null hypothesis* $\mathrm{H}_0^C = \cap_{j=1}^m \mathrm{H}_j$ with $m_0 = m$. In other words, for the FWER, weak control means control of $pr(V \geq 1 \mid \mathrm{H}_0^C)$, while strong control means control of $\max_{\Lambda_0 \subseteq \{1,\ldots,m\}} pr(V \geq 1 \mid \cap_{j \in \Lambda_0} \mathrm{H}_j)$. In general, weak control without any other safeguards is unsatisfactory. In the microarray setting, where it is very unlikely that no genes are differentially expressed, it seems particularly important to have strong control of the Type I error rate. In the remainder of this article, unless specified otherwise, probabilities and expectations are computed under arbitrary combinations of true and false hypotheses, that is, under the null hypotheses $\cap_{j \in \Lambda_0} \mathrm{H}_j$ for some arbitrary subset $\Lambda_0 \subseteq \{1, \ldots, m\}$ of size m_0.

Power. Within the class of multiple testing procedures that control a given Type I error rate at an acceptable level α, one seeks procedures that maximize *power*, that is, minimize a suitably defined Type II error rate. As with Type I error rates, the concept of power can be generalized in various ways when moving from single to multiple hypothesis testing. Three common definitions of power are: (i) the probability of rejecting at least one false null hypothesis, $pr(S \geq 1) = pr(T \leq m_1 - 1)$; (ii) the average probability of rejecting the false null hypotheses, $E(S)/m_1$, or *average power*; and (iii) the probability of rejecting all false null hypotheses, $pr(S = m_1) = pr(T = 0)$ (Shaffer 1995). When the family of tests consists of pairwise mean comparisons, these quantities have been called any–pair power, per–pair power, and all–pairs power (Ramsey 1978). In a spirit analogous to the FDR, one could also define power as $E(S/R|R > 0)pr(R > 0) = pr(R > 0) - FDR$; when $m = m_1$, this is the any–pair power $pr(S \geq 1)$. One should note again that probabilities are conditional on which null hypotheses are true and which are false.

Comparison of Type I error rates. In general, for a given multiple testing procedure, $PCER \leq FWER \leq PFER$. Thus, for a fixed criterion α for controlling the Type I error rates, the order reverses for the number of rejections R: procedures controlling the PFER are generally more conservative than those controlling either the FWER or the PCER, and procedures controlling the FWER are more conservative than those controlling the PCER. To illustrate the properties of the different Type I error rates, suppose each hypothesis H_j is tested individually at level α_j and the decision to reject or not reject this hypothesis is based solely on that test. Under the complete null hypothesis, the PCER is simply the average of the α_j and the PFER is the sum of the α_j. In contrast, the FWER is a function not of the test sizes α_j alone, but involves the *joint* distribution of the test statistics T_j

$$PCER = (\alpha_1 + \ldots + \alpha_m)/m \leq \max(\alpha_1, \ldots, \alpha_m) \ \leq \ FWER \ \leq \ PFER = \alpha_1 + \ldots + \alpha_m.$$

The FDR also depends on the joint distribution of the test statistics and, for a fixed procedure, $FDR \leq FWER$, with $FDR = FWER$ under the complete null. The classical approach to multiple testing calls for strong control of the FWER (*e.g.* Bonferroni procedure). The recent proposal of Benjamini & Hochberg (1995) controls the FWER in the weak sense and can be less conservative than FWER otherwise. Procedures controlling the PCER are generally less conservative than those controlling either the FDR or FWER, but tend to ignore the multiplicity problem altogether. The following simple example describes the behavior of the various Type I error rates as the total number of hypotheses m and the proportion of true hypotheses m_0/m vary.

A simple example. Consider random Gaussian m–vectors with mean $\mu = (\mu_1, \ldots, \mu_m)$ and identity covariance matrix I_m. Suppose we wish to test simultaneously the m null hypotheses $\mathrm{H}_j : \mu_j = 0$ against the two–sided alternatives $\mathrm{H}'_j : \mu_j \neq 0$. Given a random sample of n m–vectors from this distribution, a simple multiple testing procedure would be to reject H_j if $|\bar{X}_j| \geq z_{\alpha/2}/\sqrt{n}$, where $\bar{X}_j$ is the average of the jth coordinate for the n m–vectors, $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1-\alpha/2$, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Let $R_j = I(|\bar{X}_j| \geq z_{\alpha/2}/\sqrt{n})$, where $I(\cdot)$ is the indicator function, equaling 1 if the condition in parentheses is true, and 0 otherwise, and assume without loss of generality that the m_0 true null hypotheses are $\mathrm{H}_1, \ldots, \mathrm{H}_{m_0}$. Then $V = \sum_{j=1}^{m_0} R_j$, $R = \sum_{j=1}^{m} R_j$, and analytical formulae for the Type I error rates can easily be derived as

$$\begin{aligned}
PFER &= \sum_{j=1}^{m_0} \gamma_j, \\
PCER &= \sum_{j=1}^{m_0} \gamma_j/m, \\
FWER &= 1 - \prod_{j=1}^{m_0} (1-\gamma_j), \\
FDR &= \sum_{r_1=0}^{1} \cdots \sum_{r_m=0}^{1} \frac{\sum_{j=1}^{m_0} r_j}{\sum_{j=1}^{m} r_j} \prod_{j=1}^{m} \gamma_j^{r_j} (1-\gamma_j)^{1-r_j},
\end{aligned}$$

with the FDR convention that $0/0 = 0$ and $\gamma_j = E(R_j) = 1 - \Phi(z_{\alpha/2} - \mu_j\sqrt{n}) + \Phi(-z_{\alpha/2} - \mu_j\sqrt{n})$ denoting the chance of rejecting hypothesis H_j. In our simple example, $\gamma_j = \alpha$ for $j = 1, \ldots, m_0$, and if we further assume that $\mu_j = d/\sqrt{n}$ for $j = m_0 + 1, \ldots, m$, then the expressions for the error rates simplify to

$$\begin{aligned}
PFER &= m_0\alpha, \\
PCER &= m_0\alpha/m, \\
FWER &= 1 - (1-\alpha)^{m_0}, \\
FDR &= \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \binom{m_0}{v} \alpha^v (1-\alpha)^{m_0-v} \binom{m_1}{s} \beta^s (1-\beta)^{m_1-s},
\end{aligned}$$

where $\beta = 1 - \Phi(z_{\alpha/2} - d) + \Phi(-z_{\alpha/2} - d)$. Note that unlike the PFER, FWER, or PCER, the FDR depends on the distribution of the test statistics for the false null hypotheses through the random variable S. In general, the FDR is thus more complicated to compute. Figure 1 displays plots of the FWER, PCER, and FDR *vs.* the number of hypotheses m, for different proportions

$m_0/m = 1\,, 0.9, 0.8, 0.5, 0.2, 0.1$ of true null hypotheses, and for $\alpha = 0.05$ and $d = 1$. In general, the FWER and PFER increase sharply with the number of hypotheses m, while the PCER remains constant (the PFER is not shown on the figure because it is on a different scale). Under the complete null $m = m_0$, the FDR is equal to the FWER and both increase sharply with m. However, as the proportion of true null hypotheses m_0/m decreases, the FDR remains relatively stable as a function of m and approaches the PCER. Figure 2 displays plots of the FWER, PCER, and FDR *vs.* individual test size α, for different proportions m_0/m of true null hypotheses, and for $m = 100$ and $d = 1$. The FWER is generally much larger than the PCER, the largest difference being under the complete null $m = m_0$. As the proportion of true null hypotheses decreases, the FDR becomes closer to the PCER. Similar behavior of the error rates is displayed in Figure 3, which plots Type I error rates *vs.* expected proportion of rejected hypotheses $E(R)/m$, for different proportions m_0/m of true null hypotheses, and for $m = 100$ and $d = 1$. These plots can be used to compare the different Type I error rates one would expect for a given number of rejected hypotheses.

2.3 Adjusted p–values

Unadjusted p–values. Consider first a single hypothesis H_1, say, and a family of tests of H_1, with level–α nested rejection regions S_α such that: (a) $pr(T_1 \in S_\alpha|\mathrm{H}_1) = \alpha$ for all $\alpha \in [0, 1]$ which are achievable under the distribution of T_1, and (b) $S_{\alpha'} = \cap_{\alpha \geq \alpha'} S_\alpha$ for all α and α' for which these regions are defined in (a). Rather than simply reporting rejection or not of the hypothesis, a *p–value* connected with the test can be defined as $p_1 = \inf\{\alpha : t_1 \in S_\alpha\}$ (adapted from Lehmann (1986), p. 170, to include discrete test statistics). The p–value can be thought of as the level of the test at which the hypothesis H_1 would just be rejected. The smaller the p–value p_1, the stronger the evidence against the null hypothesis H_1. Rejecting H_1 when $p_1 \leq \alpha$ provides control of the Type I error rate at level α. In our context, the p–value can also be restated as the probability of observing a test statistic as extreme or more extreme in the direction of rejection as the observed one, that is, $p_1 = pr(|T_1| \geq |t_1||\mathrm{H}_1)$. Extending this concept to the multiple testing situation leads to the very useful notion of adjusted p–value.

Adjusted p–values. Let t_j and $p_j = pr(|T_j| \geq |t_j||\mathrm{H}_j)$ denote respectively the test statistic and p–value for hypothesis H_j (gene j), $j = 1, \ldots, m$. Just as in the single hypothesis case, a multiple testing procedure may be defined in terms of critical values for the test statistics or p–values of individual hypotheses: *e.g.* reject H_j if $|t_j| \geq c_j$ or if $p_j \leq \alpha_j$, where the critical values c_j and α_j are chosen to control a given Type I error rate (FWER, PCER, PFER, or FDR) at a prespecified level α. Alternatively, 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_j can be defined as the level of the entire test procedure at which H_j would just be rejected, given the values of all test statistics involved (Hommel & Bernhard 1999, Shaffer 1995, Westfall & Young 1993, Wright 1992, Yekutieli & Benjamini 1999). If interest is in controlling the FWER, the FWER adjusted p–value for hypothesis H_j is

$$\tilde{p}_j = \inf\{\alpha \in [0, 1] : \mathrm{H}_j \text{ is rejected at } FWER = \alpha\}.$$

The corresponding random variables for unadjusted (or raw) and adjusted p–values are denoted by P_j and $\tilde{P}_j$, respectively. Hypothesis H_j is then rejected, *i.e.*, gene j is declared differentially expressed, at FWER α if $\tilde{p}_j \leq \alpha$. Adjusted p–values for other Type I error rates are defined similarly, that is, for the FDR, $\tilde{p}_j = \inf\{\alpha : \mathrm{H}_j \text{ is rejected at } FDR = \alpha\}$ (Yekutieli & Benjamini 1999). 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 also most conveniently described in terms of their adjusted p–values and these can in turn be easily determined using resampling methods (Westfall & Young 1993).

Stepwise procedures. One usually distinguishes among three types of multiple testing procedures: 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 unadjusted p–values, that is, each hypothesis is evaluated using a critical value that is independent of the results of tests of other hypotheses. Improvement 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. In *step–down* procedures, the hypotheses corresponding to the *most* significant test statistics (*i.e.*, smallest unadjusted p–values or largest absolute test statistics) are considered successively, with further tests depending on the outcomes of earlier ones. As soon as one hypothesis is accepted, all remaining hypotheses are accepted. In contrast, for *step–up* procedures, the hypotheses corresponding to the *least* significant test statistics are considered successively, again with further tests depending on the outcomes of earlier ones. As soon as one hypothesis is rejected, all remaining hypotheses are rejected. The next section discusses single–step and stepwise procedures for control of the FWER.

2.4 Control of the family–wise error rate

2.4.1 Single–step procedures

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

$$\tilde{p}_j = \min\bigl(mp_j, 1\bigr). \tag{1}$$

Control of the FWER in the strong sense follows from Boole's inequality. Assume without loss of generality that the true null hypotheses are H_j, for $j = 1, \ldots, m_0$, then, for P_j having a $U[0,1]$ distribution under H_j

$$FWER = pr(V \geq 1) = pr\Bigl(\bigcup_{j=1}^{m_0}\{\tilde{P}_j \leq \alpha\}\Bigr) \leq \sum_{j=1}^{m_0} pr\bigl(\tilde{P}_j \leq \alpha\bigr) \leq \sum_{j=1}^{m_0} pr\bigl(P_j \leq \alpha/m\bigr) = m_0\alpha/m.$$

Closely related to the Bonferroni procedure is the Šidák procedure which is exact under the complete null for protecting the FWER when the unadjusted p–values are independently distributed as $U[0,1]$. The *single–step Šidák adjusted p–values* are given by

$$\tilde{p}_j = 1 - (1 - p_j)^m. \tag{2}$$

However, 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, for example, to co–regulation. Westfall & Young (1993) propose adjusted p–values for

less conservative multiple testing procedures which take into account the dependence structure among test statistics. The *single–step minP adjusted p–values* are defined by

$$\tilde{p}_j = pr\big(\min_{1\le l\le m} P_l \le p_j \mid \mathrm{H}_0^C\big), \tag{3}$$

where H_0^C denotes the complete null hypothesis and P_l the random variable for the unadjusted p–value of the lth hypothesis. Alternatively, one may consider procedures based on the *single–step maxT adjusted p–values* which are defined in terms of the test statistics T_j themselves

$$\tilde{p}_j = pr\big(\max_{1\le l\le m} |T_l| \ge |t_j| | \mathrm{H}_0^C\big). \tag{4}$$

The following points should be noted regarding these four procedures.

1. If the unadjusted p–values $(P_1, \ldots, P_m)$ are independent and P_j has a $U[0,1]$ distribution under H_j, 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, however, it controls the FWER for test statistics that satisfy an inequality known as Šidák's inequality: $pr(|T_1| \le c_1, \ldots, |T_m| \le c_m) \ge \prod_{j=1}^m pr(|T_j| \le c_j)$. This inequality, also known as the *positive orthant dependence property*, 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 or equal to the Šidák adjusted p–values.

3. Computing the quantities in (3) using the upper bound provided by Boole's inequality yields the Bonferroni p–values, for unadjusted p–values $P_l \sim U[0,1]$ marginally under H_l.

In other words, procedures based on the minP adjusted p–values are less conservative than the Bonferroni or Šidák (under the Šidák inequality) procedures. In the case of independent test statistics, the Šidák and minP adjustments are equivalent as discussed in item 1, above.

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 unadjusted p–values $(P_1, \ldots, P_m)$ is said to have the *subset pivotality* property if the joint distribution of the sub–vector $\{P_j : j \in \Lambda_0\}$ is identical under the restrictions $\cap_{j\in\Lambda_0}\mathrm{H}_j$ and $\mathrm{H}_0^C = \cap_{j=1}^m \mathrm{H}_j$, for all subsets Λ_0 of $\{1, \ldots, m\}$. The subset pivotality condition is important because it ensures that procedures based on adjusted p–values computed under the complete null provide strong control of the FWER. A practical consequence of this property is that resampling for computing adjusted p–values may be done conveniently under the complete null rather than the partial null hypotheses. Without subset pivotality, multiplicity adjustment is more complex.

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_j 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. In the case of non–identically distributed test statistics T_j (*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 2.6) 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 computations than the maxT p–values, because the unadjusted p–values must be computed before considering the distribution of their successive minima (Ge & Dudoit 2002).

2.4.2 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 analogs, in terms of their adjusted p–values, of the four procedures described in the previous section. Let $p_{r_1} \leq p_{r_2} \leq ... \leq p_{r_m}$ denote the *observed ordered unadjusted p–values*, and $\mathrm{H}_{r_1}, \mathrm{H}_{r_2}, \ldots, \mathrm{H}_{r_m}$ the corresponding null hypotheses. For control of the FWER at level α, the Holm (1979) procedure proceeds as follows. Define

$$j^* = \min\Big\{ j : p_{r_j} > \frac{\alpha}{m-j+1} \Big\}$$

and reject hypotheses H_{r_j}, for $j = 1, \ldots, j^* - 1$. If no such j^* exists, reject all hypotheses. The *step–down Holm adjusted p–values* are thus given by

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \Big\{ \min\big((m-k+1)\, p_{r_k}, 1\big) \Big\}. \tag{5}$$

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

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \Big\{ 1 - (1 - p_{r_k})^{(m-k+1)} \Big\}. \tag{6}$$

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

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \Big\{ pr\big(\min_{l \in \{r_k,\ldots,r_m\}} P_l \leq p_{r_k} \mid \mathrm{H}_0^C \big) \Big\}, \tag{7}$$

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

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \Big\{ pr\big(\max_{l \in \{r_k,\ldots,r_m\}} |T_l| \geq |t_{r_k}| \mid \mathrm{H}_0^C \big) \Big\}, \tag{8}$$

where $|t_{r_1}| \geq |t_{r_2}| \geq ... \geq |t_{r_m}|$ denote the *observed ordered test statistics*. Note that computing the quantities in (7) under the assumption that $P_l \sim U[0,1]$ and using the upper bound provided by Boole's inequality yields 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 the strong control of the FWER for the maxT and minP procedures the reader is referred to Westfall & Young (1993, Section 2.8). Step–down procedures such as the Holm procedure may be further improved by taking into account logically related hypotheses as described in Shaffer (1986).

2.4.3 Step–up procedures

In contrast to step–down procedures, step–up procedures begin with the least significant p–value, p_{r_m}, and are usually based on the following probability result of Simes (1986). Under the complete null hypothesis H_0^C and for independent test statistics, the ordered unadjusted p–values $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ satisfy

$$pr\big(P_{(j)} > \alpha j/m,\ \forall\ j = 1, \ldots, m \mid \mathrm{H}_0^C\big) \geq 1 - \alpha,$$

with equality in the continuous case. This inequality is known as the *Simes inequality*. In important cases of dependent test statistics, Simes showed that the probability was larger than $1-\alpha$, however this does not hold generally for all joint distributions.

The Hochberg (1988) procedure, based on the Simes inequality, can be viewed as a step–up modification of Holm's step–down procedure, since the ordered p–values are compared to the same critical values in both procedures. For control of the FWER at level α, define

$$j^* = \max\Big\{j : p_{r_j} \leq \frac{\alpha}{m - j + 1}\Big\}$$

and reject hypotheses H_{r_j}, for $j = 1, \ldots, j^*$. If no such j^* exists, reject no hypothesis. The *step–up Hochberg adjusted p–values* are thus given by

$$\tilde{p}_{r_j} = \min_{k=j,\ldots,m} \Big\{\min\big((m - k + 1)\, p_{r_k}, 1\big)\Big\}. \tag{9}$$

Related procedures include those of Hommel (1988) and Rom (1990). Step–up procedures have often been found to be more powerful than their step–down counterparts; however, it is important to keep in mind that all procedures based on the Simes inequality rely on the assumption that the result proved under independence yields a conservative procedure for dependent tests. More research is needed to determine circumstances in which such methods are applicable, and in particular, whether they are applicable for the types of correlation structures encountered in microarray experiments. Troendle (1996) proposed a permutation–based step–up multiple testing procedure which takes into account the dependence structure among the test statistics and is related to the Westfall & Young (1993) step–down maxT procedure.

2.5 Control of the false discovery rate

A different approach to multiple testing was proposed in 1995 by Benjamini & Hochberg. These authors argue that, in many situations, control of the FWER can lead to unduly conservative procedures and one may be prepared to tolerate some Type I errors, provided their number is small in comparison to the number of rejected hypotheses. These considerations led to a less conservative approach which calls for controlling the expected proportion of Type I errors among the rejected hypotheses – the *false discovery rate*, FDR. More specifically, the FDR is defined as $FDR = E(Q)$, where $Q = V/R$ if $R > 0$, and 0 if $R = 0$, *i.e.*, $FDR = E(V/R \mid R > 0)pr(R > 0)$. Under the complete null, given the definition of $0/0 = 0$ when $R = 0$, the FDR is equal to the FWER; procedures controlling the FDR thus also control the FWER in the weak sense. Note that earlier references to the FDR can be found in Seeger (1968) and Sorić (1989).

Benjamini & Hochberg (1995) derived the following step–up procedure for (strong) control of the FDR for independent test statistics. Let $p_{r_1} \leq p_{r_2} \leq ... \leq p_{r_m}$ denote the observed ordered unadjusted p–values. For control of the FDR at level α define

$$j^* = \max\Big\{j : p_{r_j} \leq \frac{j}{m}\alpha\Big\}$$

and reject hypotheses H_{r_j}, for $j = 1, \ldots, j^*$. If no such j^* exists, reject no hypothesis. Corresponding adjusted p–values are

$$\tilde{p}_{r_j} = \min_{k=j,\ldots,m} \Big\{\min\Big(\frac{m}{k}\, p_{r_k}, 1\Big)\Big\}. \tag{10}$$

Benjamini & Yekutieli (2001) proved that this procedure controls the FDR under certain dependence structures (positive regression dependency). They also proposed a simple conservative modification of the procedure which controls the false discovery rate for arbitrary dependence structures. For control of the FDR at level α, define

$$j^* = \max\Big\{j : p_{r_j} \leq \frac{j}{m\sum_{j=1}^{m} 1/j}\alpha\Big\}$$

and reject hypotheses H_{r_j}, for $j = 1, \ldots, j^*$. If no such j^* exists, reject no hypothesis. Corresponding adjusted p–values are

$$\tilde{p}_{r_j} = \min_{k=j,\ldots,m} \Big\{\min\Big(\frac{m\sum_{j=1}^{m} 1/j}{k}\, p_{r_k}, 1\Big)\Big\}. \tag{11}$$

For a large number m of hypotheses, the penalty in this conservative procedure is about $\log m$, as compared to the Benjamini & Hochberg (1995) procedure. Note that the Benjamini & Hochberg procedure can also be conservative, even in the independence case, as it was shown that for this step–up procedure $E(Q) \leq \frac{m_0}{m}\alpha \leq \alpha$. Until recently, most FDR controlling procedures were either designed for independent test statistics or did not make use of the dependency structure among the test statistics. In the spirit of the Westfall & Young (1993) resampling procedures for FWER control, Yekutieli & Benjamini (1999) proposed new FDR controlling procedures which use resampling based adjusted p–values to incorporate certain types of dependency structures among the test statistics (the procedures assume among other things that the unadjusted p–values for the true null hypotheses are independent of the p–values for the false null hypotheses).

In the microarray setting, where thousands of comparisons are performed simultaneously and a fairly large number of genes are expected to be differentially expressed, FDR controlling procedures present a promising alternative to more conservative FWER approaches. In this context, one may be willing to bear a few false positives as long as their number is small in comparison to the number of rejected hypotheses. The problematic definition of $0/0 = 0$ is also not as important in this case.

2.6 Resampling

In many situations, the joint (and marginal) distribution of the test statistics is unknown. Resampling methods (bootstrap or permutation) can be used to estimate unadjusted and adjusted p–values while avoiding parametric assumptions about the joint distribution of the test statistics. In the microarray setting, the joint distribution under the complete null hypothesis of the test statistics $(T_1, \ldots, T_m)$ can be estimated by permuting the columns of the gene expression data matrix X. Permuting entire columns of this matrix creates a situation in which the response or

covariate Y is independent of the gene expression levels, while attempting to preserve 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, and in such a case a random subset of B permutations (including the observed) may be considered. The manner in which the responses/covariates are permuted depends on the experimental design, for example, for a two–factor design, one should permute the levels of the factor of interest within the levels of the other factor (see Scheffé (1959), Section 9.3, and Section 3.2.2 for an example).

Box 1. Permutation algorithm for unadjusted p–values.

For the bth permutation, $b = 1, \ldots, B$

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

The permutation distribution of the test statistic T_j for hypothesis H_j, $j = 1, \ldots, m$, is given by the empirical distribution of $t_{j,1}, \ldots, t_{j,B}$. For two–sided alternative hypotheses, the permutation p–value for hypothesis H_j is

$$p_j^* = \frac{\sum_{b=1}^{B} I\big(|t_{j,b}| \geq |t_j|\big)}{B},$$

where $I(\cdot)$ is the indicator function, equaling 1 if the condition in parentheses is true, and 0 otherwise.

Permutation adjusted p–values for the Bonferroni, Šidák, Holm, and Hochberg procedures can be obtained by replacing p_j by p_j^* in equations (1), (2), (5), (6), and (9). The permutation unadjusted p–values can also be used for the FDR controlling procedures described in Section 2.5. For the step–down maxT adjusted p–values of Westfall & Young (1993), the null distribution of successive maxima $\max_{l \in \{r_j, \ldots, r_m\}} |T_l|$ 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 $\max_{l \in \{r_1, \ldots, r_m\}} |T_l|$).

Box 2. Permutation algorithm for step–down maxT adjusted p–values – based on Algorithms 2.8 and 4.1 in Westfall & Young (1993).

For the bth permutation, $b = 1, \ldots, B$

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

$$\begin{aligned} u_{m,b} &= |t_{r_m,b}| \\ u_{j,b} &= \max\Big(u_{j+1,b}, |t_{r_j,b}|\Big) \qquad \text{for } j = m-1, \ldots, 1, \end{aligned}$$

where r_j are such that $|t_{r_1}| \geq |t_{r_2}| \geq ... \geq |t_{r_m}|$ for the **original** data.

The adjusted p–values are estimated by

$$\tilde{p}^*_{r_j} = \frac{\sum_{b=1}^{B} I\big(u_{j,b} \geq |t_{r_j}|\big)}{B},$$

with the monotonicity constraints enforced by setting

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

The reader is referred to Ge & Dudoit (2002) for a fast permutation algorithm for estimating minP adjusted p–values.

2.7 Recent proposals for microarray experiments

Efron et al. (2000), Golub et al. (1999), and Tusher et al. (2001) have recently proposed resampling algorithms for multiple testing in microarray experiments. However, these procedures were not presented within the standard statistical framework for multiple testing. In particular, the Type I error rates considered were rather loosely defined, thus making it difficult to assess the properties of the multiple testing procedures. These recent proposals are reviewed next, within the framework introduced in Sections 2.2 and 2.3.

2.7.1 Neighborhood analysis of Golub et al.

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) (the study is described in greater detail in Section 3.2.3). In their so–called *neighborhood analysis*, the authors compute a test statistic t_j for each gene ($P(g, c)$ in their paper)

$$t_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{s_{1j} + s_{2j}},$$

where $\bar{x}_{kj}$ and s_{kj} denote respectively the average and standard deviation of the expression levels of gene j in the class $k = 1, 2$ samples. This statistic is based on an *ad hoc* definition of correlation, and resembles a t–statistic with an unusual standard error calculation (note 16 in Golub et al.). It

is not pivotal, *i.e.*, its null distribution depends on parameters of the distribution which generated the data, and a standard two–sample t–statistic should be preferred (note that this definition of pivotality is different from subset pivotality in Section 2.4). Statistics such as t_j have been used in meta–analysis to measure effect sizes (National Reading Panel 1999).

Golub et al. use the term "neighborhood" to refer to sets of genes with test statistics T_j greater in absolute value than a given critical value c, that is, sets of rejected hypotheses $\{j : T_j \geq c\}$ or $\{j : T_j \leq -c\}$ (these sets are denoted by $N_1(c, r)$ and $N_2(c, r)$ in note 16 of Golub et al.). The ALL/AML labels were permuted $B = 400$ times to estimate the complete null distribution of the numbers $R(c) = V(c) = \sum_{j=1}^{m} I(T_j \geq c)$ of false positives for different critical values c (similarly for the other tail, with $T_j \leq -c$). Figure 2 in Golub et al. contains plots of the observed $R(c) = r(c)$ and permutation quantiles of $R(c)$ against critical values c for one–sided tests [2]. A critical value c is then chosen so that the chance of exceeding the observed $r(c)$ under the complete null is equal to a prespecified level α, that is,

$$G(c) = pr\big(R(c) \geq r(c) \mid \mathrm{H}_0^C\big) = \alpha. \tag{12}$$

Golub et al. provide no further guidelines for selecting the critical value c or discussion of the Type I error control of their procedure. Like some PFER, PCER, or FWER controlling procedures, the neighborhood analysis considers the complete null distribution of the number of Type I errors $V(c) = R(c)$, however, instead of controlling $E(V(c))$, $E(V(c))/m$, or $pr(V(c) \geq 1)$, it seeks to control a different quantity, $pr(R(c) \geq r(c)|\mathrm{H}_0^C)$, which can be thought of as a p–value under H_0^C for the number of rejected hypotheses $R(c)$ and is thus a random variable. For simplicity, consider continuous test statistics and two–sided hypotheses. Then, conditional on the observed ordered absolute test statistics, $|t|_{(1)} \geq \ldots \geq |t|_{(m)}$, the function $G(c)$ is left–continuous, with discontinuities at $|t|_{(j)}$, $j = 1, \ldots, m$, that is,

$$G(c) = \begin{cases} pr\big(R(c) \geq m \mid \mathrm{H}_0^C\big) \\ pr\big(R(c) \geq j \mid \mathrm{H}_0^C\big) \\ pr\big(R(c) \geq 0 \mid \mathrm{H}_0^C\big) \end{cases} = \begin{cases} pr\big(|T|_{(m)} \geq c \mid \mathrm{H}_0^C\big), & \text{if } c \leq |t|_{(m)}, \\ pr\big(|T|_{(j)} \geq c \mid \mathrm{H}_0^C\big), & \text{if } |t|_{(j+1)} < c \leq |t|_{(j)}, \quad 1 \leq j \leq m-1, \\ 1, & \text{if } |t|_{(1)} < c. \end{cases}$$

Here, $|T|_{(j)}$ denote the ordered absolute test statistics, $|T|_{(1)} \geq \ldots \geq |T|_{(m)}$, and, in principle, different realizations of $|T|_{(j)}$ could correspond to different hypotheses. Although $G(c)$ is decreasing in c within intervals $(|t|_{(j+1)}, |t|_{(j)}]$, it is not in general decreasing overall, and there may be several values of c with $G(c) = \alpha$. Hence, one must decide on an appropriate procedure for selecting the critical value c. Two natural choices are given by stepwise procedures, as described next.

Step–down procedure. A step–down procedure would be to let $c = |t|_{(j^*-1)}$, where

$$j^* = \min\Big\{j : G(|t|_{(j)}) > \alpha\Big\} = \min\Big\{j : pr\big(|T|_{(j)} \geq |t|_{(j)} \mid \mathrm{H}_0^C\big) > \alpha\Big\},$$

and reject hypotheses $\mathrm{H}_{(j)}$, for $j = 1, \ldots, j^* - 1$. If no such j^* exists, reject all hypotheses. The corresponding adjusted p–values are

$$\tilde{p}_{(j)} = \max_{k=1,\ldots,j} \Big\{pr\big(|T|_{(k)} \geq |t|_{(k)} \mid \mathrm{H}_0^C\big)\Big\}. \tag{13}$$

[2] We are aware that our notation can lead to confusion when compared with that of Golub et al. We chose to follow the notation of Sections 2.2 and 2.3 to allow easy comparison with other multiple testing procedures described in the present article. For comparison with Golub et al. note that we use T_j to denote $P(g, c)$, c to denote r, and $r(c)$ to denote the realization of $R(c)$, that is, $|N_1(c, r)| + |N_2(c, r)|$.

Hence, the step–down adjusted p–values for the neighborhood analysis of Golub et al. are based on the p–values $pr\big(|T|_{(j)} \geq |t|_{(j)} \mid \mathrm{H}_0^C\big)$ for the order statistics. Note that, unlike the Westfall & Young maxT procedure, where the adjusted p–values are based on $pr\big(\max_{l\in\{r_j,\ldots,r_m\}} |T_l| \geq |t_{r_j}| \mid \mathrm{H}_0^C\big)$ for the *fixed* ordering $\{r_j\}$ of the observed test statistics, the maxima in equation (13) could be taken over different sets of hypotheses for different realizations of T_j. Because of this difference, it is unlikely that the step–down version of neighborhood analysis provides strong control of any Type I error rate, as the subset pivotality condition is not satisfied. A straightforward argument shows that the step–down procedure does, however, control the FWER weakly.

Step–up procedure. The corresponding step–up procedure would be to let $c = |t|_{(j^*)}$, where

$$j^* = \max\Big\{j : G(|t|_{(j)}) \leq \alpha\Big\} = \max\Big\{j : pr\big(|T|_{(j)} \geq |t|_{(j)} \mid \mathrm{H}_0^C\big) \leq \alpha\Big\},$$

and reject hypotheses $\mathrm{H}_{(j)}$, for $j = 1, \ldots, j^*$. If no such j^* exists, reject no hypothesis. The corresponding adjusted p–values are

$$\tilde{p}_{(j)} = \min_{k=j,\ldots,m} \Big\{pr\big(|T|_{(k)} \geq |t|_{(k)} \mid \mathrm{H}_0^C\big)\Big\}. \tag{14}$$

Again, it is unlikely that this procedure provides strong control of any Type I error rate, as the subset pivotality condition is not met. Furthermore, because of the step–up nature of the procedure and the non–monotonicity of the function $G(c)$, the FWER is not even controlled weakly. In this step–up procedure, hypothesis $\mathrm{H}_{(j)}$ is rejected at nominal level α whenever $pr\big(|T|_{(k)} \geq |t|_{(k)} \mid \mathrm{H}_0^C\big) \leq \alpha$ for some $k \geq j$, thus each hypothesis is given several chances at rejection.

Figures 4 and 5 display plots of the observed number of rejected hypotheses $r(c)$ and 95th permutation quantile of $R(c)$ against the critical value c for data simulated from a model described in Table 3. Also shown are plots of the "Type I error rate" $G(c)$ *vs.* c. These figures illustrate the non–monotonicity of $G(c)$ and the possibility of having several values of c with $G(c) = \alpha$. The plots of the step–down and step–up adjusted p–values demonstrate the importance of selecting the proper c in the case of several "crossings" of the observed $r(c)$ with quantiles of $R(c)$. For a fixed criterion $\alpha = 0.05$, the step–up procedure starts from the left and looks for the first time $G(c)$ dips below $\alpha = 0.05$, while the step–down procedures starts from the right and looks for the first time $G(c)$ rises above $\alpha = 0.05$. When $G(c)$ is monotone in c, the two procedures should yield the same results, otherwise, the step–up procedure generally produces a larger number of rejected hypotheses, but does not provide weak control of the FWER.

As a final remark, note that the number of permutations $B = 400$ used in Golub et al. (1999) is probably not large enough for reporting 99th quantiles in Figure 2. A better plot for Figure 2 of Golub et al. might be of the error rate $G(c) = pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$ *vs.* the critical values c, as this does not require a prespecified level α.

2.7.2 Significance Analysis of Microarrays (SAM) of Efron et al. and Tusher et al.

We consider two variants of the *Significance Analysis of Microarrays* or *SAM* multiple testing procedure, the original version in Efron et al. (2000) and the more recent Tusher et al. (2001) and Chu et al. (2000) version. Note that these manuscripts also address the question of choosing appropriate test statistics for different types of responses and covariates. Here, we focus only on

the proposed methods for dealing with the multiplicity problem and assume that a suitable test statistic is computed for each gene. The two SAM procedures are described next.

1. Compute a test statistic t_j for each gene j and define order statistics $t_{(j)}$ such that $t_{(1)} \geq t_{(2)} \geq \ldots \geq t_{(m)}$ [3].

2. Perform B permutations of the responses/covariates $y_1, \ldots, y_n$. For each permutation b compute the test statistics $t_{j,b}$ and the corresponding order statistics $t_{(1),b} \geq t_{(2),b} \geq \ldots \geq t_{(m),b}$. Note that $t_{(j),b}$ may correspond to a different gene than $t_{(j)}$.

3. From the B permutations, estimate the expected value (under the complete null) of the order statistics by $\bar{t}_{(j)} = (1/B) \sum_b t_{(j),b}$.

4. Form a Quantile–Quantile plot (so–called "SAM plot") of the observed $t_{(j)}$ *vs.* the expected $\bar{t}_{(j)}$.

5. – *Efron et al.* For a fixed threshold Δ, genes with $|t_{(j)} - \bar{t}_{(j)}| \geq \Delta$ are declared "significant", *i.e.*, the corresponding hypotheses $\mathrm{H}_{(j)}$ are rejected.
   – *Tusher et al.* For a fixed threshold Δ, let $j_0 = \max\{j : \bar{t}_{(j)} \geq 0\}$, $j_1 = \max\{j \leq j_0 : t_{(j)} - \bar{t}_{(j)} \geq \Delta\}$, and $j_2 = \min\{j > j_0 : t_{(j)} - \bar{t}_{(j)} \leq -\Delta\}$ [4]. All genes with $j \leq j_1$ are called "significant positive" and all genes with $j \geq j_2$ are called "significant negative". Define the upper cut–point $cut_{up}(\Delta) = \min\{t_{(j)} : j \leq j_1\} = t_{(j_1)}$ and the lower cut–point $cut_{low}(\Delta) = \max\{t_{(j)} : j \geq j_2\} = t_{(j_2)}$. If no such j_1 (j_2) exists, set $cut_{up}(\Delta) = \infty$ ($cut_{low}(\Delta) = -\infty$).

6. – *Efron et al.* For a given threshold Δ, the expected number of false positives is estimated by applying step 5 to each of the B permuted datasets. That is, for each permutation b, compute the number of genes with $|t_{(j),b} - \bar{t}_{(j),b}| \geq \Delta$, where $\bar{t}_{(j),b} = \sum_{b' \neq b} t_{(j),b'}/(B-1)$ is the average of the order statistics excluding the bth permutation, and average this number over permutations.
   – *Tusher et al.* For a given threshold Δ, the expected number of false positives is estimated by computing for each of the B permutations the number of genes with $t_{j,b}$ above $cut_{up}(\Delta)$ or below $cut_{low}(\Delta)$, and averaging this number over permutations.

7. A threshold Δ is chosen to control the expected number of false positives, PFER, under the complete null, at an acceptable level.

Both SAM procedures return for each value of the threshold Δ the following quantities: the number of rejected hypotheses

$$R_{efron}(\Delta) = \sum_{j=1}^{m} I\big(|T_{(j)} - \bar{t}_{(j)}| \geq \Delta\big),$$

$$R_{tusher}(\Delta) = \sum_{j=1}^{m} \Big(I\big(T_j \geq cut_{up}(\Delta)\big) + I\big(T_j \leq cut_{low}(\Delta)\big) \Big) = j_1 + m - j_2 + 1,$$

[3] The notation for the ordered test statistics is different here than in Efron et al. (2000) and Tusher et al. (2001) to be consistent with previous notation whereby we set $t_{(1)} \geq t_{(2)} \geq \ldots \geq t_{(m)}$ and $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.

[4] This is our interpretation of the description in the SAM manual (Chu et al. 2000): "For a fixed threshold Δ, starting at the origin, and moving up to the right find the first $i = i_1$ such that $d_{(i)} - \bar{d}_{(i)} \geq \Delta$". That is, we take the "origin" to be given by the index j_0.

an estimate of the expected number of false positives, PFER, under the complete null

$$\widehat{PFER}^0_{efron}(\Delta) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{m} I\big(|t_{(j),b} - \bar{t}_{(j),b}| \geq \Delta\big),$$

$$\widehat{PFER}^0_{tusher}(\Delta) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{m}\Big(I\big(t_{j,b} \geq cut_{up}(\Delta)\big) + I\big(t_{j,b} \leq cut_{low}(\Delta)\big)\Big),$$

and a "false discovery rate"

$$\widehat{FDR}^0_{efron}(\Delta) = \widehat{PFER}^0_{efron}(\Delta)/R_{efron}(\Delta),$$

$$\widehat{FDR}^0_{tusher}(\Delta) = \widehat{PFER}^0_{tusher}(\Delta)/R_{tusher}(\Delta).$$

At first glance, there does not seem to be a big difference between the two versions of SAM. Both procedures should reject the same sets of hypotheses (genes) for a given value of the threshold Δ, if $|t_{(j)} - \bar{t}_{(j)}| \geq \Delta$ whenever $j \leq j_1$ or $j \geq j_2$, that is, whenever $t_{(j)} - \bar{t}_{(j)}$ is monotone in j. A fundamental difference exists, however, in the estimation of the expected number of Type I errors, $PFER = E(V|\mathrm{H}_0^C)$, which leads to the choice of Δ in each version. In Efron et al., the PFER is estimated by applying exactly the same steps to the original and permuted data, *i.e.*, the PFER is estimated by counting the number of genes with order statistics at least Δ away from the expected order statistics. In contrast, Tusher et al. compute order statistics only for the original data to obtain global cut–offs for the test statistics. These cut–offs are actually random variables, as they depend on the observed test statistics. In the permutation, the cut–offs are kept fixed and the PFER is estimated by counting the number of genes with test statistics above/below these global cut–offs. Figure 6 gives a graphical representation of the two variants: in the Quantile–Quantile plot of the test statistics, the Efron cut–offs are parallel to the identity line and the Tusher cut–offs are horizontal lines. For a fixed threshold Δ, the two procedures produce different estimates of the PFER. Next, we derive adjusted p–values for the Efron et al. and Tusher et al. procedures. For simplicity, it is assumed that the test statistics are continuous and hence that there are no ties. Also, rather than working with the PFER, we work with the PCER, which is simply the PFER divided by the total number of hypotheses and is a number between 0 and 1.

Adjusted p–values for Efron et al. SAM procedure. For the Efron et al. procedure, let $PCER^0_{efron}(\Delta) = E(V_{efron}(\Delta) \mid \mathrm{H}_0^C)/m = \sum_{j=1}^{m} pr\big(|T_{(j)} - \bar{t}_{(j)}| \geq \Delta \mid \mathrm{H}_0^C\big)/m$ denote the expected proportion of false positives under the complete null for a given threshold Δ. Here, $\bar{t}_{(j)}$ is taken to be a fixed estimate of $E(T_{(j)}|\mathrm{H}_0^C)$. One may then express the procedure in terms of the following adjusted p–values

$$\tilde{p}_{(j)} = PCER^0_{efron}(|t_{(j)} - \bar{t}_{(j)}|), \tag{15}$$

which can be estimated by permutation. Rejection of $\mathrm{H}_{(j)}$ for $\tilde{p}_{(j)} \leq \alpha$ controls the PCER at level α in the weak sense. Since the Efron et al. procedure is based on the distribution of the order statistics $T_{(j)}$ under the complete null, the subset pivotality condition of Westfall & Young (1993) does not hold. Note that the adjusted p–values $\tilde{p}_{(j)}$ are not necessarily monotone in j, as the differences $|t_{(j)} - \bar{t}_{(j)}|$ are not necessarily monotone. Thus, a test statistic T_i could possibly have a smaller adjusted p–value than a more "extreme" test statistic T_j with $|T_j| \geq |T_i|$. A stepwise

version of the Efron et al. procedure could be devised to deal with this feature.

Adjusted p–values for Tusher et al. SAM procedure. Similarly, for the Tusher et al. procedure, let $PCER^0_{tusher}(\Delta) = E(V_{tusher}(\Delta) \mid \mathrm{H}_0^C)/m = \sum_{j=1}^m \Big(pr\big(T_j \geq cut_{up}(\Delta) \mid \mathrm{H}_0^C\big) + pr\big(T_j \leq cut_{low}(\Delta) \mid \mathrm{H}_0^C\big)\Big)/m$ denote the expected proportion of false positives under the complete null for a given threshold Δ. One may then express the procedure in terms of the following adjusted p–values

$$\tilde{p}_{(j)} = PCER^0_{tusher}(\Delta_{(j)}), \tag{16}$$

where

$$\Delta_{(j)} = \begin{cases} \max\{t_{(k)} - \bar{t}_{(k)} : j \leq k \leq j_0\}, & \text{if } j \leq j_0 \\ \max\{\bar{t}_{(k)} - t_{(k)} : j_0 < k \leq j\}, & \text{if } j > j_0 \end{cases}$$

for $j_0 = \max\{j : \bar{t}_{(j)} \geq 0\}$. Due to the possibility that $t_{(k)} - \bar{t}_{(k)} < 0$ for some $k \leq j_0$ or $\bar{t}_{(k)} - t_{(k)} < 0$ for some $k > j_0$, it may happen that $\Delta_{(j)} < 0$ (this problem is not addressed in Tusher et al.). In such cases, the null $\mathrm{H}_{(j)}$ is never rejected and we define $\tilde{p}_{(j)} = 1$. Rejection of $\mathrm{H}_{(j)}$ for $\tilde{p}_{(j)} \leq \alpha$ controls the PCER at level α in the strong sense. Note that the rejection regions $S_\Delta = (-\infty, cut_{low}(\Delta)] \cup [cut_{up}(\Delta), \infty)$ are nested, that is, $S_{\Delta'} \subseteq S_\Delta$ for $\Delta' \geq \Delta$. Furthermore, the $\Delta_{(j)}$'s are monotone in j in each of the tails, thus the adjusted p–values are monotone in j for each tail (unlike the p–values for the Efron et al. procedure). The p–values are however not symmetric, as $t_{(j)} = t$ and $t_{(j)} = -t$ could correspond to different $\Delta_{(j)}$ and cut–offs $(cut_{low}(\Delta_{(j)}), cut_{up}(\Delta_{(j)}))$.

Both SAM procedures thus aim to control the PFER (or PCER), but the Efron et al. procedure only controls this error rate in the weak sense. The only difference between the Tusher et al. version of SAM and standard procedures which reject the null H_j for $|t_j| \geq c$ is in the use of asymmetric critical values chosen from the Quantile–Quantile plot (see Braver (1975) for the use of asymmetric critical values). Otherwise, SAM does not provide any new definition of Type I error rate nor any new procedure for controlling this error rate. In summary, the SAM procedure in Efron et al. amounts to rejecting $\mathrm{H}_{(j)}$ whenever $|t_{(j)} - \bar{t}_{(j)}| \geq \Delta$, where Δ is chosen to control the PFER weakly at a given level. By contrast, the SAM procedure in Tusher et al. rejects H_j whenever $t_j \geq cut_{up}(\Delta)$ or $t_j \leq cut_{low}(\Delta)$, where $cut_{low}(\Delta)$ and $cut_{up}(\Delta)$ are chosen from the Quantile–Quantile plot and such that the PFER is controlled strongly at a given level.

Control of FDR. The term "false discovery rate" is misleading, as the definition in SAM is different than the standard definition of Benjamini & Hochberg (1995): the SAM FDR is estimating $E(V|\mathrm{H}_0^C)/R$ and not $E(V/R)$ as in Benjamini & Hochberg. Furthermore, the FDR in SAM can be greater than one (*cf.* Table 3, p. 16 in Chu et al. (2000)). The issue of strong *vs.* weak control is only mentioned briefly in Tusher et al. and the authors claim that "SAM provides a reasonably accurate estimate for the true FDR".

Additional comments. The Efron SAM procedure considers the distribution of the order statistics $T_{(j)}$ under the complete null hypothesis and rejects the null hypothesis $\mathrm{H}_{(j)}$ for large deviations of $T_{(j)}$ from its expected value under the complete null. This approach could be refined by accounting for the different variances of the order statistics under the complete null, *i.e.*, by declaring gene j differentially expressed if $|t_{(j)} - \bar{t}_{(j)}| \geq sd_{(j)}\Delta$, where $sd^2_{(j)} = \sum_b (t_{(j),b} - \bar{t}_{(j)})^2/B$. A further

refinement would be to consider p–values for the order statistics: $p_{(j)} = pr\big(|T|_{(j)} \geq |t|_{(j)} \mid \mathrm{H}_0^C\big)$. These could be estimated by permutation as

$$p^*_{(j)} = \frac{\sum_b I(|t|_{(j),b} \geq |t|_{(j)})}{B}.$$

Following such an approach would lead to procedures that are similar to the ones considered to deal with the Golub et al. (1999) proposal. Since these procedures are based on the distribution of order statistics under the complete null, they only achieve weak control of Type I error rate under consideration. However, procedures based on $|T|_{(1)}$ might yield good tests of the complete null hypothesis H_0^C and of the null hypothesis corresponding to the largest order statistic $|T|_{(1)}$. They are not appropriate for testing other hypotheses while controlling the Type I error rate strongly.

3 Data

3.1 Simulated data

Artificial gene expression profiles $\mathbf{x}$ and binary responses y were generated as in Box 3 below.

Box 3. Type I error rate and power calculations for simulated data.

1. For the ith response group, $i = 1, 2$, generate n_i independent m–vectors or "artificial gene expression profiles" $\mathbf{x}$ from the Gaussian distribution with mean μ_i and covariance matrix Σ. The m_0 "genes" for which $\mu_1 = \mu_2$ are not differentially expressed and correspond to the true null hypotheses (see Table 3 for the model parameters used in the simulation).

2. For each of the m "genes", compute a two–sample t–statistic (with equal variances in the two response groups) comparing the gene expression levels in the two response groups. Apply the multiple testing procedures of Section 2 to determine which genes are differentially expressed for prespecified Type I error rates α (see Table 2 for a summary of the multiple testing procedures applied in the simulation).

3. For each procedure, record the number R_b of genes declared differentially expressed, the numbers V_b and T_b of Type I and II errors, and the false discovery rate $Q_b = V_b/R_b$ if $R_b > 0$ and 0 if $R_b = 0$.

Repeat steps 1–3 B times and estimate the Type I error rates and average power for each of the procedures as follows

$$\begin{aligned} PCER &= \frac{\sum_b V_b/m}{B}, \\ FWER &= \frac{\sum_b I(V_b \geq 1)}{B}, \\ FDR &= \frac{\sum_b Q_b}{B}, \\ \textit{Average power} &= 1 - \frac{\sum_b T_b/(m - m_0)}{B}. \end{aligned}$$

The 18 multiple testing procedures described in Table 2 were applied to each of the simulated datasets. Unadjusted p–values for each of the genes were computed in two ways: by permutation of the n responses and from the t–distribution with $n_1 + n_2 - 2$ degrees of freedom. Table 3 lists the different parameters used in the simulation.

3.2 Microarray data

3.2.1 Apo AI experiment, Callow et al.

The apo AI experiment was carried out as part of a study of lipid metabolism and artherosclerosis susceptibility in mice (Callow et al. 2000). 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 the experiment was to identify genes with altered expression in the livers of the apo AI 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 control C57Bl/6 mice. For each of these 16 mice, target cDNA was obtained from mRNA by reverse transcription and labeled using a 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 a green–fluorescent dye, Cy3. Target cDNA was hybridized to microarrays containing 6,356 cDNA probes, including 257 related to lipid metabolism. Each of the 16 hybridizations produced a pair of 16–bit images, which were processed using the software package *Spot* (Buckley 2000). The resulting fluorescence intensities were normalized as described in Dudoit et al. (2002). For each microarray $i = 1, \ldots, 16$, the base 2 logarithm of the Cy5/Cy3 fluorescence intensity ratio for gene j represents the expression response x_{ji} of that gene in either a control or treatment mouse.

Differentially expressed genes were identified by computing two–sample Welch's t–statistics for each gene j

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

where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ denote the average expression level of gene j in the n_1 control and n_2 treatment hybridizations, respectively. Similarly, s^2_{1j} and s^2_{2j} denote the variances of gene j'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 considered the multiple testing procedures of Section 2 and estimated unadjusted and adjusted p–values based on all possible $\binom{16}{8} = 12,870$ permutations of the treatment/control labels.

3.2.2 Bacteria experiment, Boldrick et al.

Boldrick et al. (2002) performed an *in vitro* study of the gene expression response of human peripheral blood mononuclear cells (PBMCs) to infection by pathogenic bacteria. One of the experiments (dose–response dataset) monitored the effect of three factors on the expression response of PBMCs: bacteria type, dose of the bacterial infection, and time after infection. Two types of bacteria were considered: the Gram–negative *B. pertussis* and the Gram–positive *S. aureus*; four doses of the pathogens were administered: 1X, 10X, 100X, and 1000X, where X represents the number of particles per cell (X=0.002 for the Gram–positive and X=0.004 for the Gram–negative); and the gene expression response was measured at five timepoints after infection: 0.5, 2, 4, 6, and 12 hours (extra

timepoints at 1 and 24 hours were recorded for dose 100X). A total of 44 hybridizations ($2 \times 4 \times 5$ plus 1 and 24 hour measurements for dose 100X) were performed using the Lymphochip, a specialized microarray comprising 18,432 elements enriched in genes that are preferentially expressed in lymphoid cells or which are of known immunological or oncological importance. In each hybridization, fluorescent cDNA targets were prepared from PBMC mRNA (red–fluorescent dye Cy5) and a reference sample derived from a pool of mRNA from 6 immune cell lines (green–fluorescent dye Cy3). The microarray scanned images were analyzed using the *GenePix* package and the resulting intensities were preprocessed as described in Boldrick et al.. For each microarray $i = 1, \ldots, 44$, the base 2 logarithm of the Cy5/Cy3 fluorescence intensity ratio for gene j represents the expression response x_{ji} of that gene in PBMCs infected by the Gram–positive or negative bacteria for one of the 22 dose $\times$ time combinations. The analysis below is based on a subset of 2,562 genes that were well–measured in both the dose–response and the diversity datasets (*cf.* Boldrick et al. for details on the preselection of the genes).

One of the goals of this experiment was to identify genes that have a different expression response to infection by the Gram–positive and the Gram–negative bacteria. As there are clearly dose and time effects on the expression response, the null hypothesis of no bacteria effect was tested for each gene based on a paired t–statistic. For any given gene, let d_i denote the difference in the expression response to infection by the Gram–negative and Gram–positive bacteria for the ith dose $\times$ time block, $i = 1, \ldots, 22$. The paired t–statistic is defined as $t = \bar{d}/\sqrt{s_d^2/n}$, where $\bar{d}$ is the average of the $n = 22$ differences d_i and s_d^2 is the variance of these 22 differences. In order to assess the statistical significance of the results, we considered the multiple testing procedures of Section 2 and estimated unadjusted and adjusted p–values based on all possible $2^{22} = 4,194,304$ permutations of responses *within* the 22 dose $\times$ time blocks.

3.2.3 Leukemia study, Golub et al.

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 chips 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 $X = (x_{ji})$, where x_{ji} denotes the expression level for gene j in mRNA sample i.

Differentially expressed genes in ALL and AML patients were identified by computing two–sample Welch's t–statistics for each gene j as in Section 3.2.1. In order to assess the statistical significance of the results, we considered the multiple testing procedures of Section 2 and estimated unadjusted and adjusted p–values based on $500,000$ random permutations of the ALL/AML labels.

4 Results

4.1 Simulated data

Figures 7 – 10 display plots of Type I error rates and power for different multiple testing procedures in the simulation study (see Box 3, Tables 2 and 3 for a description of the procedures and simulation parameters). For each procedure, adjusted p–values were computed as detailed in Section 2 and null hypotheses were rejected whenever their adjusted p–value was less than a prespecified level α. With the exception of `SAM efron`, `Golub sd`, and `Golub su`, all procedures controlled the claimed Type I error rate in the strong sense. As expected, procedures controlling the FWER were the most conservative, followed by procedures controlling the FDR.

Procedures controlling the FWER. The simulation study allowed us to compare the performance of single–step *vs.* stepwise procedures (*i.e.*, Bonferroni *vs.* Holm and Hochberg procedures, and single–step maxT *vs.* step–down maxT procedures). Although stepwise procedures are generally less conservative than single–step procedures, we found that with a very large number of hypotheses m the difference was minute. This is to be expected, as for large m the ratios $m/(m-k+1)$ are very close to 1 for moderate k. In contrast, incorporating the dependence structure among the genes, as in the maxT procedures, led in some situations to substantial gains in power over the Bonferroni, Holm, and Hochberg procedures. The largest gains in power were achieved for small samples sizes when the unadjusted p–values used in the Bonferroni, Holm, and Hochberg procedures were estimated by permutations ($n_1 = n_2 = 5$, Figure 10).

Procedures controlling the FDR. As expected, for a fixed nominal level $\alpha = 0.05$, the two FDR procedures provided substantial increases in power compared to the more conservative FWER procedures, but were in general less powerful than procedures controlling the PCER. The Benjamini & Yekutieli (2001) FDR procedure was more conservative than the Benjamini & Hochberg (1995) procedure (up to a 30% difference in power in Figure 10) and controlled the FDR much below the nominal 5% level (the actual FDR was usually less than 1%). For the simulation models, the standard Benjamini & Hochberg procedure controlled the FDR at the nominal 5% level, in spite of the correlation between the test statistics. The gene expression levels were simulated from multivariate Gaussian distributions; thus the t–statistics have a multivariate t–distribution. Although the multivariate t–distribution does not satisfy the positive regression dependency condition of Benjamini & Yekutieli, the standard step–up procedure nonetheless seems to control the FDR in our example.

SAM procedures. The adjusted p–values considered in the comparison study are based on the PCER, which is simply the PFER estimated by SAM divided by the number of hypotheses. Although `SAM efron` controls the PCER in the weak sense, *i.e.*, under the complete null (Figures 7–8), it is clear that it does not control the PCER in the strong sense (Figures 9–10). In the simulation in Figure 9, the actual PCER is twice the nominal PCER; in some of the simulations, the actual PCER was up to three times the nominal PCER. In contrast, the `SAM tusher` procedure (implemented in the SAM software package `http://www-stat.stanford.edu/ tibs/SAM/index.html`) treats each gene separately in the permutation and thus controls the PCER in the strong sense. The bottom left panels of Figures 7 – 10 display the average of the nominal SAM FDR, $\widehat{FDR}_b^0 = \widehat{PFER}_b^0/R_b$, as well as the average of the actual SAM FDR, Q_b, over the B simulations. In some of the simulations, the nominal SAM FDR was much smaller than the actual FDR; in other instances, the SAM FDR was actually greater than 1. The two SAM versions, especially the Tusher et al. version, are very

similar in power to standard procedures which control the PCER in the strong sense (`PCER ss t`, `PCER ss perm`, as described in Table 2).

Neighborhood analysis. As shown in Figures 7 – 10, the step–down procedure controls the FWER under the complete null (weak control), but fails do so when there are false null hypotheses. The step–up version of the Golub procedure does not control any known type of error rate, not even the PCER, and can lead to very high Type I error rates.

Nominal *vs.* **permutation** *p*–**values.** Because the gene expression levels were simulated as Gaussian random variables, the two–sample t–statistics should have a t–distribution with n_1+n_2-2 degrees of freedom. The simulation suggests that procedures based on permutation p–values can be much more conservative than procedures based on the nominal p–values from the t–distribution, the largest difference being for small sample sizes ($n_1 = n_2 = 5$, Figures 8, 10). The smaller the sample sizes n_1 and n_2, the smaller the total number of possible permutations, $B = \binom{n_1+n_2}{n_1}$, and hence the larger the smallest possible unadjusted p–value, $2/B$. Procedures most affected by the discreteness of the permutation unadjusted p–values were the FDR procedures and the Bonferroni, Holm, and Hochberg procedures. Procedures based on the maxT adjusted p–values, which are based on the test statistics rather than the unadjusted p–values, did not suffer from this problem.

4.2 Microarray data

The procedures described in Section 2 were applied to the three microarray datasets of Section 3. Genes with permutation adjusted p–values $\tilde{p}_j^* \leq \alpha$ were declared differentially expressed at level α for the Type I error rate controlled by the procedure under consideration. For each dataset, ordered adjusted p–values were plotted for each procedure in panels (a) and (b) of Figures 11, 12, and 13. The number R of genes declared differentially expressed was recorded for different values of the nominal Type I error rate α and plotted against α in panels (c) and (d) of the same figures. Panel (e) displays plots of adjusted p–values (on a log scale) *vs.* t–statistics. Finally, panel (f) displays Quantile–Quantile plots of the t–statistics. Results for the Golub et al. (1999) neighborhood analysis were not plotted in these figures, because it led to rejection of virtually all hypotheses for two–sided alternatives (*i.e.*, for tests based on absolute t–statistics). The results from the neighborhood analysis are discussed in greater detail below. As expected, the number of genes declared differentially expressed, R, was the greatest for procedures controlling the PCER (SAM procedures, procedures based on unadjusted p–values) and the smallest for procedures controlling the FWER (Bonferroni, Holm, Hochberg, maxT). The Efron et al. SAM method yielded the largest number of rejected hypotheses (in the thousands for the leukemia and bacteria datasets, for nominal Type I error rates as low as 10%), but, as suggested by the simulation study, this happens at the cost of an increased Type I error rate compared to the nominal SAM error rate. Also as expected, the Tusher et al. SAM variant and the standard p–value based procedures for controlling the PCER (unadjusted permutation p–values) produced very similar results. As in the simulation study, the Benjamini & Yekutieli (2001) FDR procedure was much more conservative than the standard Benjamini & Hochberg (1995) FDR procedure. Procedures based on the step–down maxT adjusted p–values generally provided a less conservative test than either the Bonferroni, Holm, or Hochberg procedures. The Bonferroni procedure yielded similar results as its step–down (Holm) and step–up (Hochberg) analogs.

The different multiple testing procedures behaved similarly for the leukemia and bacteria datasets;

however, their behavior on the apo AI dataset was quite different due to the smaller sample sizes. Aside from the PCER procedures, only the maxT and standard Benjamini & Hochberg (1995) procedures rejected any hypothesis at levels $\alpha \leq 20\%$. 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 unadjusted p–values must be at least $2/12,870$. As a result, the Bonferroni adjusted p–values must be at least $6,356 \times 2/12,870 \approx 1$. This dataset clearly highlights the power of the maxT procedure over standard Bonferroni–like procedures or even some FDR procedures.

SAM procedures. For the SAM Efron and Tusher procedures, Figure 14 displays plots of t–statistics $t_{(j)}$, adjusted p–values $\tilde{p}^*_{(j)}$, and associated thresholds $\Delta_{(j)}$ and cut–offs $(cut_{low}(\Delta_{(j)}), cut_{up}(\Delta_{(j)}))$. For the apo AI experiment, the thresholds $\Delta_{(j)}$ and adjusted p–values $\tilde{p}^*_{(j)}$ for the Efron procedure are not monotone in j (panel (a)). The thresholds $\Delta_{(j)}$ and adjusted p–values $\tilde{p}^*_{(j)}$ for the Tusher procedure are monotone in j, but a number of $\Delta_{(j)}$s resulted in infinite upper cut–offs (red plotting symbols in panel (c)). As could be expected from the Quantile–Quantile plot of Figure 11, the cut–offs $(cut_{low}(\Delta_{(j)}), cut_{up}(\Delta_{(j)}))$ are not symmetric (panel (e)). Furthermore, the problematic issue of negative thresholds $\Delta_{(j)}$ arises near the origin for this dataset: this happens, for example, when the expected order statistics are positive and the observed order statistics are less than their expected values. In this case, the corresponding null hypotheses are never rejected by SAM Tusher; we thus assign adjusted p–values of 1 to these hypotheses. Panels (b), (d), and (f) display analogous plots for the leukemia study. In contrast to the apo AI experiment, the SAM Tusher cut–offs for these data are fairly symmetric.

Neighborhood analysis. For the leukemia study, Figure 15, panel (a), displays plots of the step–down and step–up adjusted p–values for the neighborhood analysis of Golub et al. (1999). The p–values were calculated for three different types of test statistics: absolute t–statistic $|t_j|$ (two–sided alternative), t–statistic t_j (one–sided alternative of over–expression in AML), and t–statistic $-t_j$ (one–sided alternative of over–expression in ALL). The two–sided adjusted p–values were unreasonably small, leading to rejection of virtually all hypotheses. The one–sided p–values were larger, but again led to a very large number of rejections: over a thousand for a Golub Type I error rate $G(c)$ of 0.05. Figure 15, panel (b), displays plots of the Golub Type I error rate, or p–value for the number of rejections $R(c)$, $G(c) = pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$, *vs.* critical values c, for the three different types of alternatives. The adjusted p–values for the step–down and step–up procedures in panel (a) are virtually identical because of the monotonicity of the Type I error rate $G(c)$ in panel (b). Panels (c) – (e) are plots of the observed number of rejection $R(c) = r(c)$ and permutation quantiles of $R(c)$ against critical values c for the three types of tests. One can clearly see that even for a small Type I error rate $G(c) = \alpha$, the critical value c is very small in magnitude. Similar plots were presented in Figure 2 of Golub et al. for one–sided tests; again, the critical values (given by the intersections of the curves for the observed and permutation quantiles of the number of rejections R) were very close to zero. The same qualitative behavior was observed for the other two datasets (figures not shown).

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. These eight sequences correspond to only four distinct genes: apo AI (3 copies), apo CIII (2 copies), sterol C5 desaturase (2 copies), and a novel EST (1 copy). All changes were confirmed by real–time quantitative PCR (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. Callow et al. showed that the down–regulation of apo CIII was actually due to genetic polymorphism rather than lack of apo AI. The presence of apo AI and apo CIII among the differentially expressed genes thus provides a check of the statistical method, if 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.

Bacteria experiment. In this experiment, 66 spotted DNA sequences had maxT adjusted p–values less than 0.05 and several of these sequences actually corresponded to the same genes: CD64 (3 copies), Iκ B alpha (5 copies), SHP–1 (2 copies), plasma gelsolin (2 copies) (see Appendix A). In contrast to the apo AI experiment, the spotted DNA sequences exhibited a continuum of change and we could not identify a group of genes that clearly stood out from the rest. A detailed discussion of the biological findings can be found in Boldrick et al. (2002).

Leukemia study. There was significant overlap between the gene lists in Golub et al. (1999) (p. 533 and Figure 3B) and the list of 92 genes with maxT adjusted p–values less 0.05 (see Appendix B). The reader is referred to Golub et al. for a description of these genes and their involvement in ALL and AML.

5 Discussion

In this article, we have discussed different approaches to large scale multiple hypothesis testing in the context of microarray experiments. Standard multiple testing procedures, as well as recent proposals for microarray experiments, were compared in terms of their Type I error rate control and power, using gene expression and simulated datasets.

The comparison study highlighted five desirable properties of multiple testing procedures for large multiplicity problems such as those arising in microarray experiments: (i) control of an appropriate and precisely defined *Type I error rate*; (ii) *strong* control of the Type I error rate, *i.e.*, control of this error rate under any combination of true and false null hypotheses; (iii) taking into account the *joint* distribution of the test statistics; (iv) reporting the results in terms of *adjusted p–values*; (v) availability of efficient *resampling* algorithms for nonparametric procedures.

A number of recent articles have addressed the question of multiple testing in the context of microarray experiments (Dudoit et al. 2002, Efron et al. 2000, Golub et al. 1999, Kerr et al. 2000, Manduchi et al. 2000, Tusher et al. 2001, Westfall et al. 2001). However, not all proposed solutions were cast in the standard statistical multiple testing framework and some procedures fail to provide adequate Type I error rate control. In particular, the Type I error rates considered in some of these papers were rather loosely defined, thus making it difficult to assess the properties of the multiple testing procedures. Regarding item (i), control of the per–comparison error rate (PCER) is often not adequate, as it does not really deal with the multiplicity problem. Although not stated explicitly in Efron et al. (2000) and Tusher et al. (2001), both SAM procedures are based on computing the PFER, a constant multiple of the PCER. Given the information provided in Golub et al. (1999), we determined that the Type I error rate in the neighborhood analysis is $G(c) = pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$, that is, as a p–value for the number of rejected hypotheses under the complete null (in this case, the number of Type I errors, $V(c)$). This is a rather unusual definition and a more detailed discussion of the procedure and its limitations is given below. In the microarray setting, where it is

very unlikely that no genes are differentially expressed, property (ii) of strong control of the Type I error rate is essential, whether it be the FWER, PCER, or FDR. This was demonstrated in the simulation study, where the Type I error rates for some of the procedures were no longer controlled when a subset of null hypotheses were allowed to be false. The Efron et al. (2000) version of SAM and the neighborhood analysis of Golub et al. (1999) both rely on the distribution of *ordered* test statistics under the complete null hypothesis, and therefore only provide *weak* control of the Type I error rate. Regarding point (iii), the comparison study highlighted the gains in power that can be achieved by taking into account the joint distribution of the gene expression levels when assessing statistical significance (maxT procedures *vs.* Bonferroni, Holm, and Hochberg procedures). Rather than simply reporting rejection or not of the null hypothesis of no differential expression for a given gene, we have found *adjusted p–values* (point (iv)) to be particularly useful and flexible summaries of the strength of the evidence in favor of differential expression. The adjusted p–value for a particular gene reflects the overall false positive rate for the entire experiment when genes with smaller p–values are declared differentially expressed. Adjusted p–values may also be used to summarize and compare the results from different multiple testing procedures. Finally, as mentioned in item (v), efficient resampling–based nonparametric multiple testing procedures are needed to take into account the complex dependency structures between gene expression levels. Such procedures were proposed in Westfall & Young (1993) for FWER control, however, due to the large–scale nature of microarray experiments, computational issues remain to be addressed (Ge & Dudoit 2002), in addition to methodological ones.

Procedures controlling the FWER. Results on both simulated and microarray datasets suggest that the Westfall & Young (1993) step–down maxT procedure is well–adapted for microarray experiments. Like the classical Bonferroni procedure, it provides strong control of the FWER. However, it can be substantially more powerful than the Bonferroni, Holm, and Hochberg procedures, because it takes into account the dependence structure between the test statistics. In addition, the maxT procedure performed very well compared to other procedures (including some FDR procedures), when adjusted p–values were estimated by permutation. It does not suffer as much as others from the small number of possible permutations associated with small sample sizes, because the adjusted p–values are based on the test statistics rather than the unadjusted p–values. For a detailed comparison of maxT and minP procedures, the reader is referred to Ge & Dudoit (2002). The main advantage of the minP procedure is that it provides balanced adjustments; the issue of balance is important when the test statistics for the different hypotheses are not identically distributed (Beran 1988, Westfall & Young 1993, p. 50). However, estimation of the minP p–values by resampling is more costly computationally than for maxT p–values, because the unadjusted p–values must be computed before considering the distribution of their successive minima. Also, when adjusted p–values are estimated by permutation 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.

Procedures controlling the FDR. In the microarray setting, where thousands of comparisons are performed simultaneously and a fairly large number of genes are expected to be differentially expressed, procedures controlling the FDR present a promising alternative to more conservative approaches controlling the FWER. In this context, one may be willing to bear a few false positives as long as their number is small in comparison to the number of rejected hypotheses. Most FDR controlling procedures proposed thus far either control the FDR under restrictive dependency structures (*e.g.* independence or positive regression dependency) or do not exploit the joint distribution

of the test statistics. It would thus be useful to develop FDR controlling procedures, in the spirit of the Westfall & Young (1993) minP and maxT procedures for FWER control, that strongly control the FDR and take into account the dependence structure between test statistics. Such procedures could lead to increased power, as in the case of FWER control. Initial work in this direction can be found in Yekutieli & Benjamini (1999), assuming unadjusted p–values for the true null hypotheses are independent of the p–values for the false null hypotheses. Reiner et al. (2001) applied different FDR controlling procedures to the apo AI dataset.

SAM procedures. The Efron et al. (2000) and Tusher et al. (2001) versions of SAM seem very similar at first glance. A fundamental difference exists, however, in the estimation of the expected number of Type I errors, $E(V|\mathrm{H}_0^C)$, leading to the choice of the threshold Δ. The difference lies in the use of ordered test statistics in Efron et al. to estimate this error rate under the complete null hypothesis. In the Efron et al. (2000) version, the PFER is thus only weakly controlled, while in the Tusher et al. (2001) version it is strongly controlled. The only difference between the Tusher et al. version of SAM and standard procedures which reject the null H_j for $|t_j| \geq c$ is in the use of asymmetric critical values chosen from a Quantile–Quantile plot. Otherwise, SAM does not provide any new definition of Type I error rate, nor any new procedure for controlling this error rate. There are a number of practical problems linked to the implementation of the Tusher et al. SAM procedure (software package `http://www-stat.stanford.edu/ tibs/SAM/index.html`). The user does not choose a significance level ahead of time; rather, the PFER is estimated for a fixed set of thresholds Δ. In some cases, it can be hard to select Δ for a prespecified PFER. Using the adjusted p–values derived in Section 2.7.2 provides a more flexible implementation of the procedure. A problem remains, however, with the choice of the "origin" j_0, when the differences $t_{(j)} - \bar{t}_{(j)}$ are negative to the right of the origin or positive to the left of the origin. This can lead to negative thresholds $\Delta_{(j)}$ for the adjusted p–value calculations. In such cases, we set the p–values to one, as the corresponding nulls are never rejected by SAM. As part of the SAM method Efron et al. (2000) and Tusher et al. (2001) suggest test statistics for identifying differentially expressed genes for different types of responses. These test statistics are based on standard t– or F–statistics, with a "fudge" factor in the denominator to deal with the small variance problem encountered in microarray experiments (Lönnstedt & Speed 2002). The "shrunken" statistics were not used in the comparison study of Section 4, because we wanted to focus on control of Type I error for a given choice of test statistics.

Neighborhood analysis. Although not stated explicitly in Golub et al. (1999), the error rate controlled by the neighborhood analysis is a p–value for the number of rejected hypotheses under the complete null, that is, $G(c) = pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$. A critical value c is then chosen to control this unusual error rate at a prespecified level α. Given the data, the function $G(c)$ is not in general monotone in c, and there are possibly several values of c with $G(c) = \alpha$. The non–monotonicity issue was not addressed in Golub et al.. In Section 2.7.1, we considered a step–down and a step–up version of the neighborhood analysis to deal with this problem and derived corresponding adjusted p–values. Because the neighborhood analysis is based on the distribution of the order statistics under the complete null, only weak control of the Type I error rate can be achieved. In turns out that the step–down version controls the FWER weakly, while the step–up version does not control any standard error rate, not even the PCER. Application of the neighborhood analysis to the three microarray datasets of Section 3 resulted in unreasonably long lists of genes declared differentially expressed, especially for two–sided hypotheses. This can be seen also in Figure 2 of Golub et al. (1999), where a critical value near zero is used for the test statistics and thousands of genes are declared differentially expressed. Golub et al. applied the neighborhood analysis separately for

each type of one–sided hypothesis (over–expression in AML compared to ALL and *vice versa*); it is not clear how an overall Type I error rate can be obtained.

Open questions.

In the comparison study of Section 4, only two–sided tests were considered. In practice, however, researchers are interested in determining the direction of rejection for the null hypotheses, *i.e.*, in determining whether genes are over– or under–expressed in, say, treated cells compared to untreated cells. This raises the issue of Type III error rate control, where Type III error refers to correctly declaring that a gene is differentially expressed, but deciding that it is over–expressed when in fact it is really under–expressed, or *vice versa*. Control of these errors in addition to Type I errors brings in additional complexities (Finner 1999), and will not be considered here.

We have considered thus far only one null hypothesis per gene. When comparing several treatments or in the context of factorial experiments (Section 3.2.2), one may be interested in testing several hypotheses simultaneously for each gene. For example, when monitoring the gene expression response of a particular type of cells to K treatments, one may wish to consider all $K(K-1)/2$ pairwise treatment comparisons and determine which correspond to significant treatment differences. A number of procedures are available to deal with such testing situations one gene at a time (*e.g.* procedures of Tukey and Scheffé). An open problem is the extension of these methods to the 2D–testing problem where several hypotheses are tested simultaneously for each of thousands of genes.

A related issue is the development of resampling methods for estimating adjusted p–values in the context of factorial experiments, which impose some structure on the columns of the gene expression data matrix. For the 3–factor bacteria experiment, Gram–positive and Gram–negative labels were permuted *within* the 22 dose $\times$ time blocks, to respect the blocking structure of the experiment and allow the possibility of dose and time effects on the expression response of PBMCs. Permutation is only one of several resampling approaches which can be used to estimate adjusted p–values. Bootstrap procedures, parametric and non–parametric, should also be investigated, as they may allow estimation of adjusted p–values for more specific null hypotheses.

The methods described above operate on individual genes. However, it is well known that genes are expressed in a coordinated manner, for example, through pathways or the sharing of the same transcription factors. It would be interesting to develop multiple testing procedures for identifying *groups* of differentially expressed genes, where the groups may be defined *a priori*, from the knowledge of pathways, say, or by cluster analysis. Initial work in this area can be found in Tibshirani et al. (2001).

Finally, we did not consider Bayesian approaches, which constitute an important class of methods for the identification of differentially expressed genes (Efron et al. 2000, Manduchi et al. 2000, Newton et al. 2001). In such methods, the criterion for identifying differentially expressed genes is based on the posterior probability of differential expression, *i.e.*, the probability that a particular gene is differentially expressed given the data for all genes. This is in contrast to the so–called frequentist methods reviewed in this paper, which are based on adjusted p–values, *i.e.*, on the joint distribution of the test statistics given suitably defined null hypotheses. It would be interesting to compare and, when possible, reconcile these two approaches. Efron et al. (2001) and Storey (2001) discuss

Bayesian interpretations of the FDR.

Most multiple testing procedures considered in this paper are implemented in an `R` package (Ihaka & Gentleman 1996), `multtest`, which may be downloaded from
`http://www.R-project.org`.

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., Jr, J. H., 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. Statist. Soc. B* **57**: 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency, *Annals of Statistics* . Accepted.

Beran, R. (1988). Balanced simultaneous confidence sets, *Journal of the American Statistical Association* **83**: 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 immue responses to bacteria, *Proc. Natl. Acad. Sci.* **99**(2): 972–977.

Braver, S. L. (1975). On splitting the tails unequally: a new perspective on one-versus two-tailed tests, *Educational and Psychological Measurement* **35**: 283–301.

Brown, P. O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays, *The Chipping Forecast*, Vol. 21, Supplement to Nature Genetics, pp. 33–37.

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.

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

Chu, G., Goss, V., Narasimhan, B. & Tibshirani, R. (2000). SAM "Significance Analysis of Microarrays" - Users guide and technical document, *Technical report*, Stanford University.

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., 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, *Annals of Mathematical Statistics* **29**: 1095–111.

Efron, B., Storey, J. D. & Tibshirani, R. (2001). Microarrays, empirical bayes methods, and false discovery rates. Submitted.

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

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors, *Annals of Statistics* **27**: 274–289.

Ge, Y. & Dudoit, S. (2002). Fast algorithm for resampling-based p-value adjustment in multiple testing. (In preparation).

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., 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.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance, *Biometrika* **75**: 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* **6**: 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test, *Biometrika* **75**: 383–386.

Hommel, G. & Bernhard, G. (1999). Bonferroni procedures for logically related hypotheses, *Journal of Statistical Planning and Inference* **82**: 119–128.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**: 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**: 819–837.

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

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. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology* **14**: 1675–1680.

Lönnstedt, I. & Speed, T. P. (2002). Replicated microarray data, *Statistica Sinica* **12**(1): 31–46.

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

National Reading Panel (1999). *Teaching Children to Read*, National Institute of Child Health and Human Development, National Institutes of Health.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. . & Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology* **8**: 37–52.

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.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons, *Journal of the American Statistical Association* **73**: 479–485.

Reiner, A., Yekutieli, D. & Benjamini, Y. (2001). Using resampling-based fdr controlling multiple test procedures for analyzing microarray gene expression data.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified bonferroni inequality, *Biometrika* **77**: 663–665.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., 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.

Scheffé, H. (1959). *The analysis of variance*, John Wiley & Sons.

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

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures, *Journal of the American Statistical Association* **81**: 826–831.

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

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance, *Biometrika* **73**: 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. Submitted.

Tibshirani, R., Hastie, T., Narasimhan, B., Eisen, M., Sherlock, G., Brown, P. & Botstein, D. (2001). Exploratory screening of genes and clusters from microarray experiments, *Technical report*, Department of Statistics, Stanford University.

Troendle, J. F. (1996). A permutational step-up method of testing multiple outcomes, *Biometrics* **52**: 846–859.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proc. Natl. Acad. Sci.* **98**: 5116–5121.

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.), *Statistical Methods in Molecular Biology*.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference, *Biometrics* **48**(4): 1005–1013.

Yang, Y. H., Buckley, M. J., Dudoit, S. & Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data, *Journal of Computational and Graphical Statistic* **11**(1): 108–136.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. & Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research* **30**(4): e15.

Yang, Y. H., Dudoit, S., Luu, P. & Speed, T. P. (2001). Normalization for cDNA microarray data, *in* M. L. Bittner, Y. Chen, A. N. Dorsel & E. R. Dougherty (eds), *Microarrays: Optical Technologies and Informatics*, Vol. 4266 of *Proceedings of SPIE*, pp. 141–152.

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.

Table 1: Properties of multiple testing procedures.

Procedure	Type I error rate	Strong or weak control	Stepwise	Dependence structure
Bonferroni	FWER	Strong	Single	General/Ignore
Šidák	FWER	Strong	Single	Positive orthant dependence
minP	FWER	Strong	Single	Subset pivotality
maxT	FWER	Strong	Single	Subset pivotality
Holm (1979)	FWER	Strong	Down	General/Ignore
Step–down Šidák	FWER	Strong	Down	Positive orthant dependence
Step–down minP	FWER	Strong	Down	Subset pivotality
Step–down maxT	FWER	Strong	Down	Subset pivotality
Hochberg (1988)	FWER	Strong	Up	Some dependence (Simes)
Troendle (1996)	FWER	Strong	Up	Some dependence
Benjamini & Hochberg (1995)	FDR	Strong	Up	Positive regression dependence
Benjamini & Yekutieli (2001)	FDR	Strong	Up	General/Ignore
Yekutieli & Benjamini (1999)	FDR	Strong	Up	Some dependence
Unadjusted p–values	PCER	Strong	Single	General/Ignore
SAM Tusher et al. (2001)	PFER (PCER)	Strong	Single	General/Hybrid
SAM Efron et al. (2000)	PFER (PCER)	Weak	Single	General
Golub et al. (1999), step–down	$pr(R \geq r \vert \mathrm{H}_0^C)$ (FWER)	Weak	Down	General
Golub et al. (1999), step–up	$pr(R \geq r \vert \mathrm{H}_0^C)$	Weak	Up	General

Note. By "General/Ignore", we mean that a procedure controls the claimed Type I error rate for general dependency structures, but does not explicitly take into account the joint distribution of the test statistics. For the Tusher et al. (2001) SAM version, the term "General/Hybrid" refers to the fact only the marginal distribution of the test statistics is considered when computing the PFER. The test statistics are considered jointly only to determine the cut–offs $cut_{up}(\Delta)$ and $cut_{low}(\Delta)$ from the Quantile–Quantile plot.

Table 2: Multiple testing procedures applied in simulation study.

Name	Description
`Bonf t`	Bonferroni procedure, reject H_j if $\tilde{p}_j \leq \alpha$ (equation (1)), p_j computed from t–distribution with $n_1 + n_2 - 2$ df.
`Bonf perm`	Bonferroni procedure, reject H_j if $\tilde{p}^*_j \leq \alpha$ (equation (1)), p^*_j computed by permutation as in Box 1.
`Holm t`	Holm procedure, reject H_{r_j} if $\tilde{p}_{r_j} \leq \alpha$ (equation (5)), p_j computed from t–distribution with $n_1 + n_2 - 2$ df.
`Holm perm`	Holm procedure, reject H_{r_j} if $\tilde{p}^*_{r_j} \leq \alpha$ (equation (5)), p^*_j computed by permutation as in Box 1.
`Hoch t`	Hochberg procedure, reject H_{r_j} if $\tilde{p}_{r_j} \leq \alpha$ (equation (9)), p_j computed from t–distribution with $n_1 + n_2 - 2$ df.
`Hoch perm`	Hochberg procedure, reject H_{r_j} if $\tilde{p}^*_{r_j} \leq \alpha$ (equation (9)), p^*_j computed by permutation as in Box 1.
`maxT ss`	single–step maxT procedure, reject H_j if $\tilde{p}^*_j \leq \alpha$ (equation (4)).
`maxT sd`	step–down maxT procedure, reject H_{r_j} if $\tilde{p}^*_{r_j} \leq \alpha$ (equation (8), Box 2).
`FDR BH t`	Benjamini & Hochberg (1995) procedure, reject H_{r_j} if $\tilde{p}_{r_j} \leq \alpha$ (equation (10)), p_j computed from t–distribution with $n_1 + n_2 - 2$ df.
`FDR BH perm`	Benjamini & Hochberg (1995) procedure, reject H_{r_j} if $\tilde{p}^*_{r_j} \leq \alpha$ (equation (10)), p^*_j computed by permutation as in Box 1.
`FDR BY t`	Benjamini & Yekutieli (2001) procedure, reject H_{r_j} if $\tilde{p}_{r_j} \leq \alpha$ (equation (11)), p_j computed from t–distribution with $n_1 + n_2 - 2$ df.
`FDR BY perm`	Benjamini & Yekutieli (2001) procedure, reject H_{r_j} if $\tilde{p}^*_{r_j} \leq \alpha$ (equation (11)), p^*_j computed by permutation as in Box 1.
`PCER ss t`	Reject H_j if $p_j \leq \alpha$, p_j computed from t–distribution with $n_1 + n_2 - 2$ df.
`PCER ss perm`	Reject H_j if $p^*_j \leq \alpha$, p^*_j computed by permutation as in Box 1.
`SAM efron`	Efron et al. (2000) SAM procedure (Section 2.7.2), reject $H_{(j)}$ if $\tilde{p}^*_{(j)} \leq \alpha$, estimated by permutation (equation(15)).
`SAM tusher`	Tusher et al. (2001) SAM procedure (Section 2.7.2), reject $H_{(j)}$ if $\tilde{p}^*_{(j)} \leq \alpha$, estimated by permutation (equation (16)).
`Golub sd`	Golub et al. (1999) neighborhood analysis, step–down version (Section 2.7.1), reject $H_{(j)}$ if $\tilde{p}^*_{(j)} \leq \alpha$, estimated by permutation (equation (13)).
`Golub su`	Golub et al. (1999) neighborhood analysis, step–up version (Section 2.7.1), reject $H_{(j)}$ if $\tilde{p}^*_{(j)} \leq \alpha$, estimated by permutation (equation (14)).

Table 3: Simulation parameters. Here, a_n denotes an n–vector with entries equal to a and b_n denotes the n–vector $1.5 * (1, 2, \ldots, n)/n$. I_m denotes the $m \times m$ identity matrix and S_m is the $m \times m$ covariance matrix for a random subset of m genes in the Boldrick et al. experiment described in Section 3.2.2.

Parameter	Value
Number of "genes"	$m = 500$
Mean vectors	$\mu_1 = 0_m$
	$\mu_2 = 0_m$
	$\mu_2 = [b_{m*0.1}, -b_{m*0.1}, 0_{m*0.8}]$
Covariance matrix	$\Sigma = I_m$
	$\Sigma = S_m$
Sample sizes	$n_1 = n_2 = 5$
	$n_1 = n_2 = 25$
Number of simulations	$B = 500$
Number of permutations for SAM	$B_{sam} = 1,000$ or all $\binom{n_1+n_2}{n_1}$
Number of permutations for neighborhood analysis	$B_{golub} = 1,000$ or all $\binom{n_1+n_2}{n_1}$
Number of permutations for unadjusted p–values	$B_{perm} = 25,000$ or all $\binom{n_1+n_2}{n_1}$
Nominal Type I error rate (PCER, FWER, or FDR)	$\alpha = 0.05$

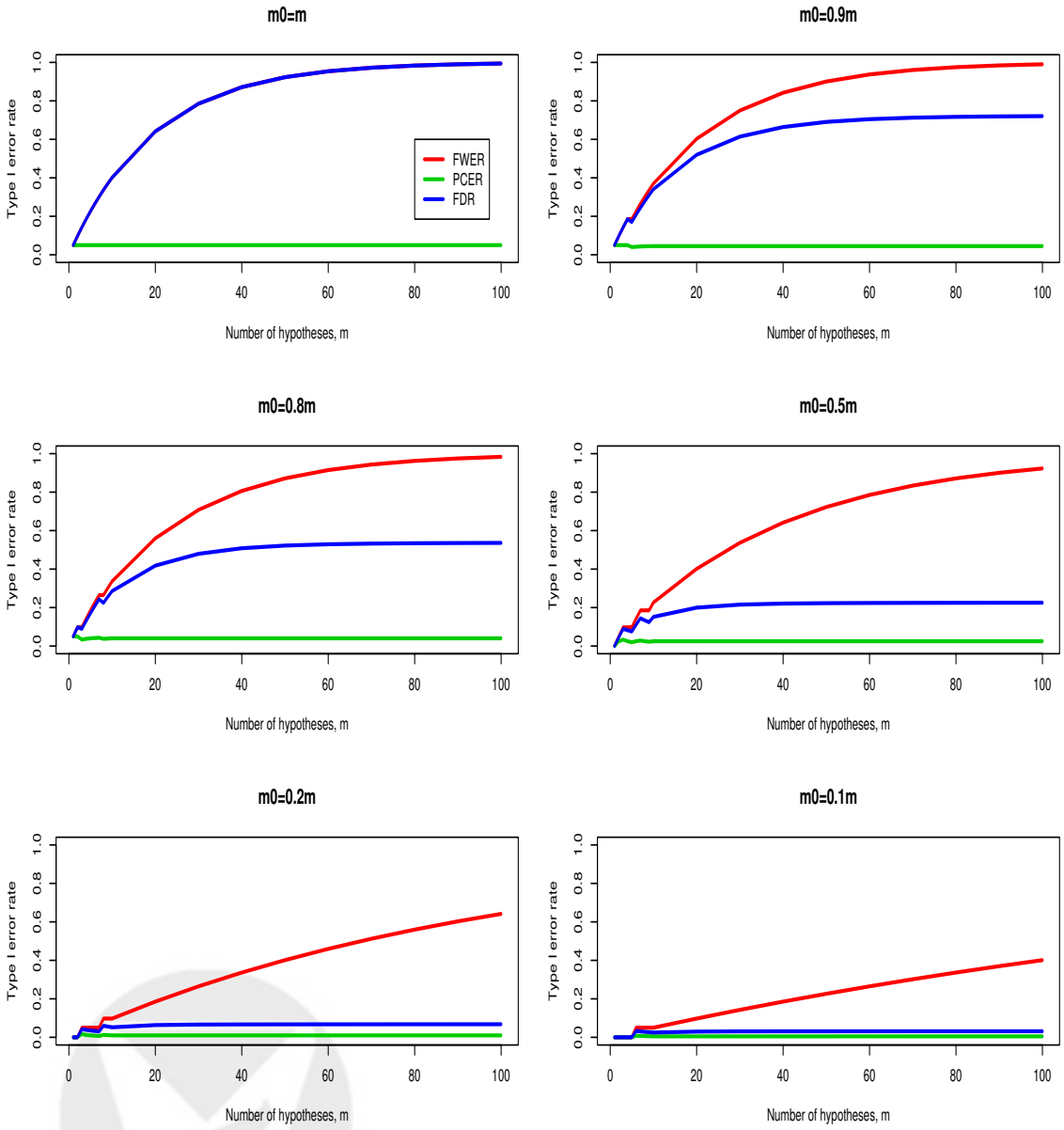


Figure 1: *Type I error rates, simple example.* Plot of Type I error rates *vs.* number of hypotheses m, for different proportions of true null hypotheses, $m_0/m = 1, 0.9, 0.8, 0.5, 0.2, 0.1$. The model and multiple testing procedures are described in Section 2.2. The individual test size is $\alpha = 0.05$ and the parameter d was set to 1. The non–smooth behavior for small m is due to the fact that it is not always possible to have exactly 90%, 80%, 50%, 20%, or 10% of true null hypotheses and rounding to the nearest integer is necessary.

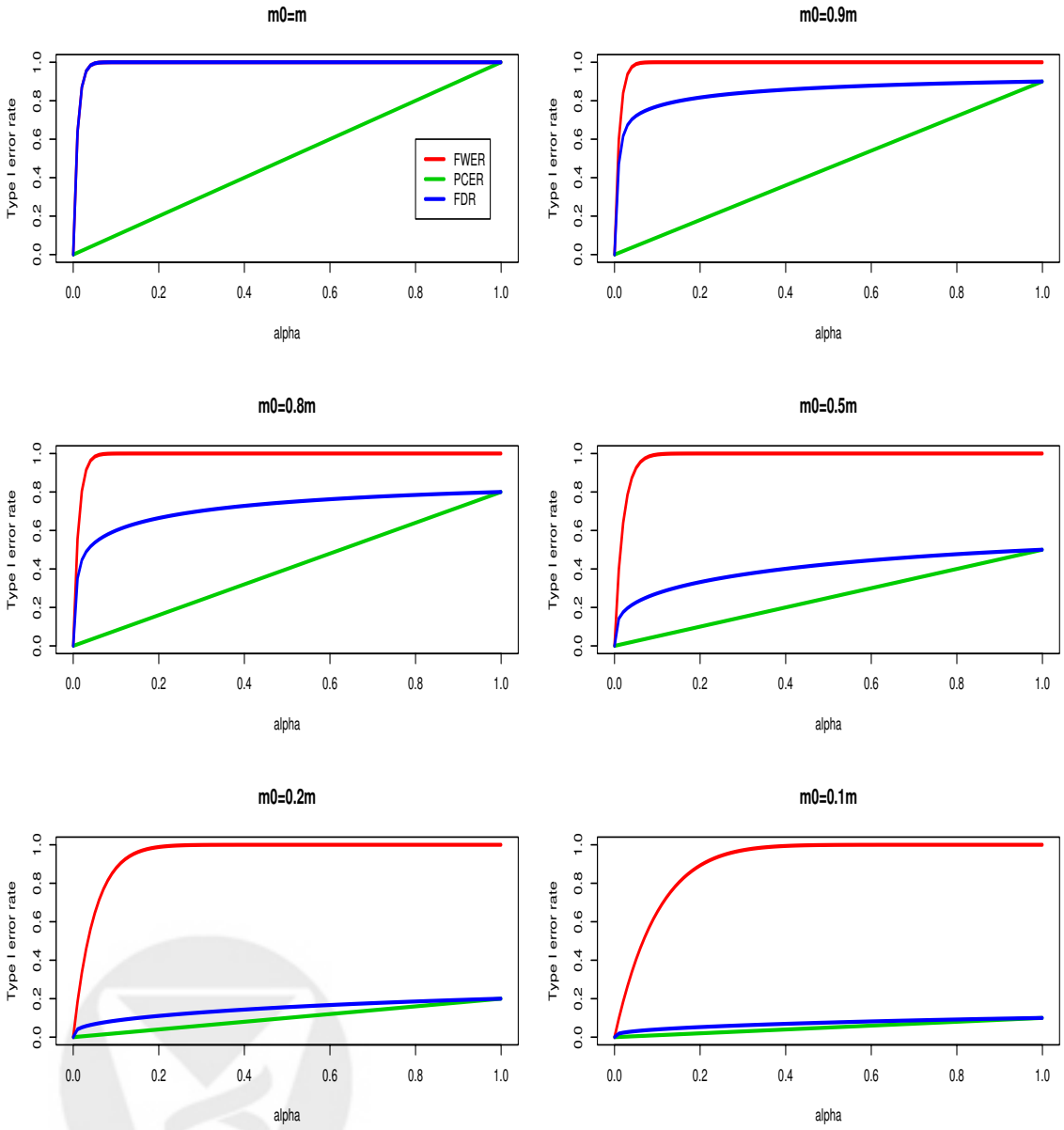


Figure 2: *Type I error rates, simple example.* Plot of Type I error rates *vs.* individual test size α, for different proportions of true null hypotheses, $m_0/m = 1\,, 0.9,\ 0.8,\ 0.5,\ 0.2,\ 0.1$. The model and multiple testing procedures are described in Section 2.2. The number of hypotheses is $m = 100$ and the parameter d was set to 1.

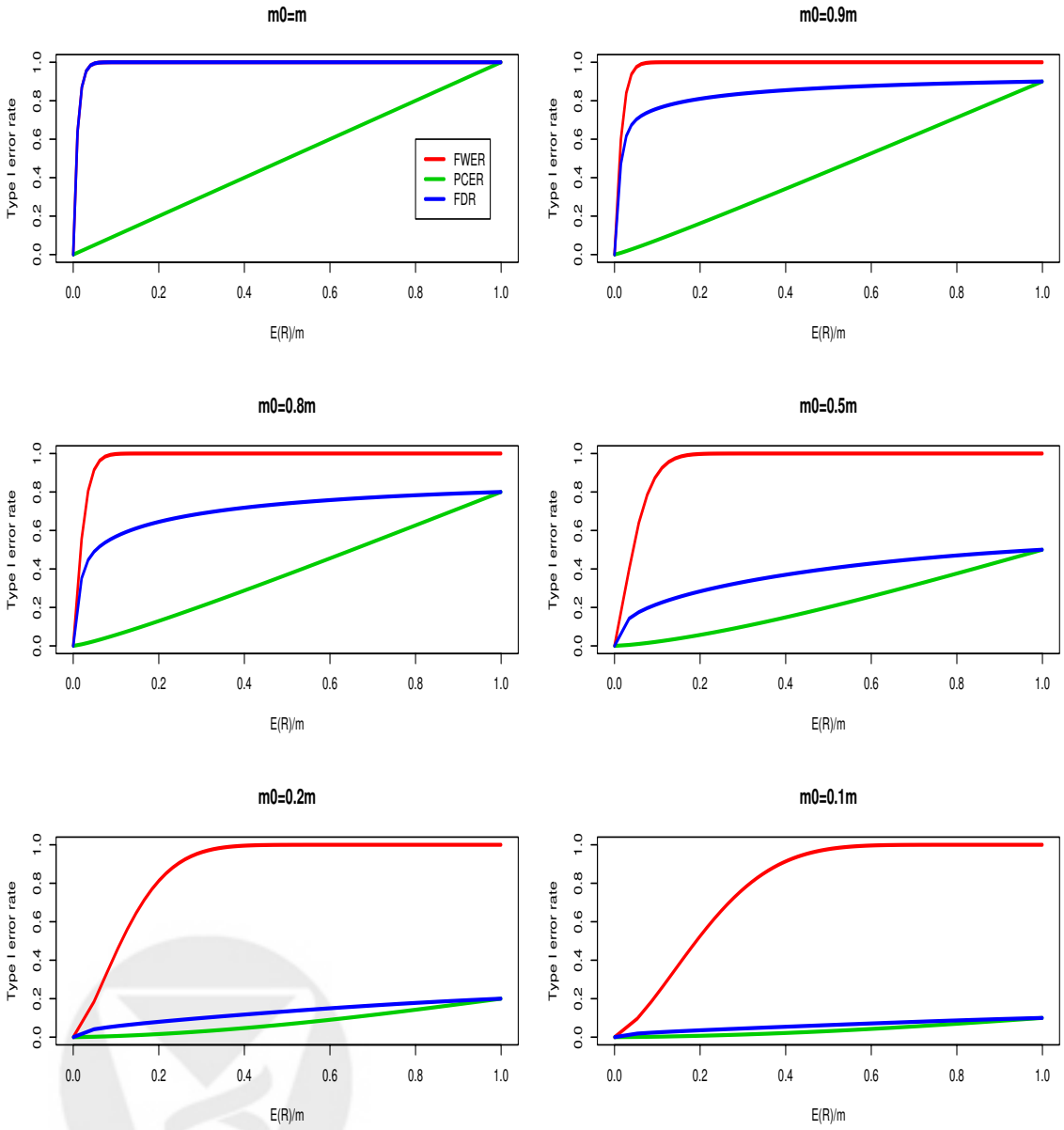


Figure 3: *Type I error rates, simple example.* Plot of Type I error rates *vs.* expected proportion of rejected hypotheses $E(R)/m$, for different proportions of true null hypotheses, $m_0/m = 1\,, 0.9, 0.8, 0.5, 0.2, 0.1$. The model and multiple testing procedures are described in Section 2.2. The number of hypotheses is $m = 100$ and the parameter d was set to 1.

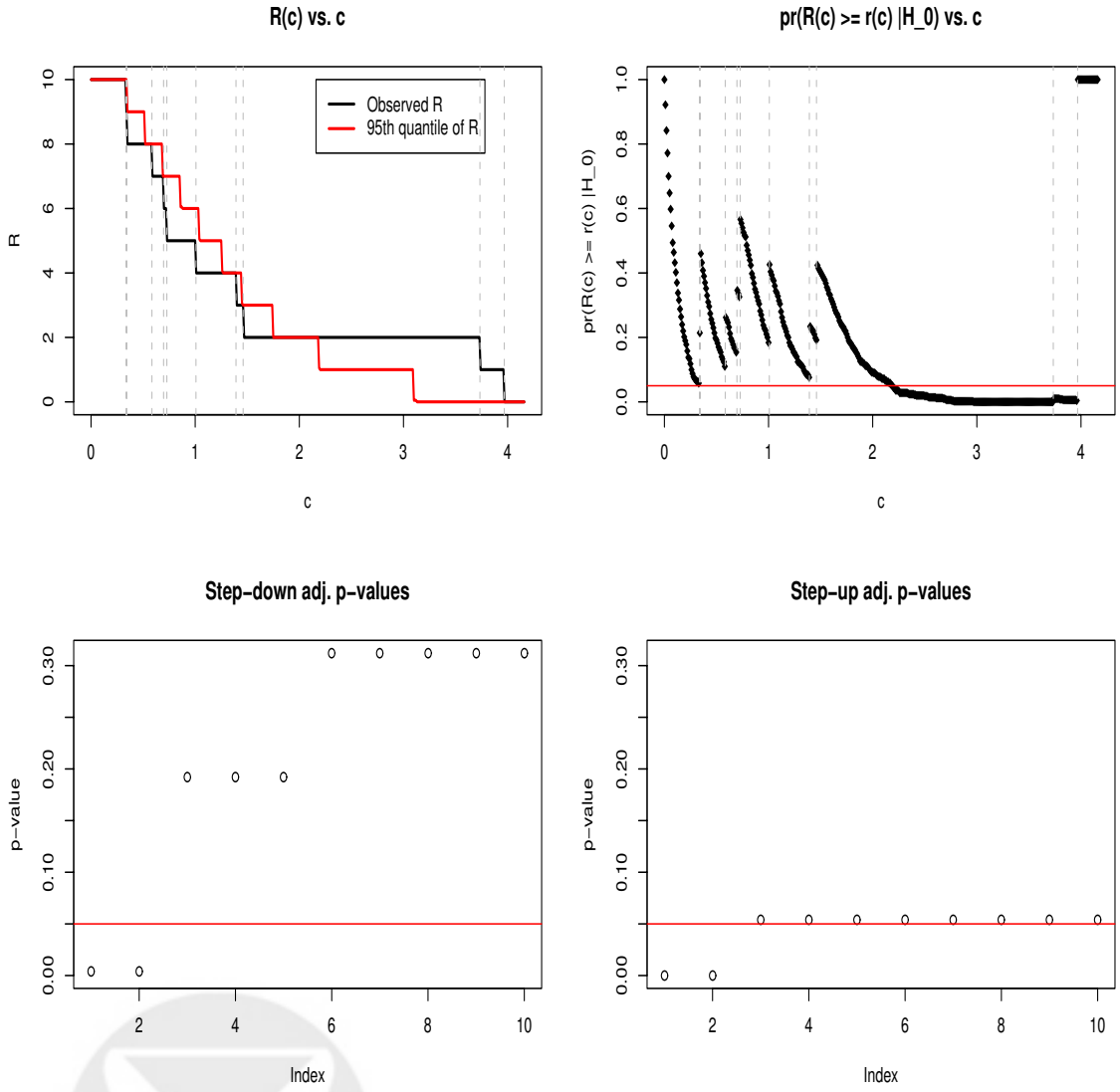


Figure 4: *Neighborhood analysis, $m = 10$ hypotheses.* Upper left panel: plot of observed number of rejected hypotheses $r(c)$ (black) and 95th quantile of $R(c)$ (red) *vs.* critical value c; vertical dashed lines correspond to observed values of $|t|_{(j)}$. Upper right panel: plot of $G(c) = pr\big(R(c) \geq r(c) \mid \mathrm{H}_0^C\big)$ *vs.* c. Lower left panel: plot of step–down adjusted p–values; the horizontal line corresponds to a 5% significance level. Lower right panel: plot of step–up adjusted p–values. Data were simulated as in Table 3, with $n_1 = n_2 = 20$, $m = 10$, $\mu_1 = 0_{10}$, $\mu_2 = [1_2, 0_8]$, $\Sigma = I_{10}$. $B = 500$ permutations of the class labels were used to estimate the quantiles of $R(c)$ and the adjusted p–values.

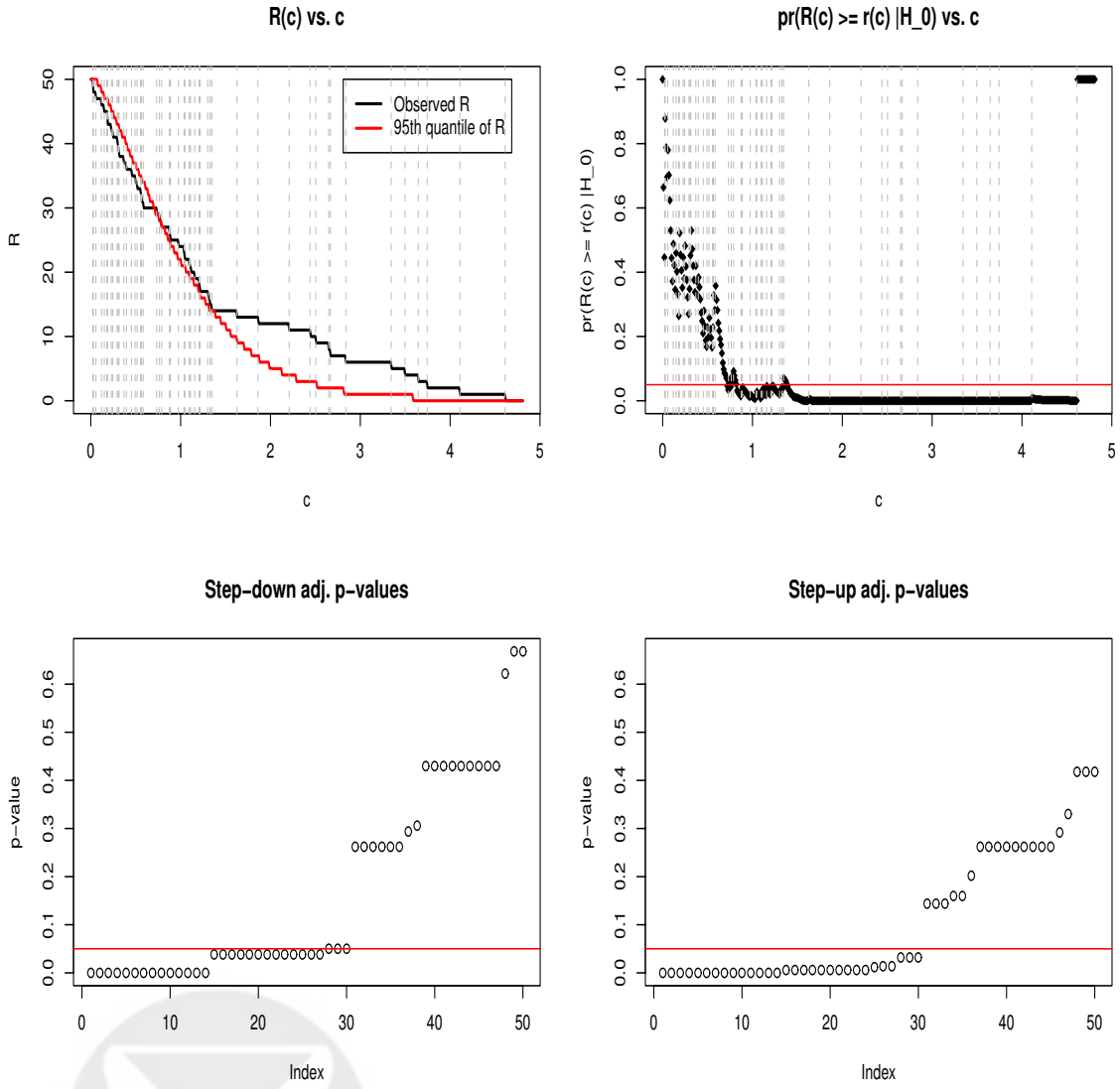


Figure 5: *Neighborhood analysis, $m = 50$ hypotheses.* Upper left panel: plot of observed number of rejected hypotheses $r(c)$ (black) and 95th quantile of $R(c)$ (red) *vs.* critical value c; vertical dashed lines correspond to observed values of $|t|_{(j)}$. Upper right panel: plot of $G(c) = pr\big(R(c) \geq r(c) \mid \mathrm{H}_0^C\big)$ *vs.* c. Lower left panel: plot of step–down adjusted p–values; the horizontal line corresponds to a 5% significance level. Lower right panel: plot of step–up adjusted p–values. Data were simulated as in Table 3, with $n_1 = n_2 = 20$, $m = 50$, $\mu_1 = 0_{50}$, $\mu_2 = [1_{10}, 0_{40}]$, $\Sigma = I_{50}$. $B = 500$ permutations of the class labels were used to estimate the quantiles of $R(c)$ and the adjusted p–values.

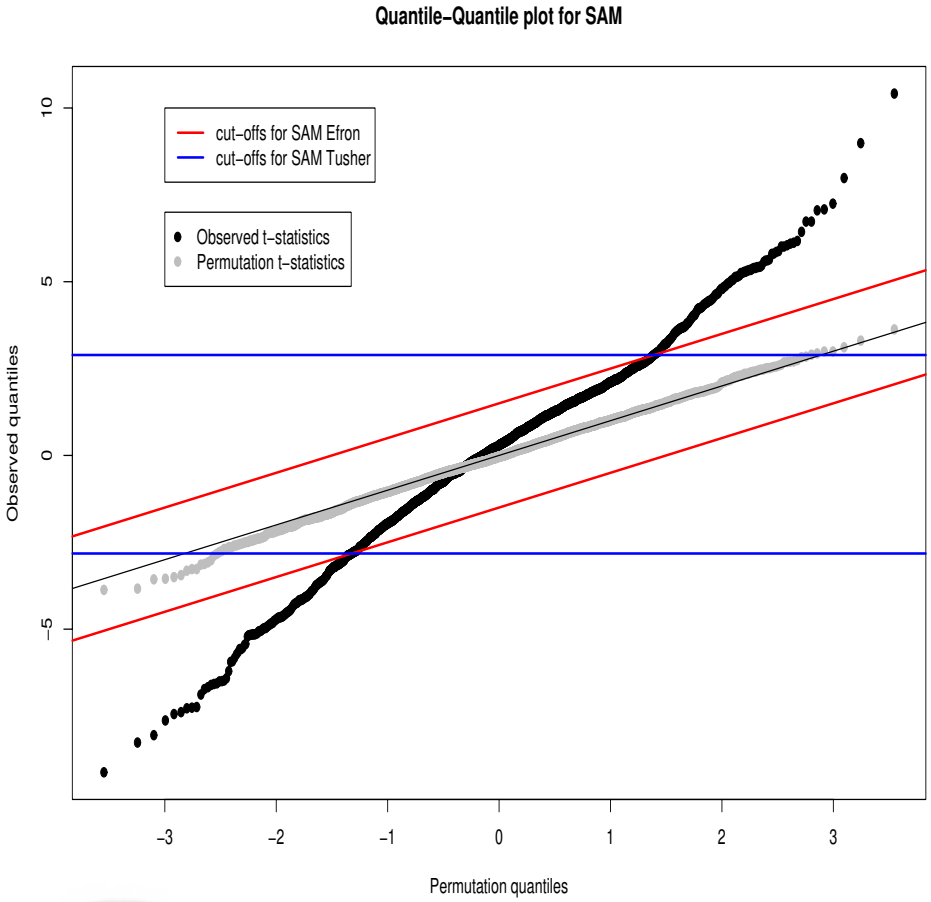


Figure 6: *SAM cut–offs.* Quantile–Quantile plot of t–statistics for the bacteria dataset. The t–statistics for the observed data are plotted using black symbols, the t–statistics for a particular permutation of the Gram–positive and negative labels are plotted using gray symbols. The red lines correspond to the SAM cut–offs for the Efron variant with $\Delta = 1.5$, the blue horizontal lines correspond to the SAM cut–offs for the Tusher variant with $\Delta = 1.5$. Note that in the permutation, the Tusher cut–offs (blue) lead to more rejected hypotheses and hence a more conservative estimate of the PFER than the Efron cut–offs (red), for the same threshold Δ.

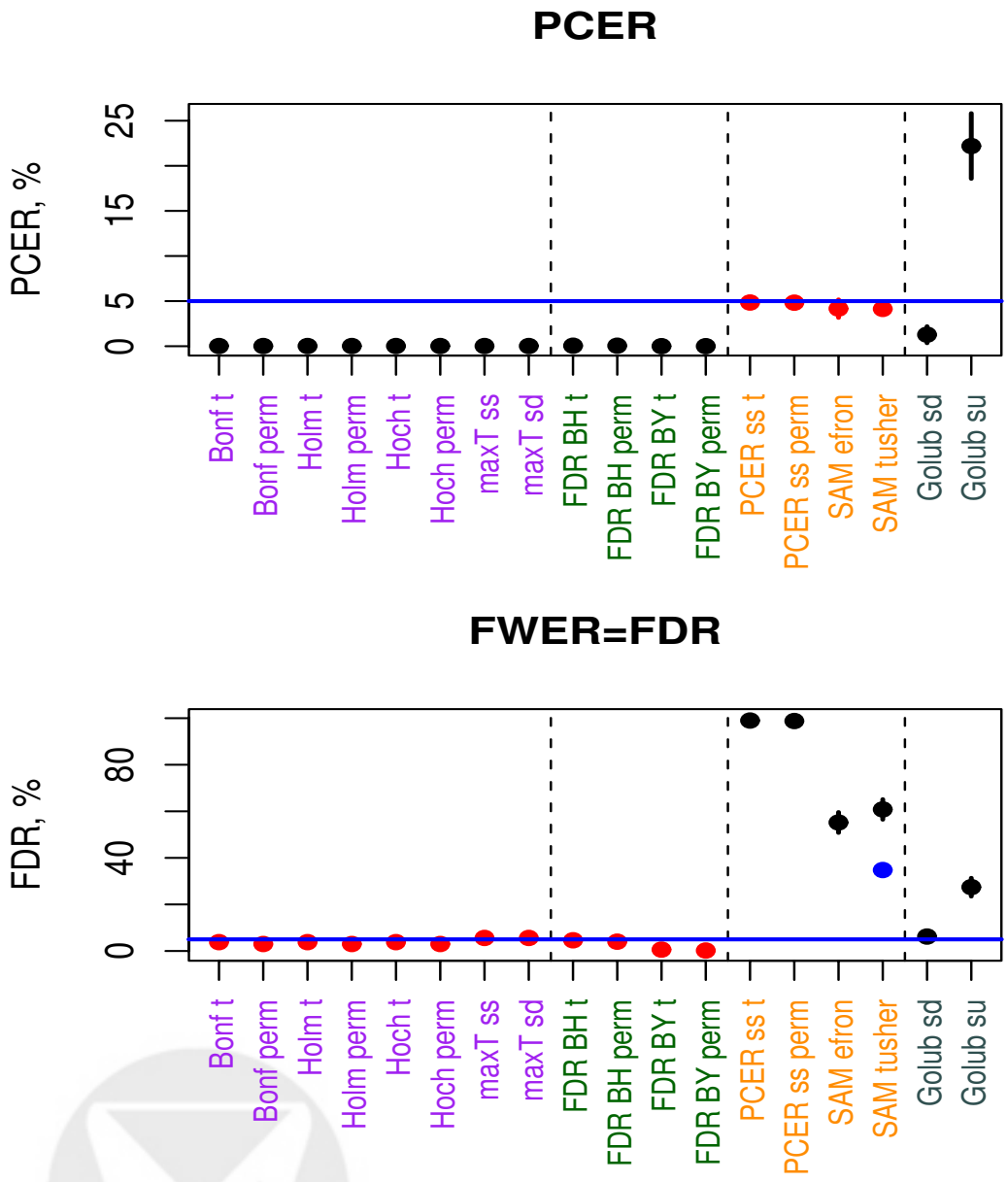


Figure 7: *Simulation 0 – Complete null.* PCER, FWER, and FDR for different multiple testing procedures. The top panel displays $PCER = \sum_b R_b/mB$ and simulation standard errors (2 SE); the bottom panel displays $FWER = \sum_b I(R_b \geq 1)/B = FDR$ and simulation standard errors (2 SE). For each definition of the Type I error rate, the procedures which are designed to control this error rate are highlighted in red. The blue line corresponds to a Type I error rate of $\alpha = 5\%$. In the FDR panel, the simulation averages of the nominal SAM FDRs are plotted in blue (for `SAM efron`, the nominal FDR is 3.25). The parameter values in Simulation 0 are $B = 500$, $B_{sam} = 1,000$, $B_{golub} = 1,000$, $B_{perm} = 25,000$, $\alpha = 0.05$, $m = 500$, $\mu_1 = \mu_2 = 0_m$, $\Sigma = S_m$, and $n_1 = n_2 = 25$.

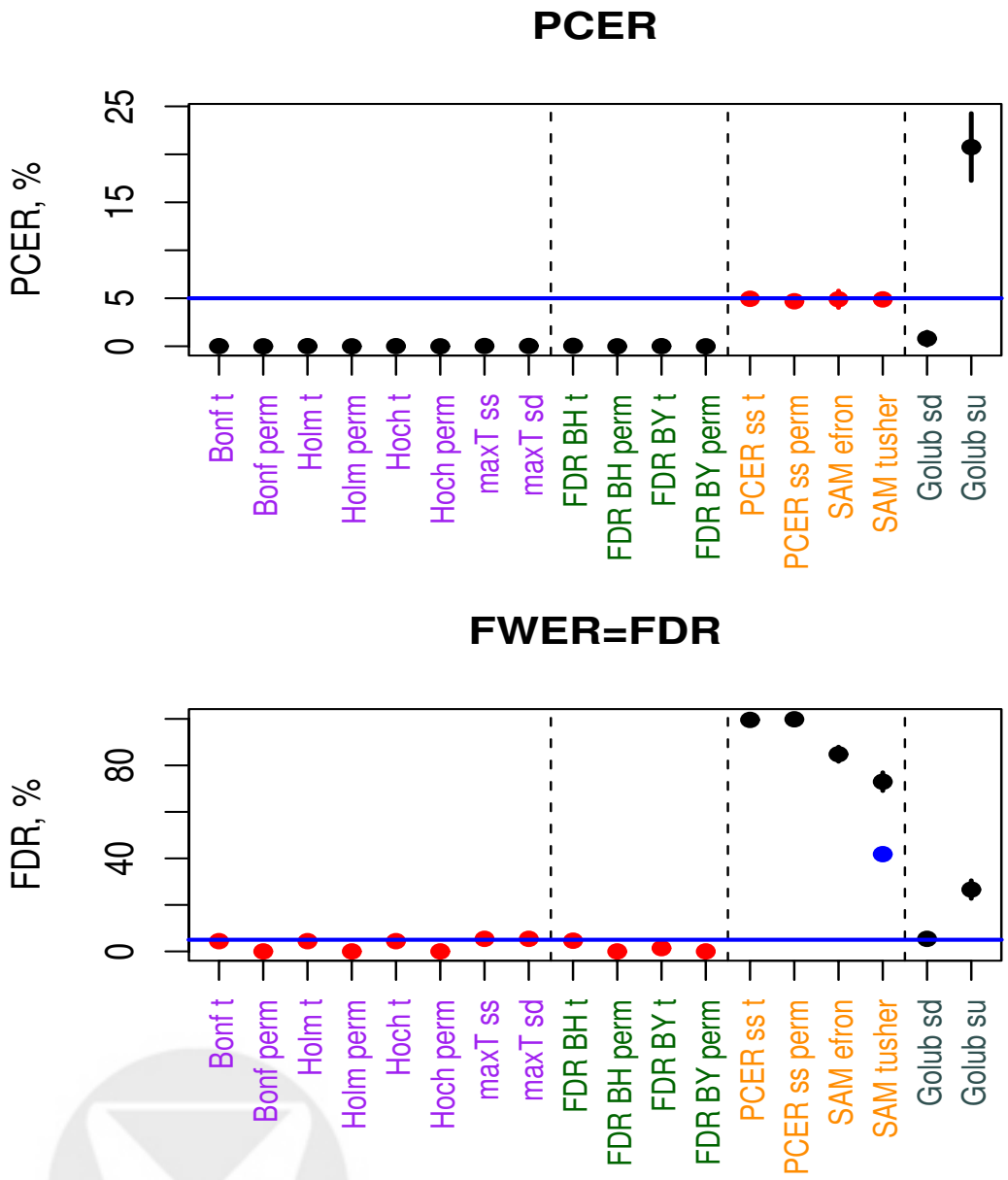


Figure 8: *Simulation 0b – Complete null.* PCER, FWER, and FDR for different multiple testing procedures. The top panel displays $PCER = \sum_b R_b/mB$ and simulation standard errors (2 SE); the bottom panel displays $FWER = \sum_b I(R_b \geq 1)/B = FDR$ and simulation standard errors (2 SE). For each definition of the Type I error rate, the procedures which are designed to control this error rate are highlighted in red. The blue line corresponds to a Type I error rate of $\alpha = 5\%$. In the FDR panel, the simulation averages of the nominal SAM FDRs are plotted in blue (for `SAM efron`, the nominal FDR is 3.68). The parameter values in Simulation 0b are $B = 500$, $B_{sam} = \binom{10}{5} = 252$, $B_{golub} = 252$, $B_{perm} = 252$, $\alpha = 0.05$, $m = 500$, $\mu_1 = \mu_2 = 0_m$, $\Sigma = S_m$, and $n_1 = n_2 = 5$.

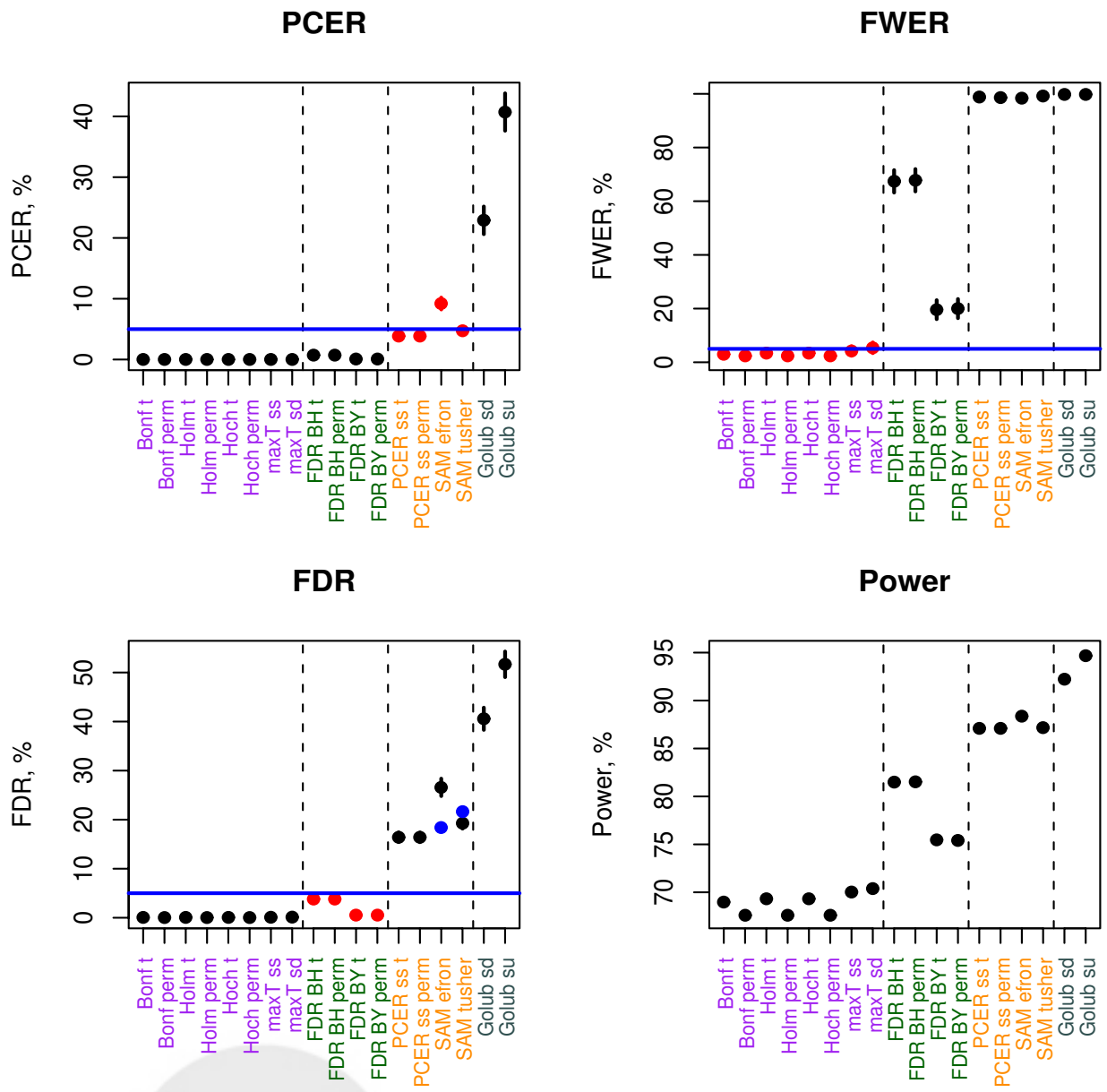


Figure 9: *Simulation 1 – 20% false nulls.* PCER, FWER, FDR, and power for different multiple testing procedures. The top left panel displays $PCER = \sum_b V_b/mB$ and simulation standard errors (2 SE); the top right panel displays $FWER = \sum_b I(V_b \geq 1)/B$ and simulation standard errors (2 SE); the bottom left panel displays $FDR = \sum_b Q_b/B$ and simulation standard errors (2 SE); the bottom right panel displays $Average\ power = 1 - \sum_b T_b/B(m - m_0)$ and simulation standard errors (2 SE). For each definition of the Type I error rate, the procedures which are designed to control this error rate are highlighted in red. The blue line corresponds to a Type I error rate of $\alpha = 5\%$. In the FDR panel, the simulation averages of the nominal SAM FDRs are plotted in blue. The parameter values in Simulation 1 are $B = 500$, $B_{sam} = 1,000$, $B_{golub} = 1,000$, $B_{perm} = 25,000$, $\alpha = 0.05$, $m = 500$, $\mu_1 = 0_m$, $\mu_2 = [b_{m*0.1}, -b_{m*0.1}, 0_{m*0.8}]$, $\Sigma = S_m$, and $n_1 = n_2 = 25$.

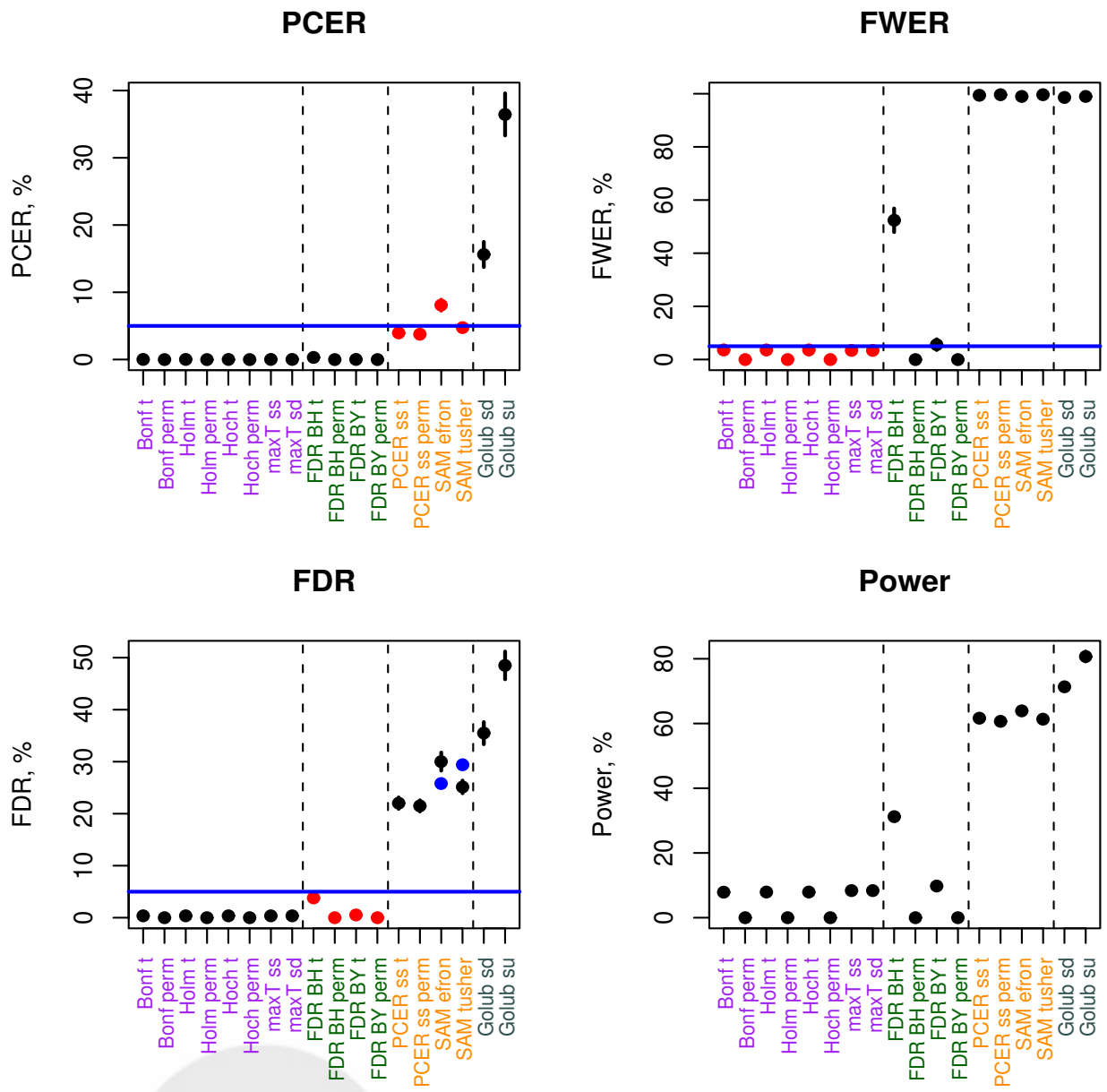


Figure 10: *Simulation 1b – 20% false nulls.* PCER, FWER, FDR, and power for different multiple testing procedures. The top left panel displays $PCER = \sum_b V_b/mB$ and simulation standard errors (2 SE); the top right panel displays $FWER = \sum_b I(V_b \geq 1)/B$ and simulation standard errors (2 SE); the bottom left panel displays $FDR = \sum_b Q_b/B$ and simulation standard errors (2 SE); the bottom right panel displays $Average\ power = 1 - \sum_b T_b/B(m - m_0)$ and simulation standard errors (2 SE). For each definition of the Type I error rate, the procedures which are designed to control this error rate are highlighted in red. The blue line corresponds to a Type I error rate of $\alpha = 5\%$. In the FDR panel, the simulation averages of the nominal SAM FDRs are plotted in blue. The parameter values in Simulation 1b are $B = 500$, $B_{sam} = \binom{10}{5} = 252$, $B_{golub} = 252$, $B_{perm} = 252$, $\alpha = 0.05$, $m = 500$, $\mu_1 = 0_m$, $\mu_2 = [b_{m*0.1}, -b_{m*0.1}, 0_{m*0.8}]$, $\Sigma = S_m$, and $n_1 = n_2 = 5$.

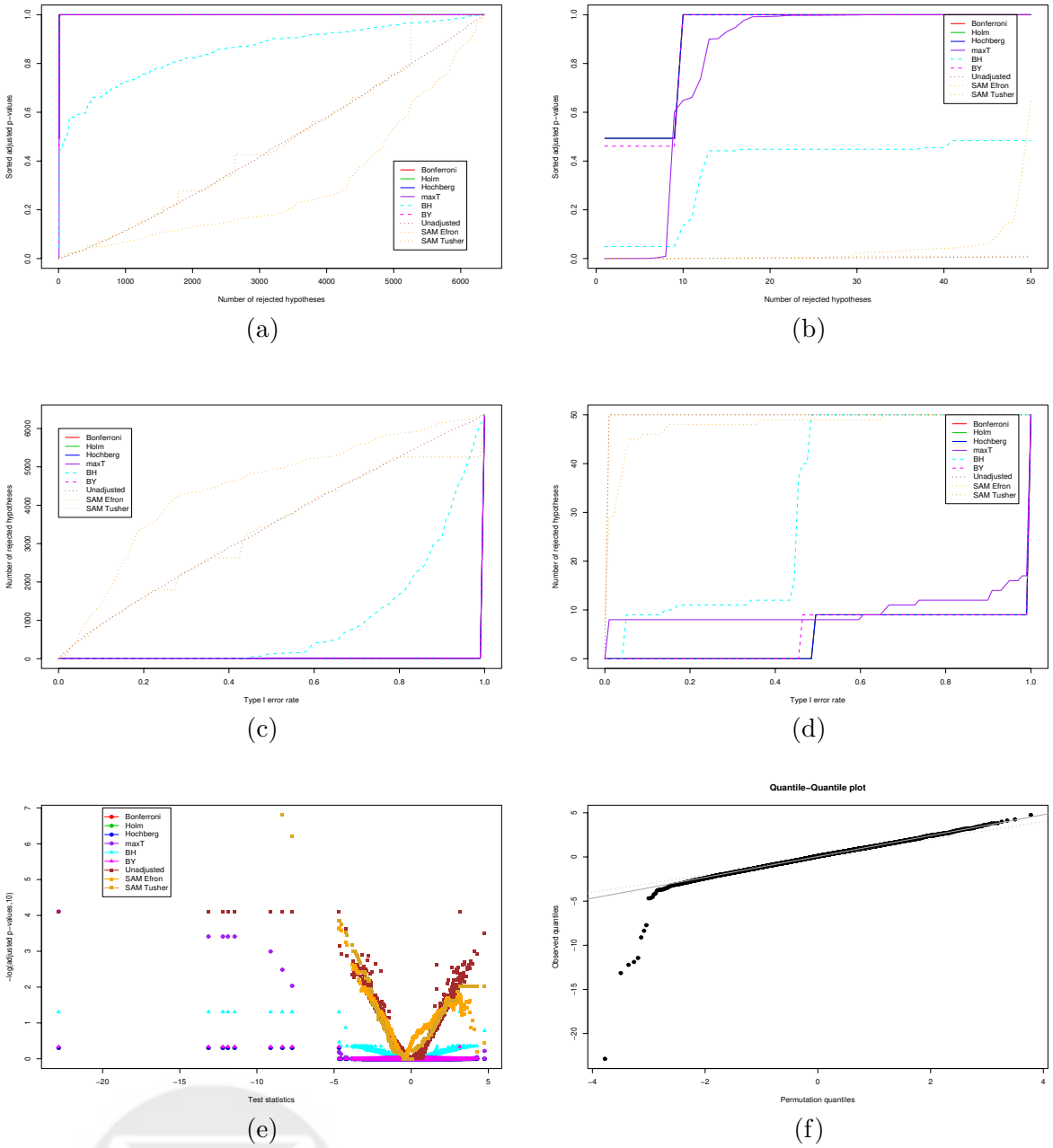


Figure 11: *Apo AI experiment.* Panels (a) and (b): Plot of sorted adjusted p–values $\tilde{p}^*_{r_j}$ *vs.* j. Panels (c) and (d): Number of genes declared differentially expressed, R, *vs.* nominal Type I error rate, α. Panels (b) and (d) are enlargements of panels (a) and (c), respectively, for the 50 genes with the smallest maxT adjusted p–values. Panel (e): Plot of adjusted p–values $-\log_{10}\tilde{p}^*_j$ *vs.* t–statistics t_j. Panel (f): Quantile–Quantile plot of t–statistics; the dotted line is the identity line and the dashed line passes through the first and third quartiles. Adjusted p–values were estimated based on all $B_{perm} = \binom{16}{8} = 12,870$ permutations of the treatment/control labels, except for the two SAM procedures for which $B_{sam} = 1,000$ random permutations were used. Note that the results for the Bonferroni, Holm, and Hochberg procedures are virtually identical, similarly for the unadjusted p–value and Tusher et al. SAM procedures.

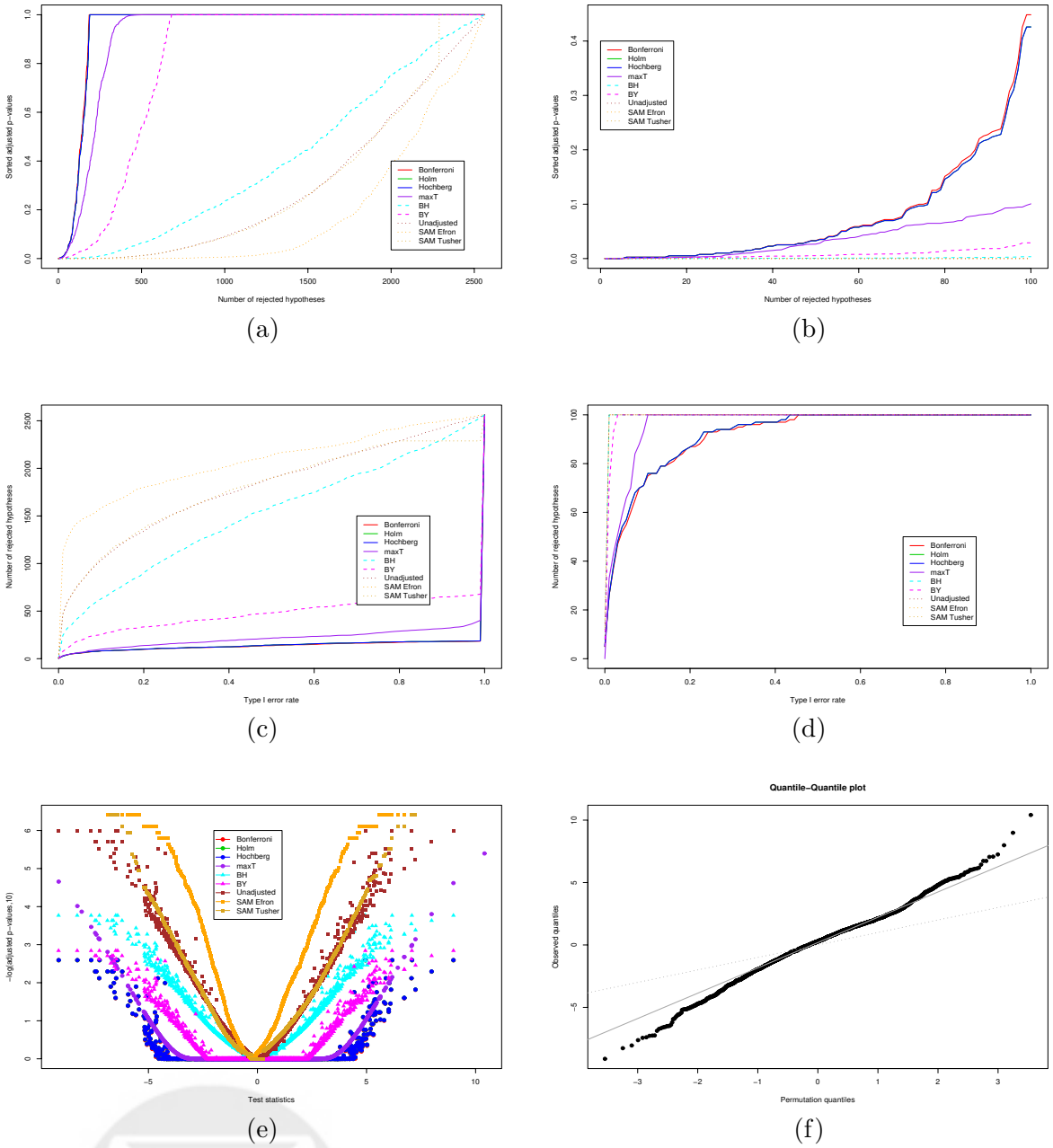


Figure 12: *Bacteria experiment.* Panels (a) and (b): Plot of sorted adjusted p–values $\tilde{p}^*_{r_j}$ *vs.* j. Panels (c) and (d): Number of genes declared differentially expressed, R, *vs.* nominal Type I error rate, α. Panels (b) and (d) are enlargements of panels (a) and (c), respectively, for the 100 genes with the smallest maxT adjusted p–values. Panel (e): Plot of adjusted p–values $-\log_{10}\tilde{p}^*_j$ *vs.* t–statistics t_j. Panel (f): Quantile–Quantile plot of t–statistics; the dotted line is the identity line and the dashed line passes through the first and third quartiles. Adjusted p–values were estimated based on all $B_{perm} = 2^{22}$ permutations of responses *within* the 22 dose $\times$ time blocks, except for the two SAM procedures for which $B_{sam} = 1,000$ random permutations were used. Note that the results for the Bonferroni, Holm, and Hochberg procedures are virtually identical, similarly for the unadjusted p–value and Tusher et al. SAM procedures.

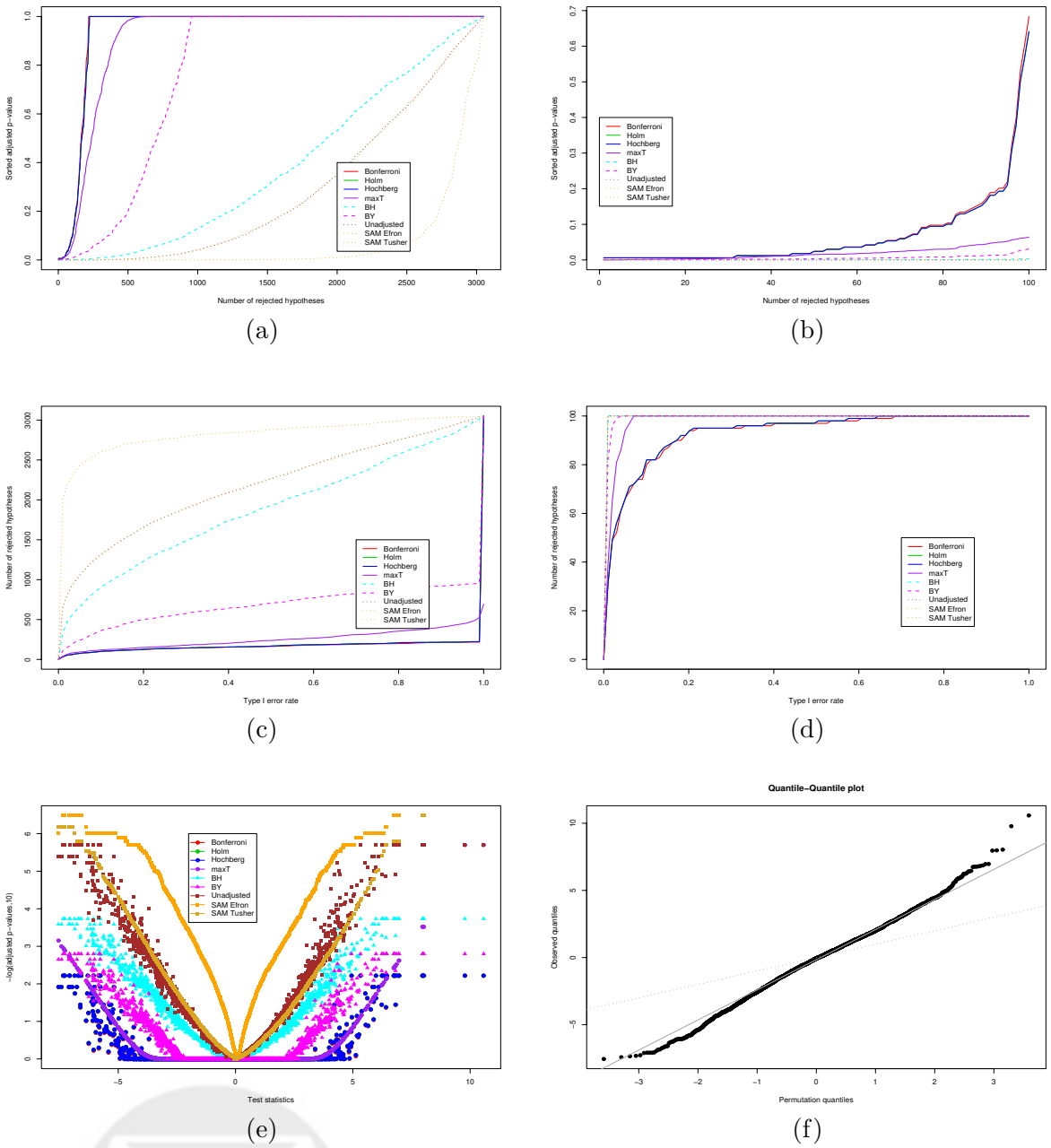
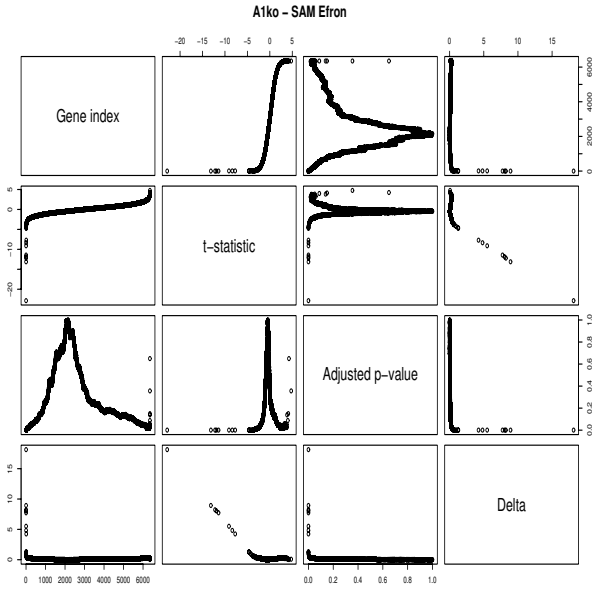


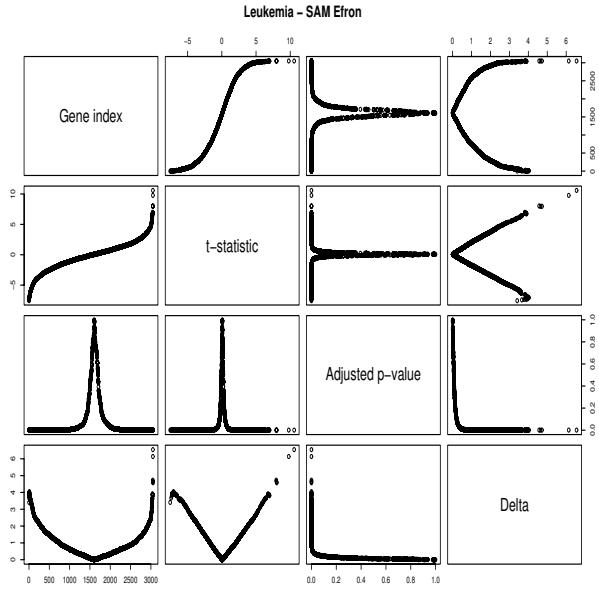
Figure 13: *Leukemia study.* Panels (a) and (b): Plot of sorted adjusted p–values $\tilde{p}^*_{r_j}$ *vs.* j. Panels (c) and (d): Number of genes declared differentially expressed, R, *vs.* nominal Type I error rate, α. Panels (b) and (d) are enlargements of panels (a) and (c), respectively, for the 100 genes with the smallest maxT adjusted p–values. Panel (e): Plot of adjusted p–values $-\log_{10}\tilde{p}^*_j$ *vs.* t–statistics t_j. Panel (f): Quantile–Quantile plot of t–statistics; the dotted line is the identity line and the dashed line passes through the first and third quartiles. Adjusted p–values were estimated based on $B_{perm} = 500,000$ random permutations of the ALL/AML labels, except for the two SAM procedures for which $B_{sam} = 1,000$ random permutations were used. Note that the results for the Bonferroni, Holm, and Hochberg procedures are virtually identical, similarly for the unadjusted p–value and Tusher et al. SAM procedures.

Apo AI experiment

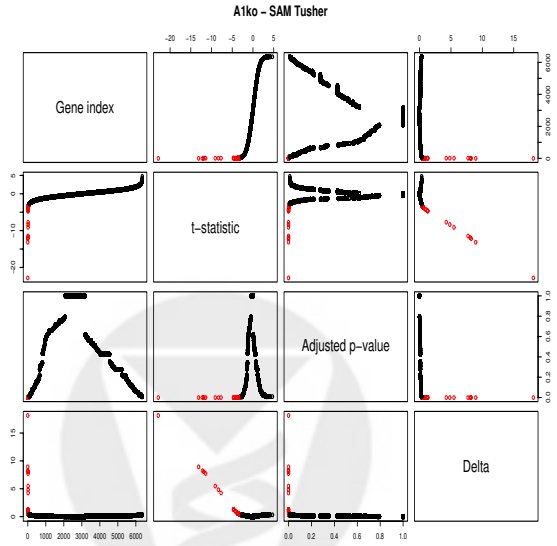
Leukemia study



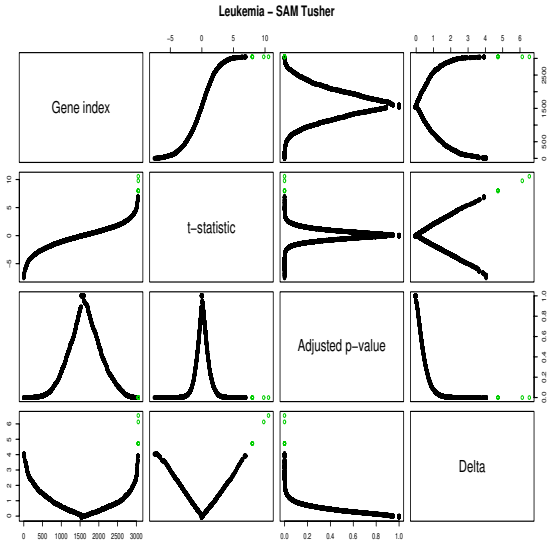
(a)



(b)

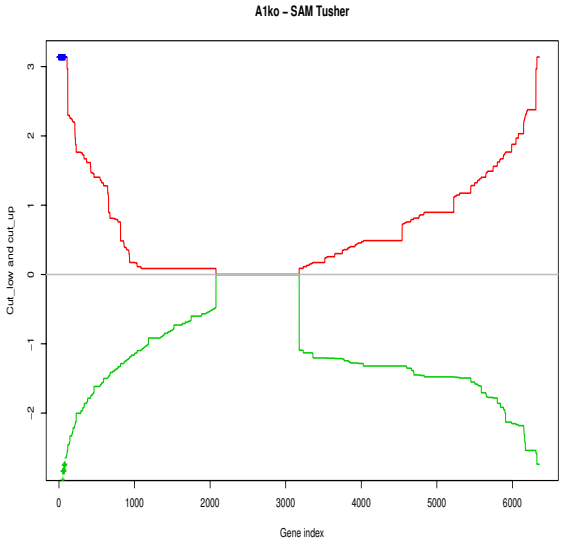


(c)



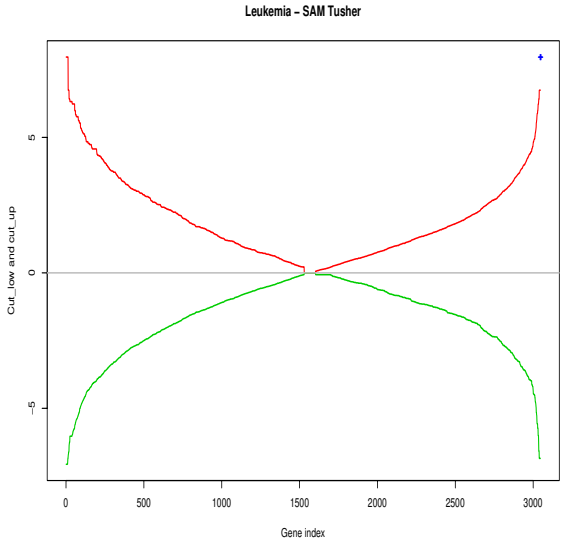
(d)

Apo AI experiment



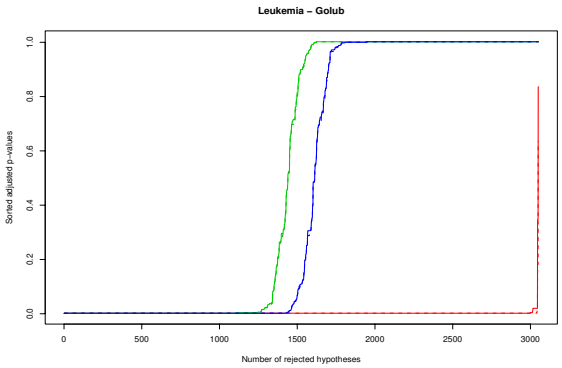
(e)

Leukemia study

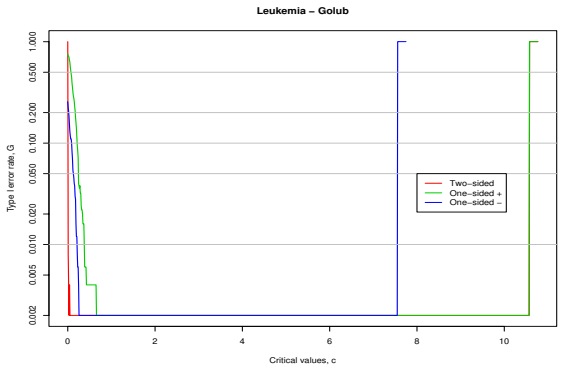


(f)

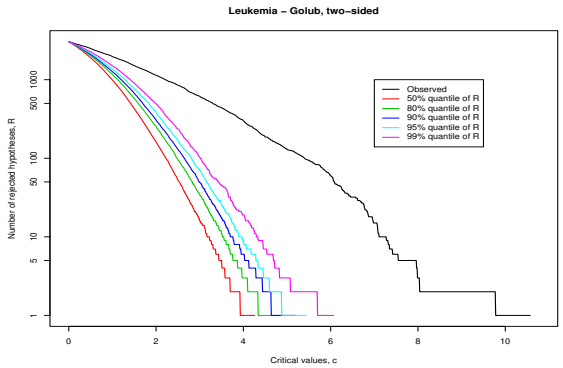
Figure 14: *Apo AI experiment and leukemia study – SAM procedures.* Panels (a) and (b): Efron et al. SAM procedure, pairs plots of test statistics $t_{(j)}$, adjusted p–values $\tilde{p}^*_{(j)}$, and thresholds $\Delta_{(j)}$ used in the calculation of adjusted p–values. Panels (c) and (d): Tusher et al. SAM procedure, pairs plots of test statistics $t_{(j)}$, adjusted p–values $\tilde{p}^*_{(j)}$, and thresholds $\Delta_{(j)}$ used in the calculation of adjusted p–values. Panels (e) and (f) : Tusher et al. SAM procedure, upper cut–offs $cut_{up}(\Delta_{(j)})$ (red) and lower cut–offs $cut_{low}(\Delta_{(j)})$ (green) corresponding to thresholds $\Delta_{(j)}$. Statistical significance was assessed based on $B_{sam} = 1,000$ random permutations of the treatment/control or ALL/AML labels.



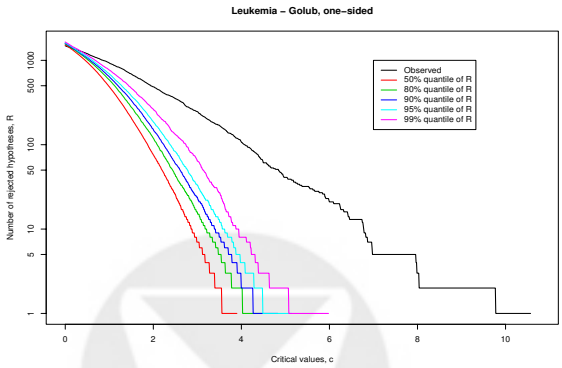
(a)



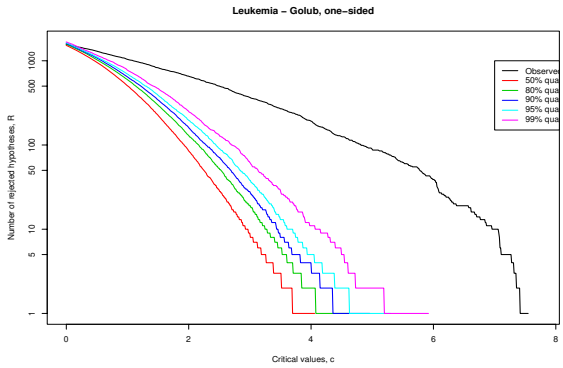
(b)



(c)



(d)



(e)

Figure 15: *Leukemia study – Neighborhood analysis.* Panel (a): plot of step–down and step–up adjusted p–values for neighborhood analysis, for two–sided (red curve) and one–sided (blue curve for over–expression in AML, green curve for over–expression in ALL) alternative hypotheses. Panel (b): plots of the Golub Type I error rates $G(c) = pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$ *versus* critical values c for the three different types of alternatives. Panels (c) – (e): plots of the observed number of rejections $R(c) = r(c)$ and permutation quantiles of $R(c)$ against critical values c for the three types of tests. Statistical significance was assessed based on $B_{golub} = 1,000$ random permutations of the ALL/AML labels.

A Differentially expressed genes in bacteria experiment

Genes with step–down maxT adjusted p–values less than 0.05 are listed below. The index in front of the gene name indicates the ordering of the p–values, where an index of 1 corresponds to the most significant p–value. Genes were separated into two sets: those with positive t–statistics, suggestive of over–expression in *S. aureus* infected cells, and those with negative t–statistics, suggestive of over–expression in *B. pertussis* infected cells.

```
Genes over--expressed in S. aureus
==================================
1 "19260 | 704783 | SHP-1=mutated in motheaten mouse=Protein tyrosine phosphatase, non-recepto
3 13213 | 1319034 | Similar to plasma gelsolin
6 "17729 | 843037 | Vitamin D (1,25- dihydroxyvitamin D3) receptor"
12 17331 | 795343 | CD64=high affinity immunogobulin gamma FC receptor I A form precursor=FC-g
14 21718 | 7387 | Unknown
15 16932 | 470615 | CD64=high affinity immunogobulin gamma FC receptor I A form precursor=FC-g
17 16687 | 297219 | cathepsin B
18 17696 | 470615 | CD64=high affinity immunogobulin gamma FC receptor I A form precursor=FC-g
27 24520 | 1319034 | Similar to plasma gelsolin
30 "17817 | 248828 | Id3=Inhibitor of DNA binding 3, dominant negative helix-loop-helix protei
31 17789 | 137231 | mitogen-responsive phosphoprotein (DOC-2)
32 "16794 | 342927 | Protein tyrosine phosphatase, non-receptor type 9 (PTPase MEG2)"
33 18273 | 1235138 | CD31=PECAM-1
34 16658 | 290563 | Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cycl
35 15851 | 768561 | MCP-1=MCAF=small inducible cytokine A2=JE=chemokine
38 16767 | 325072 | TIMP-2=Tissue inhibitor of metalloproteinase 2
39 19246 | 687129 | Hs.11500 ESTs
41 3487 | 685646 | Ro ribonucleoprotein autoantigen (Ro/SS-A)=autoantigen calreticulin
42 "17272 | 752785 | SHP-1=mutated in motheaten mouse=Protein tyrosine phosphatase, non-recept
45 17234 | 724506 | gamma-interferon inducible gene IP-30
46 16981 | 489258 | Cysteine-rich protein 2=CRP2=ESP1 protein=LIM domain protein
47 16673 | 293742 | udp glucuronosyltransferase
49 19380 | 359769 | mss4=Zn2+ binding protein/guanine nucleotide exchange factor
52 4226 | 261517 | cathepsin B
54 16551 | 236282 | WASP=Wiskott-Aldrich syndrome protein
56 2313 | 712092 | X-CGD gene involved in chronic granulomatous
disease locat ed on chromosome X=Cytochrome B-245 light chain
component of microbicidal oxidase system in phagocytes
57 17033 | 526335 | MMP-9=Matrix metalloproteinase 9=92 kD Gelatinase B=92 KD type IV collagena
58 "14664 | 1355330 | Hs.13255 Homo sapiens mRNA for KIAA0930 protein, partial cds"
59 532 | 686554 | Hs.185058 ESTs
60 21465 | 1341319 | Hs.3337 transmembrane 4 superfamily member 1
61 17660 | 345103 | protein-tyrosine kinase EPHB2v (EPHB2)
62 16867 | 376942 | Ro ribonucleoprotein autoantigen (Ro/SS-A)=autoantigen calreticulin
63 16945 | 472180 | S100 calcium binding protein A4=Placental calcium binding protein=Calvascul
64 2336 | 712278 | c-fos
65 "18453 | 1306024 | CD11C=leukocyte adhesion protein p150,95 alpha subunit=integrin alpha-X"
```

```
66 17650 | 343867 | allograft-inflammatory factor-1=interferon gamma
induced macrophage protein=Iba1=ionized calcium binding adapter
molecule 1

Genes over--expressed in B. pertussis
====================================
2 13023 | 1288054 | IkB alpha
4 16426 | 153355 | LD78 beta=almost identical to MIP-1 alpha=chemokine
5 4081 | 71 | IL-10
7 24829 | 825038 | IkB alpha
8 20417 | 825038 | IkB alpha
9 16967 | 487962 | B4-2 protein
10 16822 | 346550 | MIP-1 alpha=LD78 alpha=pAT464=Small inducible cytokine A3=macrophage infla
11 1537 | 683517 | Hs.129923 ESTs
13 4112 | 98 | CD40 ligand
16 4376 | 153355 | LD78 beta=almost identical to MIP-1 alpha=chemokine
19 15850 | 205633 | MIP1 beta=SCAY2=G-26=HC21=pAT 744=LAG-1=Act-2=H400=SIS-gamma=chemokine
20 24499 | 1288054 | Similar to IkB alpha
21 24107 | 1967950 | Unknown
22 16434 | 156520 | TSG-6=tumor necrosis factor-inducible gene
23 1538 | 683519 | IkB alpha
24 16047 | 84295 | IL-1 receptor antagonist
25 21388 | 1336549 | Hs.82554 ESTs
26 9922 | 1185708 | Hs.163214 ESTs
28 17438 | 1074540 | G-CSF=Colony-stimulating factor 3
29 "22503 | 2072763 | Hs.56009 ESTs, Weakly similar to reverse transcriptase related protein [
36 4089 | 79 | IL-6
37 842 | 703558 | Hs.172051 ESTs
40 4296 | 56 | MIP1 beta=SCAY2=G-26=HC21=pAT 744=LAG-1=Act-2=H400=SIS-gamma=chemokine
43 3760 | 684742 | adenosine deaminase
44 "16916 | 447509 | Major histocompatibility complex, class II, DN alpha"
48 24600 | 1350626 | Similar to (U64842) F25B4.2 gene product
50 16905 | 429238 | BRCA2 region EST-1
51 1515 | 683218 | CD38
53 16427 | 153768 | STAT induced STAT inhibitor-3=CIS3
55 287 | 684794 | Hs.104358 EST
```

B Differentially expressed genes in leukemia study

Genes with step–down maxT adjusted p–values less than 0.05 are listed below. The index in front of the gene name indicates the ordering of the p–values, where an index of 1 corresponds to the most significant p–value. Genes were separated into two sets: those with positive t–statistics, suggestive of over–expression in AML cells, and those with negative t–statistics, suggestive of over–expression in ALL cells.


```
Genes over--expressed in AML
```

```
============================
1 X95735_at Zyxin
2 M27891_at CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
3 M55150_at FAH Fumarylacetoacetate
4 M16038_at LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
5 L09209_s_at APLP2 Amyloid beta (A4) precursor-like protein 2
16 U50136_rna1_at Leukotriene C4 synthase (LTC4S) gene
17 Y12670_at LEPR Leptin receptor
19 Y00787_s_at INTERLEUKIN-8 PRECURSOR
22 U82759_at GB DEF = Homeodomain protein HoxA9 mRNA
23 M23197_at CD33 CD33 antigen (differentiation antigen)
24 M63138_at CTSD Cathepsin D (lysosomal aspartyl protease)
26 X62654_rna1_at ME491  gene extracted from H.sapiens gene for Me491/CD63 antigen
27 X07743_at PLECKSTRIN
33 D88422_at CYSTATIN A
34 L08246_at INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1
35 U67963_at Lysophospholipase homolog (HU-K5) mRNA
37 M28130_rna1_s_at Interleukin 8 (IL8) gene
42 M19045_f_at LYZ Lysozyme
43 U46499_at GLUTATHIONE S-TRANSFERASE, MICROSOMAL
44 D14874_at ADM Adrenomedullin
48 X04085_rna1_at Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13
61 M21551_rna1_at Neuromedin B mRNA
62 X85116_rna1_s_at Epb72 gene exon 1
67 J03801_f_at LYZ Lysozyme
68 M81695_s_at ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)
69 X17042_at PRG1 Proteoglycan 1, secretory granule
73 M62762_at ATP6C Vacuolar H+ ATPase proton channel subunit
79 M22960_at PPGB Protective protein for beta-galactosidase (galactosialidosis)
84 X61587_at ARHG Ras homolog gene family, member G (rho G)
87 X14008_rna1_f_at Lysozyme gene (EC 3.2.1.17)
91 M69043_at MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3
92 X62320_at GRN Granulin

Genes over--expressed in ALL
============================
6 M31523_at TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
7 X74262_at RETINOBLASTOMA BINDING PROTEIN P48
8 Z15115_at TOP2B Topoisomerase (DNA) II beta (180kD)
9 L47738_at Inducible protein mRNA
10 U22376_cds2_s_at C-myb gene extracted from Human (c-myb) gene, complete primary cds, and fi
11 HG1612-HT1612_at Macmarcks
12 M91432_at ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain
13 L41870_at RB1 Retinoblastoma 1 (including osteosarcoma)
14 U72936_s_at X-LINKED HELICASE II
15 X51521_at VIL2 Villin 2 (ezrin)
18 X74801_at T-COMPLEX PROTEIN 1, GAMMA SUBUNIT
```

20 J05243_at SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
21 U26266_s_at DHPS Deoxyhypusine synthase
25 M12959_s_at TCRA T cell receptor alpha-chain
28 M31211_s_at MYL1 Myosin light chain (alkali)
29 U62136_at Putative enterocyte differentiation promoting factor mRNA, partial cds
30 X15949_at IRF2 Interferon regulatory factor 2
31 U32944_at Cytoplasmic dynein light chain 1 (hdlc1) mRNA
32 L07758_at IEF SSP 9502 mRNA
36 S50223_at HKR-T1
38 M92287_at CCND3 Cyclin D3
39 U29175_at Transcriptional activator hSNF2b
40 U49020_cds2_s_at MEF2A gene (myocyte-specific enhancer factor 2A, C9 form) extracted from Hu
41 M89957_at IGB Immunoglobulin-associated beta (B29)
45 U73737_at GTBP DNA G/T mismatch-binding protein
46 M29696_at IL7R Interleukin 7 receptor
47 M94633_at GB DEF = Recombination acitivating protein (RAG2) gene, last exon
49 M83233_at TCF12 Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)
50 X59350_at CD22 CD22 antigen
51 M11722_at Terminal transferase mRNA
52 X62535_at DAGK1 Diacylglycerol kinase, alpha (80kD)
53 X63753_at SON SON DNA binding protein
54 X82240_rna1_at TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA for Tcell leukemia/
55 M77142_at NUCLEOLYSIN TIA-1
56 Z69881_at Adenosine triphosphatase, calcium
57 U79285_at GLYCYLPEPTIDE N-TETRADECANOYLTRANSFERASE
58 Y08612_at RABAPTIN-5 protein
59 U27460_at Uridine diphosphoglucose pyrophosphorylase mRNA
60 M29536_at Translational initiation factor 2 beta subunit (elF-2-beta) mRNA
63 X63469_at GTF2E2 General transcription factor TFIIE beta subunit, 34 kD
64 D38073_at MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3
65 U05259_rna1_at MB-1 gene
66 M60527_at DCK Deoxycytidine kinase
70 D14658_at KIAA0102 gene
71 U20998_at SRP9 Signal recognition particle 9 kD protein
72 U47077_at DNA-dependent protein kinase catalytic subunit (DNA-PKcs) mRNA
74 D88270_at GB DEF = (lambda) DNA for immunoglobin light chain
75 M13792_at ADA Adenosine deaminase
76 L05148_at Protein tyrosine kinase related mRNA sequence
77 D87078_at KIAA0235 gene, partial cds
78 U72342_at PLATELET-ACTIVATING FACTOR ACETYLHYDROLASE 45 KD SUBUNIT
80 D26156_s_at Transcriptional activator hSNF2b
81 X56468_at 14-3-3 PROTEIN TAU
82 D86967_at KIAA0212 gene
83 X68560_at SP3 Sp3 transcription factor
85 U38846_at Stimulator of TAR RNA binding (SRB) mRNA
86 L28010_at HnRNP F protein mRNA
88 U49844_at Protein kinase ATR mRNA

89 U31556_at E2F5 E2F transcription factor 5, p130-binding
90 L25931_s_at LBR Lamin B receptor